informatyka+

Algorytmika i programowanie

Bazy danych

Multimedia, grafika i technologie internetowe

Sieci komputerowe

Tendencje w rozwoju informatyki i jej zastosowań

Człowiek – najlepsza inwestycja

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

Warszawska Wyższa Szkoła Informatyki

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

informatyka+

Wszechnica Poranna:

Tendencje w rozwoju informatyki i jej zastosowań

Odkrywanie struktur ukrytych w danych, czyli eksploracja danych

Michał Grabowski

Człowiek – najlepsza inwestycja

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

Warszawska Wyższa Szkoła Informatyki

UNIA EUROPEJSKA
EUROPEJSKI
FUNDUSZ SPOŁECZNY

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Odkrywanie struktur ukrytych w danych, czyli eksploracja danych

Rodzaj zajęć: Wszechnica Poranna
Tytuł: Odkrywanie struktur ukrytych w danych, czyli eksploracja danych
Autor: dr hab. Michał Grabowski

Redaktor merytoryczny: prof. dr hab. Maciej M Sysło

Zeszyt dydaktyczny opracowany w ramach projektu edukacyjnego **Informatyka+** — ponadregionalny program rozwijania kompetencji uczniów szkół ponadgimnazjalnych w zakresie technologii informacyjno-komunikacyjnych (ICT).
www.informatykaplus.edu.pl
kontakt@informatykaplus.edu.pl

Wydawca: Warszawska Wyższa Szkoła Informatyki
ul. Lewartowskiego 17, 00-169 Warszawa
www.wwsi.edu.pl
rektorat@wwsi.edu.pl

Projekt graficzny: FRYCZ I WICHA

Warszawa 2009

KAPITAŁ LUDZKI
NARODOWA STRATEGIA SPÓJNOŚCI

WARSZAWSKA WYŻSZA SZKOŁA INFORMATYKI

UNIA EUROPEJSKA
EUROPEJSKI FUNDUSZ SPOŁECZNY

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Odkrywanie struktur ukrytych w danych, czyli eksploracja danych

Michał Grabowski
Warszawska Wyższa Szkoła Informatyki
michal.grabowski@twp.olsztyn.pl

Streszczenie

Na początku zostanie podana seria prostych przykładów, aby słuchacze mogli w miarę łatwo odkryć reguły ukryte w danych. Proces odkrywania reguł można interpretować jako algorytmu uczenia się systemu ze zbioru treningowego. Podane zostaną dwa przykłady danych z głębiej ukrytą strukturą, nie do zdroworozsądkowego zauważenia. Jeden z tych przykładów jest o naturze statystycznej, drugi – o naturze kombinatorycznej, zorientowany na zastosowanie drzewa decyzyjnego. Podane zostanie intuicyjne objaśnienie wykrycia rozkładu normalnego ukrytego w danych – histogram, standaryzacja wartości danych, zastosowanie rozkładu normalnego do sformułowania prognozy dotyczącej danych z przykładu o naturze statystycznej. Następnie zostanie podana definicja i przykład drzewa decyzyjnego opartego o zbiór, przyjętych jako dostępne, testów na danych. Zwrócona zostanie uwaga na znaczenie ekspresywności języka, w którym próbujemy sformułować hipotezę o strukturze ukrytej w danych. Następnie sformułowany zostanie zbiór dostępnych testów dla analizy przykładu o naturze kombinatorycznej i podane zostanie intuicyjne objaśnienie klasycznego algorytmu indukcji z danych drzewa decyzyjnego, w tym kryterium wyboru testu przez entropię. Intuicyjnie zostanie wyprowadzone z danych przykładu drugiego drzewo decyzyjne i zastosowane do sklasyfikowania danych przykładu drugiego.

Na zakończenie, ostrzeżenie, że eksploracja danych jest szeroką dziedziną oferującą dziesiątki (a może setki) algorytmów, podana zostanie również informacja o niektórych zastosowaniach algorytmów eksploracji danych.

Przedmiotem warsztatów będzie:

- Analiza statystyczna przykładowych danych z użyciem tylko arkusza Excel.
- Wspomagane arkuszem Excel wyliczenia prowadzące do skonstruowania drzewa decyzyjnego z podanego konkretnego zbioru treningowego.
- Programowanie prostych algorytmów użytecznych dla zrozumienia zjawiska nadmiernego dopasowania.

Spis treści

1. Wstęp 3

2. O pewnym problemie dużej firmy taksówkowej 3
2.1. Histogram 4
2.2. Zmienna losowa 5
2.3. Rozkład normalny 8
2.4. Ocena wiarygodności hipotezy 10

3. Drzewa decyzyjne 13
3.1. Zastosowanie drzewa decyzyjnego do klasyfikacji danych 13
3.2. Błąd klasyfikatora i walidacja krzyżowa 18
3.3. Zjawisko nadmiernego dopasowania 20

4. Propozycja przeprowadzenia prostych badań 21

5. Niektóre dziedziny zastosowań metod eksploracji danych 22

Literatura 23

1 WSTĘP

Odkrywanie struktury ukrytej w danych zaprezentujemy na dwóch przykładach. Jeden z tych przykładów (rozdział drugi, *O pewnym problemie dużej firmy taksówkowej*) będzie miał naturę statystyczną, a drugi (rozdział trzeci, *Drzewa decyzyjne*) – kombinatoryczną. Przykład o naturze kombinatorycznej należy do dziedziny **systemów uczących się** (ang. *machine learning*). Dziedzina ta zajmuje się metodami konstruowania ogólnych klasyfikatorów ze zbiorów trenujących. Podamy tu minimalny aparat pojęciowy dotyczący drzew decyzyjnych, konieczny do postawienia pewnego problemu do samodzielnego zbadania przez uczniów, którzy umieją i lubią programować. Przykład o naturze statystycznej jest zastosowaniem klasycznej metody dopasowania rozkładu prawdopodobieństwa do danych. Statystykę włączyliśmy do prezentowanego materiału dlatego, że jest ona bardzo ważna w eksploracji danych. Na przykład, statystyczne systemy uczące się to bardzo ważny dział eksploracji danych, chyba o najsilniejszych zastosowaniach.

Eksploracja danych jest obszerną dziedziną informatyki, oferującą dziesiątki metod i chyba setki algorytmów. Naszym celem jest tylko opowiedzenie o tej dziedzinie na poziomie intuicyjnym i przekonanie słuchaczy, że istnieją poważne jej zastosowania. Nie należy tego materiału traktować jako wstępu do eksploracji danych, wstępu, który ma dać pewien (choćby minimalny) ogląd tej dziedziny.

2 O PEWNYM PROBLEMIE DUŻEJ FIRMY TAKSÓWKOWEJ

Wyobraźmy sobie, że zarząd dużej firmy taksówkowej potrzebuje następującej informacji: jaki procent zamówień (w perspektywie kilkuletniej) będzie dotyczyć odległości co najmniej 40-stu kilometrów? Mówimy tu o odległościach między miejscem zamówienia a zajezdnią przy lotnisku. Powiedzmy, że informacja ta jest istotna przy podejmowaniu decyzji, jakiego rodzaju samochody i w jakich proporcjach zakupić. Firma dopiero zaczyna swoją działalność i nie ma danych o zamówieniach z poprzednich lat. Można tylko obserwować bieżącą działalność firmy.

Czy możesz zaproponować jakiś sposób rozwiązania,
bez zatrudniania drogiego specjalisty?

Następujący sposób wydaje się być bardzo naturalny. Weźmy, przykładowo, 100 losowych próbek odległości miejsce zamówienia-zajezdnia z bieżącej tygodniowej działalności. Policzmy, ile tych wartości to co najmniej 40 km i mamy stosowny procent.

Powiedzmy, że pobrano losową populację 100 próbek odległości miejsce zamówienia-zajezdnia, pokonywanych w jedną stronę przez taksówki. Wyniki są podane w tab. 1 (ułożenie próbek w tablicy kwadratowej nie ma żadnego znaczenia).

Tabela 1.
Próba pierwsza losowych odległości

21	5	36	54	7	14	43	7	3	24
22	41	56	2	10	39	48	2	43	14
36	29	49	8	42	35	23	35	41	28
56	30	22	9	46	15	58	28	21	56
37	40	14	55	41	36	23	55	53	52
69	35	51	62	41	15	65	15	29	26
60	29	22	70	44	42	16	48	77	44
29	22	28	41	48	3	54	55	42	68
30	29	11	46	17	69	68	4	76	28
31	32	47	56	21	55	36	52	48	83

Zadanie 1. Wpisz dane z tab. 1 do arkusza Excel i po posortowaniu danych oblicz stosowny procent.

Zapewne otrzymasz, że 47% kursów taksówek miało długość co najmniej 40 km. Jeżeli nie sprawdziłeś tego, to jednak sprawdź. Czy można zaufać tej prognozie? Pewien ostrożny członek zarządu zdecydował, że należy jeszcze raz dokonać pomiaru losowych próbek odległości. Wyniki losowych próbek z kolejnego tygodnia są zamieszczone w tab. 2.

Odpowiedni procent dla drugiej próby losowej wyniósł 40%. Sprawdź to. Różnica z poprzednią prognozą to 7% – powiedzmy, że nie do zaakceptowania z punktu widzenia celów zarządu. To naturalne rozwiązanie okazuje się być wysoce niepewne. Co można zaproponować? Może branie prób losowych z kolejnych tygodni i wzięcie średniej prognozy? Nie jest to całkiem wiarygodna propozycja. Pomyśl, co się dzieje, jeżeli jest pewna sezonowość: latem jest trochę więcej zamówień z dalszych odległości, zimą trochę mniej. Aby radykalnie wyeliminować propozycję brania prób losowych z kolejnych tygodni, załóżmy, że stosowna informacja jest potrzebna „na wczoraj" i nie ma czasu przeciągać na tygodnie obliczenia rozwiązania.

Czy w takim razie w ogóle można rozwiązać nasze zadanie? Można. Wiodącą ideą jest następująca myśl: w zebranych danych ukryta jest jakaś struktura, jakieś prawo. Jeżeli odkryjemy tę strukturę, to prawo, to będziemy mogli sformułować sensowną prognozę procentu odległości większych równych od 40-stu kilometrów.

Zadanie wykrycia ukrytej w danych struktury było rozważane jeszcze w czasach przed powstaniem komputerów. Bardzo ważne metody wykryła statystyka matematyczna zajmująca się zastosowaniem rachunku prawdopodobieństwa w analizie danych. Powszechna dostępność stosunkowo dużych mocy obliczeniowych komputerów dała silny impuls do rozwoju nowych dziedzin informatyki: **eksploracji danych** (ang. *data-mining*), **systemów uczących się** (ang. *machine-learning*), **hurtowni danych** (ang. *date warehouses*) i do dalszego rozwoju statystyki.

Drogi uczniu, będziemy próbowali wykryć strukturę w podanych stu danych, używając tylko arkusza Excel. Użyjemy metody proponowanej przez statystykę matematyczną. Metody statystyczne próbują „wygładzić" skończony zbiór danych, dopasowując do danych stosowny rozkład prawdopodobieństwa. Pasujący do danych rozkład prawdopodobieństwa to prawo ukryte w danych, którego będziemy poszukiwać i w oparciu o nie obliczymy rozwiązanie naszego zadania. Nie wszystkie istotne kwestie będą do końca zmatematyzowane. Będziemy odwoływać się do intuicji.

Wykrywanie ukrytych w danych struktur i praw o naturze innej niż statystyczna to przedmiot badań dziedzin informatyki, takich jak: **systemy uczące się** (ang. *machine-learning*), **teoria zbiorów przybliżonych** (ang. *rough-sets theory*), **systemy wielo-agentowe** (ang. *multi-agent systems*), **sieci neuronowe** (ang. *neural nets*), **algorytmy genetyczne** (ang. *genetic algorithms*). Możesz poznać te zagadnienia, gdy wybierzesz studia matematyczne lub informatyczne na dobrej uczelni.

Zrealizowanie naszego zamierzenia wykrycia struktury w naszych stu danych wymaga poznania pojęć, takich jak:

- histogram danych,
- zmienna losowa – ciągła lub dyskretna,
- rozkład prawdopodobieństwa zmiennej losowej,
- wartość średnia i odchylenie standardowe zmiennej losowej,
- rozkład normalny,
- miara χ-kwadrat dopasowania danych do rozkładu.

Po kolei omówimy te pojęcia bez ścisłych podstaw matematycznych, odwołując się w dużym stopniu do intuicji. Nie będzie to bardzo trudne. No cóż, na razie będzie troszeczkę tzw. teorii, a mniej ręcznej roboty klikania w arkusz. Na pociechę, pojawiające się pośrednie zadania będziemy rozwiązywać używając arkusza Excel.

2.1 HISTOGRAM

W przypadku danych dyskretnych, **histogram** danych daje informację o częstości wystąpienia danej wartości wśród danych. W przypadku danych ciągłych, histogram daje informację o liczbie danych należących do zadanych przedziałów.

Przed przystąpieniem do konstruowania histogramu danych należy zdecydować, czy dane są ciągłe czy dyskretne. Histogram dla danych ciągłych konstruuje się trochę inaczej niż histogram dla danych dyskretnych. Wprawdzie nasze dane są liczbami naturalnymi, lecz zasadniczo odległości mogą być dowolnym ułamkiem. Decydujemy więc, że nasze dane są ciągłe.

Przykładem **danych dyskretnych** są wyniki ankiety studenckiej oceny pewnego semestralnego wykładu w skali dyskretnej od –5 do 5, tzn. że możliwe oceny należą do zbioru {–5, –4, –3, –2, –1, 0, 1, 2, 3, 4, 5}, a przykładem **danych ciągłych** są wyniki pomiaru np. temperatury albo odległości.

Konstrukcja histogramu danych ciągłych

- Posortuj dane.
- Podziel posortowane dane na przedziały. W przypadku 100 danych powszechną praktyką jest wzięcie od 10 do 15 przedziałów. Równie powszechną praktyką jest branie takich przedziałów, że przypada co najmniej od 5 do 8 danych na przedział. W naszym przypadku po prostu bierzemy przedziały potencjalnie po 7 danych: [0, 7), [7, 14), [14, 21), [21, 28), [28, 35), [35, 42), [42, 49), [49, 56), [56, 63), [63, 70), [70, 77), [77, 84).
- Oblicz, ile danych wpada do pierwszego przedziału, do drugiego przedziału, do kolejnych przedziałów, aż po ostatni – to jest właśnie **początkowy histogram**.
- Łączymy przylegające przedziały, do których wpadło mniej niż 5 danych i dostajemy **wynikowy histogram**.

Zadanie 2. Posługując się arkuszem Excel oblicz histogram początkowy bez złączenia przylegających przedziałów zawierających mniej niż 5 danych.

Wskazówka. Popatrz kolejno na posortowane dane i zlicz, ile danych wpada do kolejnych przedziałów [0, 7), [7, 14), [14, 21),... Ambitniejsi uczniowie mogą te obliczenia zaprogramować w języku Visual Basic for Excel jako funkcję $h(i)$ obliczającą liczbę danych wpadających do przedziału nr i.

Powinieneś otrzymać histogram początkowy przedstawiony w tab. 3.

Tabela 3.
Histogram bez łączenia przedziałów zawierających mniej niż 5 danych

Nr. przedz.	[lewy kr.	prawy kr.)	liczba danych
0	0	7	6
1	7	14	7
2	14	21	8
3	21	28	11
4	28	35	13
5	35	42	15
6	42	49	14
7	49	56	11
8	56	63	7
9	63	70	5
10	70	77	2
11	77	83	2

Histogram po złączeniu przylegających przedziałów, zawierających mniej niż 5 danych, jest przedstawiony w tab. 4. Połączyliśmy trzy przedziały [63, 70), [70, 77), [77, 83) dlatego, że połączenie dwóch przedziałów [70, 77), [77, 83) dało tylko 4 dane w połączonym przedziale – nadal mniej niż 5. Prawy kraniec ostatniego przedziału ustaliliśmy na 999 jako reprezentację $+\infty$. Chodzi o to, że wszystkie dane większe od lub równe 63 traktujemy już jako dane wpadające do jednego przedziału $[63,+\infty)$.

Gdy chcemy wyjaśnić sens pojęcia histogramu i co więcej, rozwiązać opisany problem dużej firmy taksówkowej, musimy odwołać się do pojęcia zmiennej losowej i, co gorsza, do jej rozkładu prawdopodobieństwa.

2.2 ZMIENNA LOSOWA

Drogi uczniu, zapewne nie przepadasz za prawdopodobieństwem i pojęciami z nim związanymi, a w szczególności za zmiennymi losowymi. Postaramy się wyjaśnić potrzebne pojęcia na prostych przykładach, nie wchodząc w matematyczne definicje.

Tabela 4.
Histogram po złączeniu przylegających przedziałów zawierających mniej niż 5 danych

Nr. przedz.	[lewy kr.	prawy kr.)	liczba danych
0	0	7	6
1	7	14	7
2	14	21	8
3	21	28	11
4	28	35	13
5	35	42	15
6	42	49	14
7	49	56	11
8	56	63	7
9	63	999	9

Przykład 1. Słynna zmienna dyskretna rzut_kostką. Czy domyślasz się, jaki jest zbiór wartości, które może przyjąć ta zmienna? Oczywiście tym zbiorem jest {1, 2, 3, 4, 5, 6}. A jaki jest rozkład prawdopodobieństwa tej zmiennej, czyli z jakim prawdopodobieństwem zmienna dyskretna przyjmuje poszczególne swoje wartości? Zapewne czujesz, że jeżeli kostka jest rzetelna to rozkład prawdopodobieństwa tej zmiennej jest jednostajny: każda wartość ze zbioru {1, 2, 3, 4, 5, 6} może być przyjęta z jednakowym prawdopodobieństwem 1/6.

Przykład 2. Dzienna sprzedaż jednostek towaru *x* w pewnym sklepie. Czy domyślasz się, jaki jest zbiór wartości, które może przyjąć ta dyskretna zmienna losowa? Zbiorem tym jest cały zbiór liczb naturalnych N = {0, 1, 2,...}. Drogi uczniu, zapewne słusznie zauważysz, że astronomicznie wielkie wartości nie będą wartościami tej zmiennej, więc dlaczego cały zbiór liczb naturalnych jest przyjęty jako zbiór wartości zmiennej? Odpowiedź jest następująca: prawdopodobieństwo przyjęcia astronomicznie wielkich wartości przez tą zmienną jest równe 0. O rozkładzie prawdopodobieństwa tej zmiennej nic nie możemy powiedzieć poza tym, że prawdopodobieństwo przyjęcia bardzo dużych wartości jest 0. Więcej substancjalnych informacji o rozkładzie prawdopodobieństwa tej zmiennej statystycy potrafią podać badając histogramy dostatecznie licznych prób wartości tej zmiennej.

Przykład 3. Odległość miejsca zamówienia taksówki od zajezdni, czyli zmienna losowa, której próbę stu wartości analizujemy. Jak już wspomnieliśmy, zasadniczo odległość może być dowolnym ułamkiem, więc zbiorem wartości, które może przyjąć ta zmienna, jest zbiór liczb rzeczywistych **R**. Jest to ciągła zmienna losowa. Drogi uczniu, jakie jest prawdopodobieństwo, że zmienna przyjmie wartość, przykładowo, $z = 27{,}835281735901209827$ km? Dla ciągłych zmiennych losowych możemy sensownie pytać nie o to, jakie jest prawdopodobieństwo przyjęcia jakiejś konkretnej wartości, lecz o to, jakie jest prawdopodobieństwo przyjęcia przez zmienną wartości z mniejszej od zadanego $x \in \mathbf{R}$. Rozkład prawdopodobieństwa zmiennej losowej ciągłej, w odróżnieniu od zmiennej dyskretnej, daje informację nie o tym, jakie jest prawdopodobieństwo przyjęcia danej konkretnej wartości $x \in \mathbf{R}$, lecz o tym, jakie jest prawdopodobieństwo przyjęcia wartości mniejszej od danej konkretnej wartości $x \in \mathbf{R}$. Co możemy powiedzieć o rozkładzie prawdopodobieństwa zmiennej odległość? Na razie tylko tyle, że prawdopodobieństwo tego, że zmienna losowa odległość przyjmie wartość mniejszą od bardzo dużej liczby *x* jest bliskie 1 i prawdopodobieństwo, że zmienna losowa odległość przyjmie wartość mniejszą od bardzo małej liczby *x* jest bliskie 0. Dokładniejsze zbadanie rozkładu prawdopodobieństwa tej zmiennej jest naszym celem.

Uczniowie odważni i zdecydowani na wszystko mogą przeczytać następujące definicje dokładniej wyjaśniające omawiane pojęcia, ale nie jest to konieczne do zrozumienia dalszej części tego opracowania.

Zmienna losowa ciągła

Formalnie rzecz biorąc, ciągła zmienna losowa jest funkcją o wartościach rzeczywistych, określoną na pewnej przestrzeni z prawdopodobieństwem *X*:

$$\varphi: X \to \mathbf{R}.$$

Zwykle przestrzeń X nie jest wyraźnie wyspecyfikowana, jedynie możemy obserwować wartości zmiennej losowej φ. Wymaga się, by dla każdej liczby rzeczywistej $x \in \mathbf{R}$ następujące prawdopodobieństwo było dobrze określone:

$$f(x) = \Pr(\{e \in X \mid \varphi(e) \leq x\}).$$

Funkcja f to właśnie **rozkład prawdopodobieństwa ciągłej zmiennej losowej** φ. Wartość $f(x)$ jest prawdopodobieństwem tego, że wartość zmiennej losowej φ będzie mniejsza lub równa x.

Nie od rzeczy jest byś zajrzał do Twojego podręcznika matematyki do rozdziałów o rachunku prawdopodobieństwa i poznał stosowne definicje i przykłady, ale nie jest to konieczne.

W naszym przypadku zakładamy, że obserwowane odległości są wartościami pewnej ciągłej zmiennej losowej.

Zmienna losowa dyskretna

Niech A będzie pewnym podzbiorem (skończonym lub nie) zbioru liczb całkowitych Z. Dyskretna zmienna losowa jest funkcją określoną na pewnej przestrzeni X z prawdopodobieństwem, o wartościach w zbiorze A.

$$\varphi: X \to A.$$

Wymagamy, by dla każdego elementu $a \in A$ następujące prawdopodobieństwo było dobrze określone:

$$f(a) = \Pr(\{x \in X \mid \varphi(x) = a\}).$$

Funkcja f to **rozkład prawdopodobieństwa dyskretnej zmiennej losowej** φ. Wartość $f(a)$ jest prawdopodobieństwem tego, że dyskretna zmienna losowa φ przyjmie wartość a. Tak jak w przypadku ciągłej zmiennej losowej, przestrzeń X zwykle nie jest wyspecyfikowana. Możemy jedynie obserwować wartości dyskretnej zmiennej losowej. Można zajrzeć do podręcznika, aby poznać więcej przykładów dyskretnych zmiennych losowych.

Liczba odwiedzin w ciągu dnia pewnej strony internetowej w kolejnych dniach to kolejny przykład wartości dyskretnej zmiennej losowej.

Wartość średnia, wariancja i odchylenie standardowe zmiennej losowej

Intuicyjnie mówiąc, **wariancja** (czy **odchylenie standardowe**) zmiennej losowej jest liczbą mierzącą, jak bardzo wartości zmiennej różnią się od wartości średniej – czy to w kierunku wartości większych, czy mniejszych od wartości średniej zmiennej losowej. Dla naszych celów nie są niezbędne ścisłe definicje wymienionych wyżej pojęć. Zamiast tych definicji użyjemy tzw. **estymatorów** wymienionych wielkości: wyliczanie z losowej próby wartości zmiennej przybliżenia jej średniej, przybliżenia jej wariancji i odchylenia standardowego. Prawdziwą wartość średnią zmiennej będziemy oznaczać symbolem μ, a odchylenie standardowe symbolem δ.

Wróćmy do naszej próby stu odległości miejsca zamówienia od zajezdni. Liczność próby oznaczymy jako $n = 100$. Średnią μ naszej zmiennej losowej przybliżymy jako wartość średnią

$$\underline{x} = (\text{suma odległości } x \text{ danej losowej populacji odległości})/n.$$

Wariancję w naszej zmiennej losowej przybliżymy jako następującą sumę:

$$w = \sum_{x \in \text{próba odległości}} (x - \underline{x})^2/(n-1)$$

Różnica między wartością x i średnią $\underline{x}$ jest podniesiona do kwadratu. Czy domyślasz się, dlaczego? Dlatego, że w sumie bardziej mają „ważyć" wartości x odległe od średniej $\underline{x}$ niż wartości bliskie średniej $\underline{x}$. Dzielenie

jest przez $n - 1$ a nie przez n dlatego, by przyjęte przybliżenie wariancji było bardziej wiarygodne. Wynika to z podstaw matematycznych naszej metody, które to podstawy pomijamy.

Odchylenie standardowe δ naszej zmiennej losowej przybliżamy jako pierwiastek kwadratowy s z przybliżenia w wariancji.

$$s = \sqrt{w}.$$

Przyjmujemy teraz dość arbitralne założenie: wielkości $\underline{x}$, w, s dobrze przybliżają prawdziwą wartość średnią μ, prawdziwą wariancję i odchylenie standardowe δ naszej zmiennej losowej.

Zadanie 3. Oblicz w arkuszu Excel wartość $\underline{x}$ dla naszej próby stu odległości.

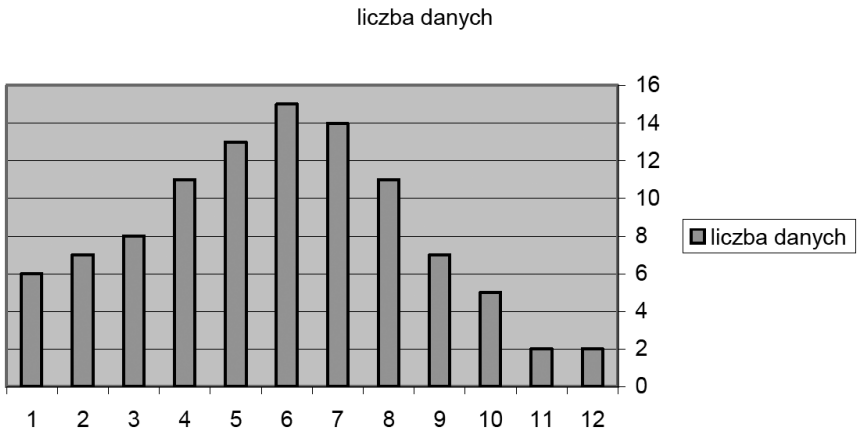
Zadanie 4. Znajdź w pomocy arkusza Excel, wśród funkcji statystycznych funkcję obliczającą odchylenie standardowe serii danych i oblicz wartość s naszej próby odległości. Można to obliczenie także zaprogramować samemu w języku Visual Basic for Excel na podstawie podanych wzorów na $\underline{x}$, w i s.

Nasze dane traktujemy jako próbę wartości zmiennej losowej, której wartościami są odległości miejsca zamówienia od zajezdni. Histogram jest pewnym przybliżeniem rozkładu prawdopodobieństwa zmiennej losowej. Zauważ, że jeżeli podzielisz wartości naszego histogramu przez 100 (liczność próby), to otrzymasz liczby, które można interpretować jako przybliżenie prawdopodobieństwa tego, że wartość zmiennej losowej będzie należała do stosownego przedziału.

Zadanie 5. Jakie jest przybliżone prawdopodobieństwo tego, że zdarzy się zamówienie o odległości miejsca zamówienia od zajezdni większej równej niż 28 i mniejszej niż 35?

Zadanie 6. Przedstaw histogram początkowy jako wykres (słupkowy).

Wykres słupkowy początkowego histogramu, przed łączeniem przylegających przedziałów zawierających mniej niż 5 danych jest przedstawiony na rys.1. Liczby poniżej słupków oznaczają kolejne przedziały: 1 – przedział [0, 7),2 – przedział [7, 14) itd.



Rysunek 1.
Wykres słupkowy histogramu Próby 1 (Tabela 1) odległości

Wyprzedzając trochę omówienie rozkładu normalnego powiemy, że kształt, w jaki układają się słupki histogramu sugeruje, że dane pasują do rozkładu normalnego. Zauważ, że wartość średnia należy do przedziału, do którego wpada najwięcej danych.

Zadanie 7. Oblicz histogram i wykres dla drugiego zestawu stu odległości.

Zapewne stwierdzisz, że dostałeś prawie taki sam histogram! Zaczyna się ujawniać struktura ukryta w naszych danych. Dane z obu zestawów mają coś wspólnego: pasują do pewnego rozkładu prawdopodobieństwa, który to rozkład chcemy wykryć.

2.3 ROZKŁAD NORMALNY

Mówiąc niedokładnie, ale obrazowo, jeżeli wartości zmiennej losowej bliskie średniej są bardziej prawdopodobne i prawdopodobieństwo wartości symetrycznie maleje, gdy wartości są coraz dalsze od średniej, to zapewne rozkład prawdopodobieństwa zmiennej losowej jest tzw. **normalny**. Po prostu, wartości zmiennej bliskie średniej są bardziej prawdopodobne niż te bardziej odległe od średniej.

Zadanie 8. Zbierz dane wzrostu uczniów, powiedzmy z dwóch, trzech klas, wpisz je do arkusza Excel i oblicz histogram zebranych danych oraz naszkicuj jego wykres. Sprawdź, czy można podejrzewać, że rozkład prawdopodobieństwa zmiennej losowej 'wzrost ucznia' jest normalny.
Uwaga. Należycie dobierz przedziały histogramu.

Histogram naszych stu danych odległości sugeruje, że mamy do czynienia z rozkładem normalnym. Odległości bliskie średniej $\underline{x}$ są bardziej prawdopodobne niż odległości dalsze od średniej.

Funkcja gęstości rozkładu normalnego

Uczniowie odważni i zdecydowani na wszystko mogą nie zasłaniać oczu i spojrzeć na słynny wzór gęstości prawdopodobieństwa rozkładu normalnego:

$$f(x) = \frac{1}{\sqrt{2\pi}\delta} e^{-\frac{1}{2}\left(\frac{x-\mu}{\delta}\right)^2}$$

gdzie δ jest odchyleniem standardowym zmiennej losowej φ o rozkładzie o powyższej gęstości, a μ jest średnią tej zmiennej.

Jaki jest sens pojęcia gęstości prawdopodobieństwa? Gęstość jest zdefiniowana tak, że pole p powierzchni pod krzywą wykresu gęstości prawdopodobieństwa na lewo od prostej odciętej x równe jest wartości w punkcie x rozkładu prawdopodobieństwa zmiennej losowej φ. Inaczej mówiąc:

p = prawdopodobieństwo tego, że zmienna φ przyjmie wartość mniejszą lub równą x.

Zauważ, że we wzorze gęstości prawdopodobieństwa różnica $x - \mu$ reprezentuje odchylenie wartości zmiennej od średniej. Odchylenie to ma być wyrażone w jednostkach δ, czyli równych odchyleniu standardowemu, stąd we wzorze jest wyrażenie $(x - \mu)/\delta$.

Przykład 4. Przypuśćmy, że wartość średnia μ pewnej zmiennej losowej to 10, a odchylenie standardowe δ to 2. Odchylenie wartości $x = 5$ od średniej to $x - \mu = -5$. Wynik jest ujemny, bowiem odchylenie wartości x jest w kierunku mniejszych od średniej wartości. Jak wartość –5 wyraża się w jednostkach $\delta = 2$? Wartość –5 =

$(x - \mu)/\delta$ odchyleń standardowych $\delta = (5 - 10)/2$ odchyleń standardowych $\delta = -2.5$ *2. Standaryzowana, czyli liczona w jednostkach δ wartość –5 to –2.5.
Jeżeli dokonamy zmiany współrzędnych:

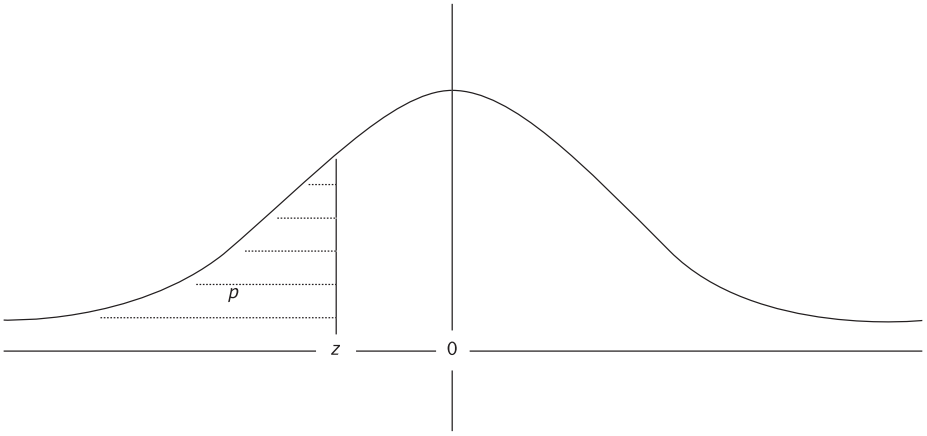
$x' = (x - \mu)/\delta$ – standaryzowana wartość x,

to wzór na gęstość prawdopodobieństwa przyjmuje postać standardową:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

czyli rozkład normalny, gdy średnia $\mu = 0$ i odchylenie standardowe $\delta = 1$.

Wykres standaryzowanego rozkładu normalnego i interpretacja powierzchni pod krzywą



Rysunek 2.
Wykres gęstości standardowego rozkładu normalnego

Na rys. 2, cała powierzchnia pod krzywą to 1, czyli 100%. Powierzchnia p na lewo od z jest równa prawdopodobieństwu p, że zmienna losowa o rozkładzie standardowym normalnym przyjmie wartość $\leq z$. Wartości powierzchni p na lewo od zadanego z są do odczytania z tablic wartości standardowego rozkładu normalnego. Można je także wyliczać w arkuszu Excel stosując funkcję statystyczną arkusza, obliczającą pole powierzchni na lewo od zadanego z pod krzywą gęstości rozkładu normalnego.

Jakie znaczenie ma powierzchnia p na lewo od z w kontekście naszej próby losowej stu odległości (i równie dobrze każdej innej próby losowej wartości zmiennej)? Powierzchnia p na lewo od z to w przybliżeniu procent elementów losowych obserwacji dających wartość zmiennej losowej mniejszej lub równej z, czyli procent obserwacji o wartości mniejszej lub równej z.

Mamy zatem

$p*$liczba_obserwacji = teoretyczna oczekiwana liczba obserwacji
o wartości $\leq z$ dla losowej populacji obserwacji
o zadanej liczności równej liczba_obserwacji.

Powyższa własność będzie użyta w ocenie wiarygodności hipotezy, że nasze dane stu odległości pasują do rozkładu normalnego.

Przykład 5. Wyliczyliśmy z naszej próby losowej odległości, że $\underline{x} = 36.56$ a $s = 19.25$. Przypominamy, że $\underline{x}$ oznacza średnią wartość naszej próby stu odległości, s jest wyliczonym z tej próby przybliżeniem odchylenia standardowego naszej zmiennej ‘odległość’.

Standaryzowana wartość prawego krańca przedziału [0, 7) to (7 – 36.56)/19.25 = –1.536. Odczytana z tablicy wartość powierzchni standardowego rozkładu normalnego na lewo od $z = –1.536$ jest równa $p = 0.0623$.

p*liczba_obserwacji = 6.23%∗100 = 6.23

jest to teoretyczna, oczekiwana liczba odległości ≤ 7 na 100 obserwacji, o ile odległości faktycznie mają rozkład normalny o średniej bliskiej 36.56 i odchyleniu standardowym bliskim 19.25.

Zadanie 9. Znajdź w pomocy arkusza Excel wśród funkcji statystycznych funkcję służącą do obliczania wartości standardowego rozkładu normalnego i oblicz stosowną wartość dla $z = –1.536$.

Jeżeli przyjmiemy hipotezę, że nasze sto danych odległości pasują do rozkładu normalnego, to jesteśmy już bardzo blisko rozwiązania naszego problemu wiarygodnej oceny procentu odległości od miejsca zamówienia do zajezdni, większych od lub równych 40 kilometrów.

Standaryzowana wartość 40 jest równa (40 – 36.56)/19.25 = 0.1787. Pole na lewo od wartości 0.1787 standaryzowanego rozkładu normalnego, odczytane z tablicy, jest równe 0.5709. Pole na prawo od wartości 0.1787 standaryzowanego rozkładu normalnego wynosi więc 1 – 0.5709 = 0.4291, czyli 42.9% stanowi oczekiwany procent zamówień z odległości powyżej 40 kilometrów.

Zadanie 10. Powtórz cały proces obliczania prognozy dla drugiego zestawu stu odległości.

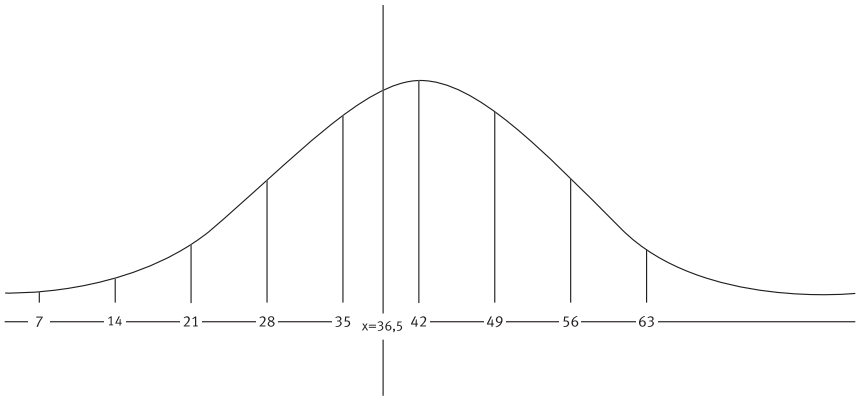
Zapewne otrzymałeś nieco inną wartość, co bierze się stąd, że średnia drugiej próby i obliczone z drugiej próby przybliżenie odchylenia standardowego są nieco inne niż dla pierwszego zestawu danych. A więc standaryzowana wartość 40-stu też będzie nieco inna. Jednak różnica tych prognoz to o rząd wielkości mniej niż 7% różnicy naiwnych prognoz wyliczonych z tych dwóch prób!

2.4 OCENA WIARYGODNOŚCI HIPOTEZY

Zajmiemy się tutaj zagadnieniem oceny wiarygodności hipotezy, że nasze sto danych odległości pasują do rozkładu normalnego, posługując się miarą χ-kwadrat dopasowania danych do rozkładu.

Drogi uczniu, teraz zaczynają się „schody". Powiemy na pocieszenie, że znajomość zagadnień prezentowanych w tej części nie jest konieczna do sformułowania naszej prognozy, która już została sformułowana, jak i do zrozumienia części trzeciej dotyczącej *Drzew Decyzyjnych*. Poniższe intuicyjne rozważania są dla tych, którzy chcieli by zobaczyć, jak nasze wzrokowe wrażenie, że kształt słupków histogramu pasuje do krzywej gęstości rozkładu normalnego, można mniej więcej sformalizować. Zrozumienie dokładnej formalizacji tego zagadnienia wymaga głębszych studiów statystyki.

Wykres gęstości rozkładu normalnego zmiennej losowej o rozkładzie normalnym z wartością średnią bliską 36.56 i odchyleniem standardowym 19.25 ma kształt mniej więcej taki, jak na rys. 3.



Rysunek 3.
Wykres gęstości zmiennej losowej o rozkładzie normalnym z wartością średnią bliską 36.56 i odchyleniem standardowym 19.25

Niech *pp* będzie polem paska wyznaczonego przez wartości $y \le z$, gdzie y, z są krańcami kolejnego przedziału. Przykładowo, $y = 28, z = 35$. Pole *pp* paska obliczamy następująco. Najpierw obliczamy standaryzowane wartości *y'*, *z'* wartości *y* i *z*. Następnie obliczamy różnicę wartości standaryzowanego rozkładu normalnego dla wartości *y'* i *z'*. Otrzymana różnica to pole *pp* paska. Wynika to z interpretacji pola pod krzywą gęstości rozkładu normalnego na lewo od prostej odciętej *x*.

*pp**liczba_obserwacji = *E* = teoretyczna liczba obserwacji o wartości pomiędzy *y* i *z*.

O = obserwowana w próbie liczba obserwacji o wartościach pomiędzy *y* i *z*.

Wartość *O* to wartość histogramu dla przedziału [*y*, *z*).

Zadanie 11. Zakładając, że nasze dane pasują do rozkładu normalnego oblicz, jakie jest teoretyczne prawdopodobieństwo tego, że zmienna losowa *odległość* przyjmie wartość większą lub równą 28 i mniejszą od 35? Jaka jest teoretyczna liczba obserwacji o wartościach pomiędzy 28 i 35?

Zadanie 12. Wypełnij w arkuszu Excel tabelę 5.

Tabela 5.
Wartości oczekiwane i obserwowane

kr. przedz	std. wart.	std. norm.	pasek	*E*	*O*
7					
14					
21					
28					
35					
42					
49					
56					
63					
999					

W tabeli 5, pierwsza kolumna zawiera prawe krańce kolejnych przedziałów histogramu. Druga kolumna ma zawierać standaryzowane wartości prawych krańców kolejnych przedziałów histogramu. Trzecia kolumna ma zawierać pole powierzchni pod krzywą gęstości standardowego rozkładu normalnego na lewo od standaryzowanego krańca kolejnego przedziału histogramu. Czwarta kolumna ma zawierać pole powierzchni pod krzywą gęstości standardowego rozkładu normalnego między standaryzowanymi wartościami krańców kolejnego przedziału histogramu, czyli równe różnicy bieżącej i poprzedniej wartości w kolumnie trzeciej. Piąta kolumna ma zawierać teoretyczne oczekiwane liczby obserwacji o wartościach z kolejnego przedziału histogramu, czyli liczby równe odpowiadającej wartość w kolumnie czwartej pomnożonej przez liczbę obserwacji (czyli przez 100). Szósta kolumna ma zawierać liczby obserwacji Próby 1 o wartościach z kolejnego przedziału histogramu.

Powinieneś otrzymać tabelę z wartościami jak w tab. 6.

W ostatnim wierszu, dla krańca przedziału 999 przez pasek rozumiemy pole powierzchni na prawo od standaryzowanej wartości 63. Ta powierzchnia równa się 1 minus pole powierzchni na lewo od standaryzowanej wartości 63, czyli wynosi 1 – 0.9147 = 0.0853.

Tabela 6.
Wartości oczekiwane i obserwowane

kr. przedz	std. wart.	std. norm.	pasek	*E*	*O*
7	–1,535584	0,0621	0,0621	6,21	6
14	–1,171948	0,121	0,0589	5,89	7
21	–0,808312	0,209	0,088	8,8	8
28	–0,444675	0,33	0,121	12,1	11
35	–0,081039	0,4681	0,1381	13,81	13
42	0,2825974	0,6103	0,1422	14,22	15
49	0,6462338	0,7422	0,1319	13,19	14
56	1,0098701	0,8438	0,1016	10,16	11
63	1,3735065	0,9147	0,0709	7,09	7
999			0,0853	8,53	9

Miara χ-kwadrat dopasowania danych do rozkładu
Następujący wzór

$$\chi^2 = \sum_{O,\,E \text{ dla kolejnych przedziałów histogramu}} (O - E)^2 / E$$

omówimy na poziomie intuicji, nie dotykając jego podstaw matematycznych.

Spróbuj odpowiedzieć na pytanie, czy z punktu widzenia dobrego dopasowania danych do rozkładu normalnego lepiej jest, gdy wartość χ^2 jest duża, czy gdy jest mała. Jeżeli wartość χ^2 jest mała, to składniki sumy są małe, bowiem składniki sumy są dodatnie. Jeżeli składniki sumy są małe, to wartości *O, E* są bliskie sobie. Jeżeli wartości *O, E* są bliskie sobie, to dane pasują do rozkładu normalnego – bowiem wartości obserwowane *O* mało się różnią od teoretycznych wartości *E*, które powinny być obserwowane, jeżeli rozkład prawdopodobieństwa zmiennej jest dokładnie rozkładem normalnym. Zatem odpowiedź na nasze pytanie brzmi: jest lepiej, gdy wartość χ^2 jest mała.

Kiedy możemy uznać, że wartość χ^2 jest „mała", czyli, że dane pasują do rozkładu normalnego? Jest to trudne zagadnienie i omówimy je tylko z grubsza. Z podstaw matematycznych omawianego wzoru na χ^2 wyliczono pewną prostokątną tablicę wartości χ^2. Numer rzędu tej tablicy odpowiada tzw. **poziomowi ufności**, numer kolumny odpowiada tzw. **stopniowi swobody** procesu liczenia wartości χ^2. Te trudne pojęcia omówimy tylko na poziomie intuicji. Jeżeli określimy jakoś liczbę stopni swobody sumy χ^2 i przyjmiemy pewien poziom ufności, to możemy odczytać z tablicy pewną wartość chi. Jeżeli wyliczona z próby wartość χ^2 jest mniejsza od tablicowej wartości chi, to uznajemy, że wartość χ^2 jest „mała" i że dane próby pasują do rozkładu normalnego.

Poziom ufności
Jeżeli zdecydowanie nie chcemy rozprawiać o „kwantylach rzędu α rozkładu χ^2 z *n* stopniami swobody", to zostaje nam jedna z popularnych i bardzo nieprecyzyjnych interpretacji poziomu ufności:

> **poziom ufności** α oznacza, że ryzyko przyjęcia błędnego wniosku,
> że dane pasują do rozkładu normalnego, jest mniejsze niż α.

To samo powiedziane inaczej: wniosek, że dane pasują do rozkładu normalnego, jest pewny w stopniu co najmniej $1 - \alpha$. Co to naprawdę oznacza, możesz dowiedzieć się głębiej studiując podstawy statystyki matematycznej, między innymi owe kwantyle rozkładu χ^2.

Liczba stopni swobody
Jest to jedna z najtrudniejszych do intuicyjnego wytłumaczenie koncepcji w statystyce. Spróbujemy objaśnić ją na prostym przykładzie. Rozważmy wzór

$$a = \sum_{1 \le i \le 4} x_i$$

Załóżmy, że zostało ustalone, że $a = 10$. Ile jest stopni swobody omawianego wzoru? No jasne, że 3. Jeżeli przyjmiemy dowolne trzy wartości zmiennych x_1, x_2, x_3, x_4 to wartość czwartej z nich jest już wyznaczona, bowiem wartości wszystkich tych zmiennych sumują się do $a = 10$.

Podobnie, jak w przypadku pojęcia poziomu ufności, określimy tylko intuicyjnie liczbę stopni swobody naszego wzoru

$$\chi^2 = \sum_{O, E \text{ dla kolejnych przedziałów histogramu}} (O - E)^2 / E$$

W sumie χ^2 jest 10 składników. Wszystkie dziesięć wartości obserwowanych O sumują się do 100, tak samo jak wartości oczekiwane E. Ile jest zatem swobodnych wartości $O - E$? Tylko 9, bowiem wszystkie 10 różnic $O - E$ sumują się do zera. Z naszych 100-tu danych wyliczyliśmy dwa parametry $\underline{x}$ oraz s, które zostały użyte (poprzez standaryzację krańców przedziałów histogramu) do wyliczenia kolejnych wartości E. „Swoboda" wyliczania E jest pomniejszona o 2. Przyjęta przez nas liczba stopni swobody to $10 - 1 - 2 = 7$. Nie jest to jasne? Drogi uczniu, nie przejmuj się. Jesteś w dobrym towarzystwie wielu specjalistów, którym zdarzało się źle wyliczać liczbę stopni swobody rozważanych wzorów.

Wróćmy do naszego przypadku zestawu stu danych. Z wyliczonej tab. 6 odczytujemy kolejne wartości O, E i liczymy wartość χ^2.

$\chi^2 = \sum (O - E)^2/E = (6 - 6{,}21)^2/6{,}21 + (7 - 5{,}89)^2 / 5{,}89 + \ldots + (9 - 8{,}53)^2/8{,}53$
$\chi^2 = 0{,}625538.$

Zadanie 13. Oblicz w arkuszu Excel wartość χ^2 według powyższego wzoru, wykorzystując wcześniej obliczone wartości O i E w tablicy 6.

Liczba stopni swobody wynosi $10 - 2 - 1$, czyli jest równa 7, a poziom istotności wynosi 0,05, czyli 5%. Tablicowa wartość chi dla siedmiu stopni swobody i poziomu istotności 5% wynosi 14,06713, czyli dużo więcej niż obliczona z populacji wartość statystyki χ^2. Wnioskujemy, że nasze dane dobrze pasują do rozkładu normalnego. Uznajemy, że nasza prognoza 42.9% jest wiarygodna.

Zadanie 14. Oblicz miarę χ^2 dla drugiego zestawu stu danych.

3 DRZEWA DECYZYJNE

3.1 ZASTOSOWANIE DRZEWA DECYZYJNEGO DO KLASYFIKACJI DANYCH

Indukcja drzew decyzyjnych jest jedną z klasycznych metod, stosowanych w systemach uczących się. Najpierw przedstawimy podstawowe pojęcia na poziomie intuicyjnym.

Załóżmy, że stany pogody opisane w tabeli zostały poklasyfikowane przez eksperta do dwóch kategorii: 0 (nie nadaje się do gry w golfa) i 1 (nadaje się do gry w golfa). Tabela ta jest przykładem **zbioru treningowego**. Zadaniem algorytmu uczenia się jest uogólnić klasyfikację podaną w zbiorze treningowym na wszystkie pozostałe obiekty (tutaj na wszystkie pozostałe stany pogody, w szczególności na stan w wierszu 15).

Zmierzamy do przedstawienia idei klasycznego algorytmu uczenia się, **indukcji drzew decyzyjnych**. Załóżmy, że mamy do dyspozycji pewien zbiór S testów. Dla ustalenia uwagi przyjmijmy, że do zbioru S należą tylko testy równościowe, np.

Przykład 6. Tabela stanów pogody
Tabela 7.
Stany pogody

x	Aura	Temperatura	wilgotność	wiatr	klasyfikacja
1	słoneczna	ciepła	duża	słaby	0
2	słoneczna	ciepła	duża	silny	0
3	pochmurna	ciepła	duża	słaby	1
4	deszczowa	umiarkowana	duża	słaby	1
5	deszczowa	zimna	normalna	słaby	1
6	deszczowa	zimna	normalna	silny	0
7	pochmurna	zimna	normalna	silny	1
8	słoneczna	umiarkowana	duża	słaby	0
9	słoneczna	zimna	normalna	słaby	1
10	deszczowa	umiarkowana	normalna	słaby	1
11	słoneczna	umiarkowana	normalna	silny	1
12	pochmurna	umiarkowana	duża	silny	1
13	pochmurna	ciepła	normalna	słaby	1
14	deszczowa	umiarkowana	duża	silny	0
15	deszczowa	ciepła	duża	słaby	?

$$t_{\text{aura}}(x) = \begin{cases} \text{słoneczna}, & \text{gdy atrybutem aura obiektu } x \text{ jest słoneczna} \\ \text{pochmurna}, & \text{gdy atrybutem aura obiektu } x \text{ jest pochmurna} \\ \text{deszczowa}, & \text{gdy atrybutem aura obiektu } x \text{ jest deszczowa} \end{cases}$$

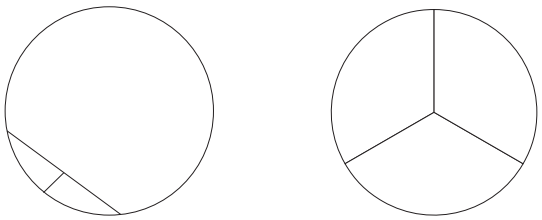
Ile testów równościowych możemy przyjąć dla naszego zbioru treningowego stanów pogody? Oczywiście cztery testy t_{aura}, $t_{\text{temperatura}}$, $t_{\text{wilgotność}}$ i t_{wiatr}. Każdy test generuje pewien podział zbioru treningowego. Na przykład test t_{aura} dzieli nasz zbiór treningowy na trzy klasy obiektów, których atrybutem aura jest odpowiednio słoneczna, pochmurna i deszczowa. Następnie, na każdym z tych zbiorów zadany jest podział na kategorie, zgodnie ze zbiorem treningowym. Przykładowo, dla testu t_{aura} dostajemy trzy podziały:

(a) podział obiektów *X zbioru* treningowego z atrybutem aura = słoneczna na te zakwalifikowane do kategorii 0 i na te zakwalifikowane do kategorii 1,
(b) podział obiektów *X* zbioru treningowego z atrybutem aura = pochmurna na te zakwalifikowane do kategorii 0 i na te zakwalifikowane do kategorii 1,
(c) podział obiektów *X* zbioru treningowego z atrybutem aura = deszczowa na te zakwalifikowane do kategorii 0 i na te zakwalifikowane do kategorii 1.

Zadanie 15. Podaj dokładnie podziały (a), (b) i (c).

Entropia podziału

Entropia ma być miarą nieporządku w podziale. Przykładowo, rozważmy następujące podziały P_1 i P_2 przedstawione na rys. 4.



Rysunek 4.
Dwa podziały

Jeżeli popatrzymy na podziały P_1, P_2 z punktu widzenia tego, że musimy zdecydować się na jedną z trzech (w tych podziałach) kategorii, to P_1 reprezentuje większy porządek (czyli mniejszą entropię) niż P_2. Dlaczego? Dlatego, że jeżeli mamy zdecydować się na zakwalifikowanie wszystkich elementów zbioru, na którym rozpięty jest podział P_1 do pewnej kategorii, to decydujemy się na tą wyraźnie najliczniejszą (dominującą) w podziale P_1 kategorię. W przypadku podziału P_2 nie mamy dominującej kategorii.

Jeżeli mówimy o zbiorach nieskończonych, to za dominującą kategorię uważamy tę, o znacząco większym w porównaniu z innymi kategoriami, prawdopodobieństwie wylosowania elementu z tej dominującej kategorii.

Jak mierzyć to, że w podziale pojawiają się dominujące (w sensie liczności) kategorie? Mierzy to wartość zwana **entropią** zdefiniowana podanym wzorem. Jeżeli p_i jest prawdopodobieństwem wylosowania obiektu z i-tej kategorii ($1 \le i \le n$), to wzór na **entropię** podziału na n kategorii jest następujący:

$$E = \sum_{i \text{ przebiegające kategorie}} -p_i * \log(p_i)$$

Uważa się, że im większa jest entropia, tym większy jest nieporządek. W przypadku skończonych zbiorów, na których rozpięty jest podział, prawdopodobieństwo p_i oznacza po prostu stosunek liczności zbioru obiektów kategorii i do liczności całego zbioru, na którym rozpięty jest podział. Przykładowo, jeżeli na zbiorze X zadany jest podział na zbiory $A_1, A_2, \ldots, A_n$ to wzór na entropię przybiera następującą postać:

$$E = \sum_{1 \le i \le n} -(|A_i|/|X|) * \log(|A_i|/|X|)$$

Zadanie 16. Dla $n = 2$ i dla $n = 3$ przygotuj w arkuszu Excel tabelę do obliczania entropii n sumujących się do jedynki prawdopodobieństw. Ambitniejsi uczniowie mogą napisać program który dla podanej liczby n i podanych n prawdopodobieństw $p_1, \ldots, p_n$ (sumujących się do jedynki) obliczy entropię tych prawdopodobieństw wedle powyższego wzoru.

Entropia testu względem zbioru treningowego

Jak już wspomnieliśmy, każdy test t generuje podział zbioru treningowego T na zbiory, powiedzmy, $T_1, \ldots, T_m$, gdzie m jest liczbą możliwych wartości testu. Każdy zbiór T_i dziedziczy ze zbioru treningowego T podział na kategorie. Entropia testu to ważona suma entropii podziałów zbiorów T_i na kategorie. W ważonej sumie entropii tych podziałów bardziej mają ważyć entropie podziałów rozpiętych na liczniejszych zbiorach T_i. Na przykład, entropią testu t_{aura} względem zbioru treningowego zadanego przez tabelę 7 stanów pogody jest:

entropia podziału (a)*stosunek liczności zbioru na którym rozpięty jest podział (a) do liczności całego zbioru treningowego +

entropia podziału (b)*stosunek liczności zbioru na którym rozpięty jest podział (b) do liczności całego zbioru treningowego +

entropia podziału (c)*stosunek liczności zbioru na którym rozpięty jest podział (c) do liczności całego zbioru treningowego.

Konkretnie, test t_{aura} dzieli zbiór treningowy stanów pogody {1, 2,..., 14} na następujące zbiory:

$T_1 = \{1, 2, 8, 9, 11\}$ – stany z atrybutem aura = słoneczna,

$T_2 = \{3, 7, 12, 13\}$ – stany z atrybutem aura = pochmurna,

$T_3 = \{4, 5, 6, 10, 14\}$ – stany z atrybutem aura = deszczowa.

Na każdym z tych zbiorów mamy następujące podziały na kategorie (najpierw stany z kategorią 0, potem stany z kategorią 1):

{{1, 2, 8}, {9, 11}} – załóżmy, że entropią tego podziału jest liczba e_1,

{Ø, {3, 7, 12, 13}} – tylko obiekty z kategorią 1, zbiór obiektów ze zbioru {3, 7, 12, 13} z kategorią 0 jest pusty. Entropią podziału na jeden zbiór jest $e_2 = 0$.

{{6, 14}, {4, 5, 10}} – załóżmy, że entropią tego podziału jest liczba e_3.

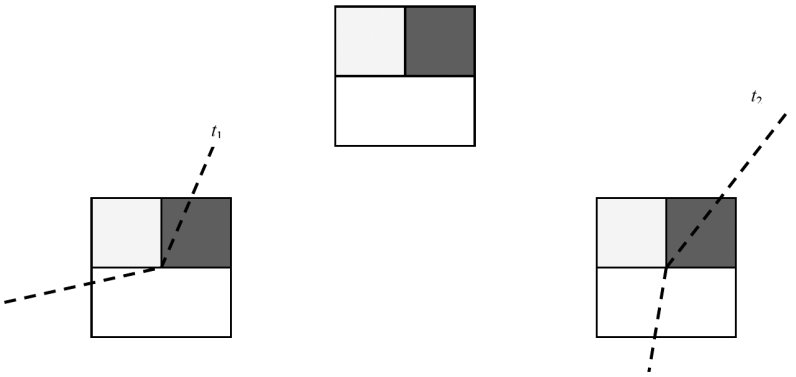
Entropia testu t_{aura} względem zbioru treningowego stanów pogody jest równa

$$(5/14)*e_1 + (4/14)*e_2 + (5/14)*e_3$$

Drogi uczniu, jeżeli zastosujesz przygotowane tabele bądź program z zadania 16 do obliczenia entropii e_1 i e_3, to możesz wyliczyć konkretną wartość liczbową entropii testu t_{aura} względem zbioru treningowego stanów pogody.

Następujący przykład ma zobrazować graficznie pojęcie entropii testu względem zbioru treningowego i ma pomóc w dobrym „wyczuciu" tego pojęcia.

Przykład 7. Przedstawienie graficzne intuicji entropii testu względem zbioru treningowego.



Rysunek 5.
Zobrazowanie intuicji entropii testu

Kwadrat na rys.5 reprezentuje pewien zbiór treningowy, z elementami kategorii żółty, biały, czerwony. Pola ilustrują liczność (bądź prawdopodobieństwo) poszczególnych kategorii. Test t_1 dzieli zbiór treningowy na dwa zbiory z dominującymi kategoriami (żółta w jednym zbiorze podziału, biała w drugim), test t_2 nie. Entropia testu t_1 jest zapewne znacząco mniejsza niż entropia testu t_2.

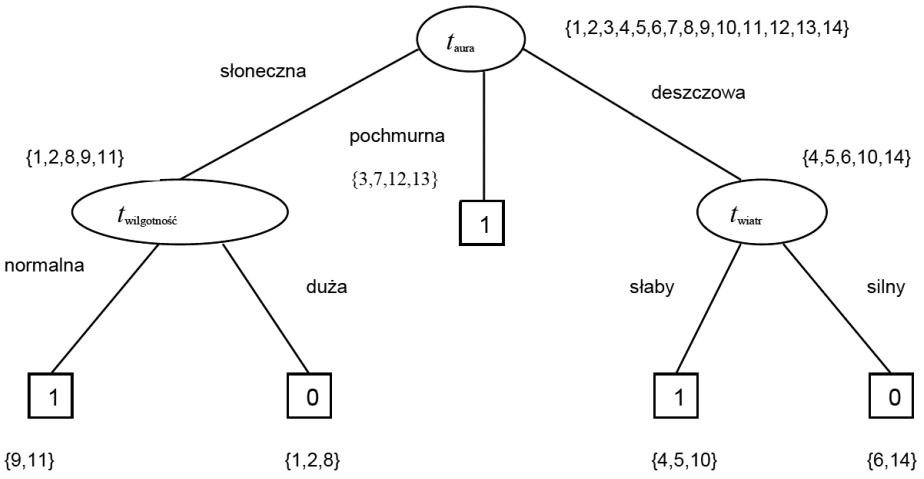
Zadanie 17. Wylicz w arkuszu Excel entropię testu $t_{temperatura}$ względem zbioru treningowego stanów pogody. Wykorzystaj narzędzia przygotowane w zadaniu 16.

Nie będziemy w tym miejscu przytaczać formalnej definicji entropii dowolnego testu. Ważnym jest widzieć, że zbiór treningowy rozpada się na kolejne mniejsze zbiory treningowe przyporządkowane „odnóżkom" testu. W naszym przykładzie te kolejne zbiory treningowe to podziały (a), (b), (c).

Przykład drzewa decyzyjnego dla zbioru treningowego stanów pogody

W każdym wierzchołku drzewa na rys. 6 jest wpisany jeden z dostępnych testów. Wierzchołek ma tyle następników, ile jest możliwych wartości testu wpisanego w ten wierzchołek. Przy korzeniu jest wypisany początkowy zbiór treningowy. Przy kolejnych wierzchołkach są wypisane kolejne mniejsze zbiory treningowe, na które jest dzielony bieżący zbiór treningowy przez test zapisany w wierzchołku. W liściach drzewa są wpisane kategorie klasyfikacji stanów pogody. Drzewo to klasyfikuje stan 15 w tabeli do kategorii 1. Obliczenie przebiega następująco. Wartością testu t_{aura} dla stanu 15 jest deszczowa. Zatem kolejnym testem, który zastosujemy do stanu 15 jest t_{wiatr}. Wartością testu t_{wiatr} dla stanu 15 jest słaby. Dochodzimy w drzewie do liścia z kategorią 1 i do tej kategorii jest zaklasyfikowany stan 15.

Zadanie 18. Czy jest sensowne, by na jednej ścieżce od korzenia do liścia pewien test wystąpił dwukrotnie?



Rysunek 6
Przykład drzewa decyzyjnego

Idea algorytmu indukcji drzew decyzyjnych

Algorytm indukcji drzew decyzyjnych jest **algorytmem rekurencyjnym**. Drogi uczniu, możesz dowiedzieć się, co to jest algorytm rekurencyjny uczęszczając pilnie na zajęcia z algorytmiki. Tutaj powiemy tylko tyle, że algorytmy rekurencyjne przetwarzają daną wejściową *d* wedle następującego schematu:

1. **jeżeli** dana *d* jest *prosta* (tzw. dno rekursji) **to** algorytm od razu oblicza wynik.
2. **jeżeli** dana *d* jest *złożona*, **to**
   2.1 algorytm oblicza dekompozycję danej *d* na *prostsze* dane d_1,..., d_n.
   2.2 odwołując się do siebie samego algorytm oblicza rozwiązania r_1, r_2,..., r_n dla *prostszych* danych d_1,..., d_n (wywołania rekurencyjne).
   2.3 z częściowych rozwiązań r_1, r_2,..., r_n algorytm oblicza rozwiązanie *r* dla danej *d*.

Algorytm indukcji drzew decyzyjnych rekurencyjnie tworzy kolejne wierzchołki drzewa decyzyjnego. Do każdego kolejnego poziomu rekursji są przekazywane następujące dane:

- bieżący zbiór treningowy *T*,
- bieżący zbiór *S* dostępnych testów,
- domniemana kategoria *k*.

Funkcja buduj(*T, S, k*) zwraca korzeń drzewa decyzyjnego zbudowanego dla zbioru treningowego *T,* zbioru dostępnych testów *S* i domniemanej kategorii *k*. Zdefiniowanie należytego zbioru dostępnych testów jest na ogół bardzo subtelnym zagadnieniem. Nie ma tu ogólnych reguł – po prostu należy dobrze wykorzystywać specyfikę konkretnego analizowanego przypadku.

algorytm buduj(*T, S, k*):
jeżeli *T* jest pusty **to** zwróć liść z wpisaną kategorią domniemaną *k //dno rekursji*
w przeciwnym przypadku
jeżeli w *T* jest tylko jedna kategoria **to** zwróć liść z wpisaną tą jedyną w *T* kategorią *//dno rekursji*
w przeciwnym przypadku
jeżeli *S* jest pusty **to** zwróć liść z wpisaną tą kategorią, która jest najliczniejsza w zbiorze *T //dno rekursji*
w przeciwnym przypadku *// zbiory T i S są niepuste*
{ 1. utwórz kolejny węzeł *n*;
2. ze zbioru *S* wybierz, wedle przyjętego kryterium wyboru testu, test *t* i wpisz go do utworzonego węzła *n*;
3. jako *k* przyjmij najliczniejszą w *T* kategorię ;
4. oblicz zbiory treningowe T_1,..., T_m, na które test *t* dzieli zbiór *T,* gdzie *m* jest liczbą możliwych wartości testu *t*;
5. *// budowanie następników węzła n*
dla wszystkich *i* = 1,..., *m* **wykonaj**

i-ty następnik węzła n := buduj(T_i, $S - \{t\}$, k); *// wołania rekurencyjne*
6. zwróć węzeł n jako wynik funkcji;
};

Sercem algorytmu indukcji drzew decyzyjnych jest kryterium wyboru testu dla kolejnych wierzchołków budowanego drzewa decyzyjnego. Klasyczny algorytm przez entropię dla każdego generowanego wierzchołka konstruowanego drzewa decyzyjnego wybiera test o minimalnej entropii liczonej względem kolejnego zbioru treningowego pojawiającego się przy tym wierzchołku. Zatem wybieramy test, który dzieli bieżący zbiór treningowy na zbiory, w których pojawiają się dominujące kategorie. To, w jakim stopniu w zbiorach podziału wyznaczonego przez test pojawiają się dominujące kategorie, mierzy entropia testu. Oprócz wyboru testu przez entropię stosowanych jest także wiele innych kryteriów wyboru testu.

Zmierzamy do pokazania, że przykładowe drzewo z rys. 6 jest wynikiem działania algorytmu indukcji drzewa decyzyjnego, zastosowanego do przykładowej tab. 7, zawierającej stany pogody, oraz zbioru S testów równościowych t_{aura}, $t_{temperatura}$, $t_{wilgotność}$, t_{wiatr}. Obliczenia wykonane za pomocą algorytmu buduj dla naszego zbioru treningowego stanów pogody przebiegają następująco.

1. Dla początkowego zbioru treningowego {1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14} oblicz entropie dostępnych testów t_{aura}, $t_{temperatura}$, $t_{wilgotność}$, t_{wiatr}. Okazuje się, że testem o minimalnej entropii jest t_{aura}. Do początkowego wierzchołka (korzenia) drzewa wstawiany jest test t_{aura}.
2. Test t_{aura} dzieli początkowy zbiór treningowy na zbiory $T_1 = \{1, 2, 8, 9, 11\}$, $T_2 = \{3, 7, 12, 13\}$, $T_3 = \{4, 5, 6, 10, 14\}$. Zbiór T_2 (wartość testu pochmurna) zawiera elementy tylko kategorii 1, więc jako drugi następnik korzenia jest generowany liść z tą kategorią,
3. (Wywołanie rekurencyjne) Dla zbioru treningowego T_1 oblicz entropie dostępnych testów $t_{temperatura}$, $t_{wilgotność}$, t_{wiatr}. Okazuje się, że testem o minimalnej entropii względem zbioru T_1 jest $t_{wilgotność}$. Do generowanego pierwszego następnika korzenia wstawiamy test $t_{wilgotność}$.
4. Test $t_{wilgotność}$ dzieli zbiór T_1 na zbiory $T_{11} = \{9, 11\}$, $T_{12} = \{1, 2, 8\}$. Zbiór T_{11} zawiera tylko elementy kategorii 1, zbiór T_{12} zawiera tylko elementy kategorii 0. Generowane są dwa liście jako następniki bieżącego węzła, ze stosownymi kategoriami.
5. (Wywołanie rekurencyjne) Dla zbioru treningowego T_3 oblicz entropie dostępnych testów $t_{temperatura}$, $t_{wilgotność}$, t_{wiatr}.. Okazuje się, że testem o minimalnej entropii względem zbioru T_3 jest t_{wiatr}. Do generowanego trzeciego następnika korzenia wstawiamy test t_{wiatr}.
6. Test t_{wiatr} dzieli zbiór T_3 na zbiory $T_{31} = \{4, 5, 10\}$, $T_{32} = \{6, 14\}$. Zbiór T_{31} zawiera tylko elementy kategorii 1, zbiór T_{32} zawiera tylko elementy kategorii 0. Generowane są dwa liście jako następniki bieżącego węzła (trzeciego następnika korzenia), ze stosownymi kategoriami.

Zadanie 19. Napisz program, który na wejściu dostaje zbiór treningowy i podział zbioru treningowego jako wynik testu na tym zbiorze. Na wyjściu obliczana jest entropia takiego testu względem podanego zbioru treningowego. Zastosuj swój program do wyliczenia drzewa decyzyjnego ze zbioru treningowego stanów pogody.

Wskazówka. Zadeklaruj dużą tablicę dwuwymiarową o wartościach typu rzeczywistego do zapisywania podawanego na wejściu zbioru treningowego. Wartości atrybutów tzw. nominalne (np. pochmurna, słoneczna, deszczowa) koduj liczbowo. Podzbiory zbioru treningowego możesz reprezentować jako zbiory numerów wierszy twojej dwuwymiarowej tablicy. Na zajęciach z algorytmiki zapewne poznasz różne sposoby reprezentowania zbioru liczb w programie, np. listy. Podział zbioru treningowego możesz reprezentować jako tablicę jednowymiarową list reprezentujących zbiory numerów wierszy tablicy dwuwymiarowej reprezentującej zbiór treningowy.

3.2 BŁĄD KLASYFIKATORA I WALIDACJA KRZYŻOWA

Wyobraźmy sobie, że dla jednego dużego i złożonego zbioru treningowego zostały wyliczone na podstawie różnych metod dwa różne drzewa decyzyjne. Należy zdecydować, które z nich wybrać. Na podstawie jakie-

go kryterium dokonamy wyboru drzewa? Są metody szacowania prawdopodobieństwa błędnej klasyfikacji przez drzewo. Oczywiście wybierzemy to drzewo, którego oszacowane prawdopodobieństwo błędnej klasyfikacji będzie mniejsze.

W tej części wprowadzimy konieczne definicje prowadzące do podstawowych pojęć związanych z błędem klasyfikatora. Omówimy elementarne podstawy metody zwanej **walidacją krzyżową** oceniania błędu klasyfikatora. Pojęcie **błędu klasyfikatora** jest istotne dla zrozumienia ważnego zjawiska zwanego **nadmiernym dopasowaniem**. Przy okazji, podamy formalną definicję entropii testu. Będzie to potrzebne tym, którzy zdecydują się zaprogramować w pełnej ogólności algorytm indukcji drzewa decyzyjnego.

Będziemy mówić o zbiorze X wszystkich możliwych przykładów (które mogą być klasyfikowane), o pojęciu w zbiorze przykładów X i kategoriach tego pojęcia, o zbiorze H hipotez (albo inaczej klasyfikatorów), o błędzie hipotezy względem pojęcia. Omówimy intuicyjnie te pojęcia na przykładach, potem podamy formalne definicje, wyjąwszy przypadek błędu hipotezy względem pojęcia. Dla nieskończonych zbiorów przykładów X błąd hipotezy względem pojęcia omówimy tylko intuicyjnie.

Przykład 8. Wróćmy do naszego przykładu 6 stanów pogody. Są trzy możliwe wartości atrybutu aura, trzy możliwe wartości atrybutu temperatura, dwie możliwe wartości atrybutu wilgotność i dwie możliwe wartości atrybutu wiatr. Zbiór X wszystkich możliwych stanów pogody ma 3*3*2*2 = 36 elementów. W zbiorze X stanów pogody rozważane jest pojęcie ocena-pogody, które dzieli zbiór X na dwie kategorie tego pojęcia: 0 – nie nadaje się do gry w golfa i 1 – nadaje się do gry w golfa. Pojęcie ocena-pogody tak naprawdę jest funkcją

$$\text{ocena-pogody}: X \to \{0, 1\}$$

Nie mamy definicji pojęcia ocena-pogody – tylko „kawałek" tego pojęcia zapisany jest jako zbiór treningowy dany tabelą 7. Nie mniej uważamy, że jest jakaś funkcja ocena-pogody dobrze oddająca intuicję dobrej (czy złej) pogody do gry w golfa. Możemy wybrać jakąś funkcję

$$h: X \to \{0, 1\}$$

zdefiniowaną przez pewne drzewo decyzyjne, w wierzchołkach którego mogą pojawić się tylko testy t_{aura}, $t_{\text{temperatura}}$, $t_{\text{wilgotność}}$, t_{wiatr}, jako hipotezę dla pojęcia ocena-pogody. Błąd hipotezy h względem pojęcia ocena-pogody jest zdefiniowany następująco:

$$\text{err}(h, \text{ocena-pogody}) = |\{x \in X \mid h(x) \neq \text{ocena-pogody}(x)\}| / |X|$$

czyli stosunek liczby przykładów błędnie klasyfikowanych przez hipotezę h do liczby wszystkich możliwych przykładów. Jako zbiór H możliwych hipotez przyjmujemy zbiór funkcji definiowalnych przez pewne drzewo decyzyjne w którego wierzchołkach mogą pojawić się testy tylko ze zbioru $\{t_{\text{aura}}, t_{\text{temperatura}}, t_{\text{wilgotność}}, t_{\text{wiatr}}\}$. Na ogół tak jest, że zbiór H hipotez jest wyznaczony przez pewien język, w którym funkcje ze zbioru H mogą być zdefiniowane. W tym przypadku język drzew decyzyjnych dla zbioru dostępnych testów $\{t_{\text{aura}}, t_{\text{temperatura}}, t_{\text{wilgotność}}, t_{\text{wiatr}}\}$. Podsumowując, mamy tu następujące rzeczy:

- zbiór X przykładów – zbiór wszystkich stanów pogody,
- zbiór C kategorii pojęcia ocena-pogody, $C = \{0, 1\}$,
- pojęcie ocena-pogody: $X \to C$
- zbiór hipotez H równy zbiorowi funkcji $h: X \to C$ definiowalnych przez pewne drzewo decyzyjne dla przyjętego zbioru dostępnych testów.
- błąd hipotezy h względem pojęcia ocena-pogody.

Zadanie 20. Podaj jakiś zbiór treningowy dla pojęcia pogoda-śródziemnomorska i oblicz dla tego zbioru treningowego drzewo decyzyjne algorytmem indukcji drzewa decyzyjnego dla zbioru dostępnych testów $\{t_{\text{aura}}, t_{\text{temperatura}}, t_{\text{wilgotność}}, t_{\text{wiatr}}\}$.

Przykład 9. Każdy przykład x z rozpatrywanego zbioru X wszystkich przykładów samochodów jest opisany przez atrybuty:

klasa: atrybut o wartościach miejski, mały, kompakt, duży,
nośność: atrybut o wartościach niska, średnia, duża,
osiągi: atrybut o wartościach słabe, przeciętne, dobre,
niezawodność: atrybut o wartościach mała, przeciętna, duża.

Rozważamy pojęcie rodzaj-samochodu z kategoriami m-o – miejski osobowy, o – osobowy, d – dostawczy. Zbiór kategorii pojęcia to $C = \{\text{m-o, o, d}\}$ (tym razem trzy a nie dwie kategorie). Pojęcie rodzaj-samochodu jest funkcją

$$\text{rodzaj-samochodu: } X \to C.$$

Zbiór H hipotez to zbiór funkcji $h: X \to C$ definiowalnych przez pewne drzewo decyzyjne dla zbioru dostępnych testów równościowych $t_{\text{klasa}}, t_{\text{nośność}}, t_{\text{osiągi}}, t_{\text{niezawodność}}$. Błąd hipotezy względem pojęcia rodzaj-samochodu jest zdefiniowany analogicznie jak w przykładzie 8, bowiem zbiór X jest skończony.

Przykład 10. Jest to przykład z nieskończonym zbiorem X wszystkich przykładów. Niech X będzie zbiorem wszystkich punktów $x = (a, b)$ płaszczyzny takich, że $0 \le a \le 1$ i $0 \le b \le 1$ (czyli X jest kwadratem jednostkowym na płaszczyźnie). Zbiór X jest zbiorem przykładów. Pojęciem p jest prostokąt o punktach narożnych $x_1 = (1/\sqrt{3}, 1/\sqrt{6})$, $x_2 = (1/\sqrt{2}, 1/\sqrt{6})$, $x_3 = (1/\sqrt{2}, 1/\sqrt{8})$, $x_4 = (1/\sqrt{3}, 1/\sqrt{8})$ – wszystkie punkty narożne prostokąta p o współrzędnych niewymiernych. Pojęcie p dzieli zbiór X na dwie kategorie: 0 – punkty nie należące do prostokąta p, 1 – punkty należące do prostokąta p.

Zbiór H hipotez to zbiór prostokątów h, których punkty narożne należą do zbioru X i mają wszystkie współrzędne wymierne. Błąd hipotezy h względem pojęcia p określimy tylko intuicyjnie, bowiem zbiór X jest nieskończony.

$\text{err}(h, p)$ = prawdopodobieństwo wylosowania przykładu $x \in X$ takiego,
że $h(x) \neq p(x)$

W tym przypadku możemy przyjąć, że $\text{err}(h, p)$ = pole różnicy symetrycznej zbiorów h i p.

Ogólne definicje

Niech X będzie zbiorem przykładów a C będzie zbiorem rozpatrywanych kategorii. **Pojęciem** jest pewna funkcja

$$\begin{array}{c} c: X \to C \\ \uparrow \quad\quad \uparrow \\ \text{pojęcie} \quad \text{zbiór kategorii} \end{array}$$

Niech H będzie pewnym zbiorem funkcji: dla $h \in H$, $h: X \to C$. Funkcje te nazywamy **hipotezami** (albo klasyfikatorami). Na ogół, przestrzeń hipotez jest wyznaczona przez pewien język, w których te funkcje mogą być zdefiniowane. Na przykład, funkcje definiowalne przez drzewa decyzyjne dla ustalonego zbioru dostępnych testów. Pojęcie, dla którego chcemy znaleźć klasyfikator, na ogół nie jest dokładnie definiowalne w przyjętym języku definiowania klasyfikatorów. Innym przykładem języka definiowania klasyfikatorów są sieci neuronowe.

Błąd hipotezy (klasyfikatora) h względem pojęcia c intuicyjnie określamy następująco:

$\text{er}(h, c)$ = prawdopodobieństwo wylosowania przykładu x takiego, że $h(x) \neq c(x)$.

Jeżeli zbiór X jest skończony, to możemy formalnie zdefiniować błąd hipotezy h względem pojęcia c.

$er(h, c) = |\{x \in X \mid h(x) \neq c(x)\}|/|X|$.

Zbiorem treningowym dla pojęcia c jest pewien zbiór T przykładów wraz z przypisanymi tym przykładom, zgodnie z pojęciem c, kategoriami. O zbiorze treningowym mówimy że jest **zbiorem etykietowanych przykładów**. Dla $d \in C$ oznaczmy $T(d) = \{x \in T \mid x$ ma w T kategorię $d\}$.

Testem jest pewna funkcja $t: X \to W = \{r_1, \ldots, r_m\}$. O elementach $r_1, \ldots, r_m$ mówimy, że są to **odnóżki testu**.

Entropia testu t względem zbioru treningowego T

Dla $i = 1, \ldots, m$ definiujemy $T_i = \{x \in X \mid x \in T \wedge t(x) = r_i\}$, czyli T_i jest i-tym zbiorem podziału zbioru T przez test t. Definiujemy

$E(T_i) = \sum_{d \in C} -(|T_i(d)|/|T_i|) * \log(|T_i(d)/|T_i|)$ – entropia podziału zbioru T_i na kategorie,

$E(t, T) = \sum_{1 \leq i \leq m} (|T_i|/|T|) * E(T_i)$ – entropia testu t względem zbioru treningowego T.

Podaliśmy formalny wzór na entropię testu, bowiem to może być potrzebne tym, którzy spróbują rozwiązać postawiony przez nas problem.

Ocenianie błędu klasyfikatora – walidacja krzyżowa

Zwykle przestrzeń X wszystkich możliwych przykładów i docelowe pojęcie c nie są dokładnie zdefiniowane. Dysponujemy tylko zbiorem treningowym, w którym „odbija" się intuicja eksperta klasyfikującego dane trenujące. Aby ocenić prawdopodobieństwo dokonania błędnej klasyfikacji przez klasyfikator, postępujemy następująco. Bierzemy pewną część T' (np. jedną piątą) danego zbioru treningowego T i traktujemy tę część jako nowy zbiór treningowy. Stosujemy nasz algorytm uczenia się (np. indukcję drzewa decyzyjnego) do zbioru T' jako zbioru treningowego. Wynikowy klasyfikator stosujemy do wszystkich elementów x zbioru $T - T'$ i zliczamy, ile razy nasz klasyfikator daje inną kategorię elementu x niż ta zapisana w zbiorze T dla elementu x. Uznajemy, że ułamek

zliczona liczba błędnych klasyfikacji na elementach x z $T = T'$/(liczność zbioru $T - T'$)

przybliża prawdopodobieństwo błędnej klasyfikacji klasyfikatora obliczonego np. z całego zbioru treningowego T. Postępowanie takie nazywa się **walidacją krzyżową**.

Walidacja krzyżowa jest metodą zdrowo-rozsądkową, której zasadność jest oparta na założeniach, które bardzo trudno jest sprawdzić.

1. Należy założyć, że struktura ukryta w zbiorze treningowym T dla pojęcia c dobrze przybliża strukturę ukrytą w całej przestrzeni X i w pojęciu c. Innymi słowy, zbiór T ma być w stosownym sensie reprezentatywny dla przestrzeni X i pojęcia c.
2. Zbiór T' wybrany do walidacji krzyżowej powinien być reprezentatywny dla zbioru T – struktura ukryta w T' powinna dobrze przybliżać strukturę ukrytą w T.

Nie podano jeszcze satysfakcjonującej teorii precyzującej dokładne znaczenie założeń 1 i 2. Stosuje się różne *ad hoc* podpórki liczenia wariancji po zbiorach T' błędu klasyfikacji lub tzw. k-krotną walidację krzyżową. Cóż, czekamy aż pojawi się jakiś młody zdolny człowiek, który znajdzie właściwy aparat pojęciowy i objaśni wszystkim dokładne znaczenie założeń 1 i 2, w tym pojęcie reprezentatywności zbioru treningowego.

3.3 ZJAWISKO NADMIERNEGO DOPASOWANIA

Zagadnienie to omówimy na prostym i trochę sztucznym przykładzie. Nie ma jeszcze teorii oferującej powszechnie przyjęty aparat pojęciowy wyjaśniający tę kwestię.

Jako przestrzeń przykładów X weźmy zbiór $N \times N$ par liczb naturalnych. Niech c będzie pojęciem zdefiniowanym następująco:

$$c(n, k) = \begin{cases} 1 & \text{gdy } k = 2n \text{ lub } k = 2n+1 \\ 0 & \text{w przeciwnym przypadku.} \end{cases}$$

Rozważamy zbiór treningowy *T* dla pojęcia *c*, przedstawiony w tab. 8.
Zauważmy, że kategorie przykładów określone w zbiorze treningowym *T* są zgodne z pojęciem *c*. Załóżmy, że językiem definiowania klasyfikatorów są drzewa decyzyjne z następującymi dostępnymi testami t_1, t_2, t_3.

$$t_1(n, k) = \begin{cases} \text{Tak,} & \text{gdy } n, k \text{ są parzyste} \\ \text{Nie,} & \text{w przeciwnym przypadku.} \end{cases}$$

$$t_2(n, k) = \begin{cases} \text{Tak,} & \text{gdy } n \text{ jest parzysta i } k = 2n \text{ lub } n \text{ jest nieparzysta} \\ \text{Nie,} & \text{w przeciwnym przypadku.} \end{cases}$$

$$t_3(n, k) = \begin{cases} \text{Tak,} & \text{gdy } n \text{ jest nieparzysta i } k = 2n + 1 \text{ lub } n \text{ jest parzysta} \\ \text{Nie,} & \text{w przeciwnym przypadku.} \end{cases}$$

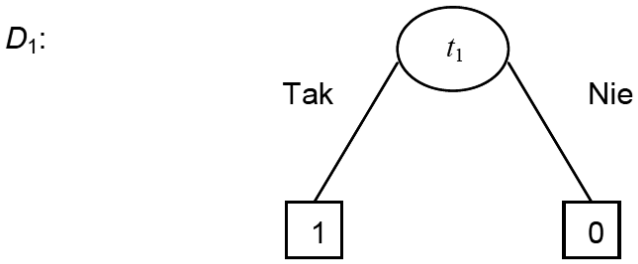
Zbiór {Tak, Nie} to zbiór możliwych wartości testów t_1, t_2, t_3.

Intuicyjnie rzecz biorąc, zbiór treningowy *T* nie jest reprezentatywny dla pojęcia *c*, bowiem nie uwzględnia on np. więcej przypadków (*n*, 2*n*+1) klasyfikowanych przez pojęcie *c* do kategorii 1. Algorytm indukcji drzew decyzyjnych z kryterium wyboru testu przez entropię wyznaczy drzewo D_1 pokazane na rys. 7.

Tabela 8.
Zbiór treningowy dla pojęcia *c*

x	*n*	*k*	klasyfikacja
1	4	8	1
2	6	12	1
3	3	15	0
4	6	13	1
5	2	4	1
6	10	20	1
7	3	15	0
8	8	16	1
9	5	7	0
10	12	24	1
11	13	1	0
12	22	44	1
13	16	32	1
14	14	28	1
15	5	9	0
16	30	60	1
17	9	7	0
18	7	13	0
19	11	15	0

Zauważmy, że entropia testu t_1 względem zbioru treningowego *T* jest bardzo bliska 0, bowiem test t_1 dzieli nasz zbiór treningowy *T* na dwa zbiory, w których prawie wszystkie elementy (poza jednym) mają taką samą kategorię, czyli są w tych zbiorach zdecydowanie dominujące kategorie. Prawdopodobieństwo tego, że drzewo D_1 będzie błędnie klasyfikować jest duże. To bardzo zły klasyfikator dla pojęcia *c*, ale niemal idealnie zgadza się ze zbiorem treningowym *T*. O klasyfikatorach, które dobrze klasyfikują prawie wszystkie elementy zbioru treningowego, ale w ogólności źle klasyfikują dane, powiadamy, że są **nadmiernie dopasowane do**



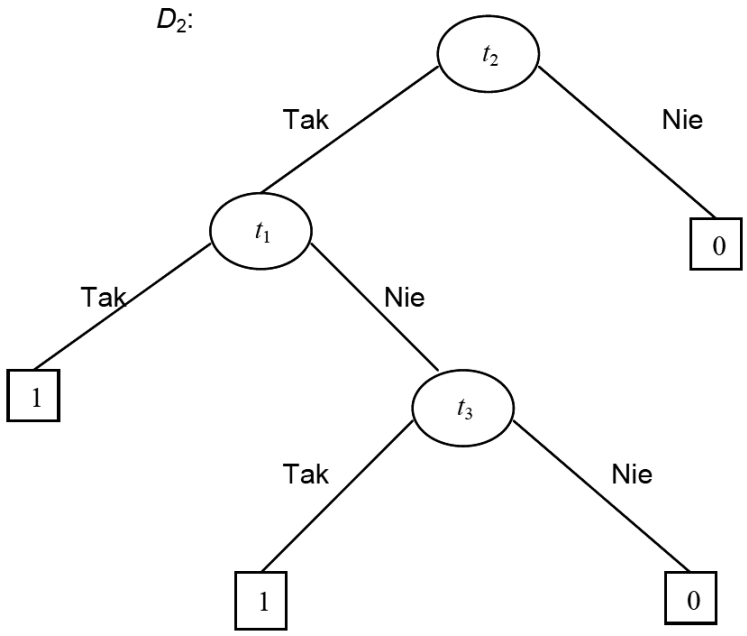
Rysunek 7.
Klasyfikator D_1

zbioru treningowego. Klasyfikator D_1 jest nadmiernie dopasowany do zbioru *T*, bowiem zbiór *T* nie obrazuje należycie struktury pojęcia *c*.

Zadanie 21. Wykorzystaj program służący do liczenia entropii testu (zadanego jako podział podanego zbioru treningowego) do obliczenia entropii testów t_1, t_2, t_3 względem zbioru treningowego *T*.

Drzewo D_2 przedstawione na rys. 8 jest lepszym klasyfikatorem pojęcia *c* niż drzewo D_1, gdyż w porównaniu z D_1 będzie znacznie częściej dobrze klasyfikować dane. Klasyfikator D_2 nie jest tak nadmiernie dopasowany do zbioru treningowego jak klasyfikator D_1.

Zadanie 22. Napisz program, który wylosuje 100-elementowy zbiór przykładów ze zbioru $N \times N$ z mniej więcej równa liczbą elementów kategorii 1 i kategorii 0 i obliczy, ile razy klasyfikator D_1 popełnia błąd na elementach z wylosowanego zbioru, a ile razy D_2 popełnia błąd na tych elementach.



Rysunek 8.
Klasyfikator D_2

4 PROPOZYCJA PRZEPROWADZENIA PROSTYCH BADAŃ

Przypuszczamy, że algorytm indukcji drzewa decyzyjnego z kryterium wyboru testu przez entropię bardzo często oblicza nadmiernie dopasowany klasyfikator. Stawiamy tezę, że jeżeli zmienimy kryterium wyboru testu na zupełnie losowy wybór testu, to obliczane w ten sposób drzewa decyzyjne często nie będą miały znacząco gorszego prawdopodobieństwa błędnej klasyfikacji niż drzewa decyzyjne obliczone z kryterium wyboru testu przez entropię, co prawdopodobnie bierze się stąd, że losując test unikamy nadmiernego dopasowania.

Losowy wybór testu to idea, która ma sens w metodzie tworzenia klasyfikatorów zwanej **lasem losowym**. Lasy losowe znajdują silne zastosowania w poważnych (nawet bardzo poważnych) analizach danych biologicznych.

Drogi uczniu, proponujemy Ci przeprowadzenie, być może wraz z kolegami, własnych badań i eksperymentu programistycznego wedle następującego scenariusza.

1. Ściągnij ze strony http://archive.ics.uci.edu/ml/ pliki z oferowanymi tam zbiorami treningowymi. Użyj także przekazanych przez nas plików heart_disease.txt, iris.txt, diabets.txt, wine.txt. jako zbiorów treningowych.
2. Dla każdego z badanych plików napisz program, który:
   2.1. wczyta zbiór treningowy z pliku,
   2.2. wśród zadeklarowanych funkcji programu będą funkcje reprezentujące zbiór *S* dostępnych testów na danych, stosownie do specyfiki konkretnego zbioru treningowego,
   2.3. program obliczy drzewo decyzyjne z kryterium wyboru testu przez entropię i drzewo decyzyjne z kryterium losowego wyboru testu,
   2.4. zgodnie z metodą walidacji krzyżowej zostaną obliczone prawdopodobieństwa błędnej klasyfikacji dla jednego i drugiego drzewa i te prawdopodobieństwa zostaną wyświetlone jako wyniki obliczeń.

Wyniki obliczeń mogą wesprzeć bądź wręcz obalić postawioną przez nas tezę. Jest to ambitne zadanie wymagające sporej zręczności w programowaniu, ale chyba jeszcze w zasięgu bardzo dobrych uczniów, fanatyków programowania, tzw. *high level fellows*. Staniecie „twarzą w twarz" z problemem właściwego ograniczenia zbioru dostępnych testów. Po prostu wszystkich możliwych testów w niektórych przypadkach będzie bardzo dużo – tak dużo, że w rozsądnym czasie program nie obliczy wynikowego drzewa decyzyjnego. Gdyby wyniki obliczeń wsparły naszą tezę, mielibyście piękny przykład na to, że sformułowany tutaj warunek stopu algorytmu indukcji drzewa decyzyjnego nie zawsze jest adekwatny. Analiza warunku stopu poprowadzi was do zagadnień unikania nadmiernego dopasowania poprzez tzw. przycinanie drzewa.

5 NIEKTÓRE DZIEDZINY ZASTOSOWAŃ METOD EKSPLORACJI DANYCH

Przedstawiamy w skrócie wybrane zastosowania metod eksploracji danych

Automatyczna klasyfikacja plam słonecznych

Astronomowie klasyfikują plamy słoneczne do kilku kategorii na podstawie zestawów danych generowanych przez automatyczne aparatury obserwatoriów słońca. Metody teorii zbiorów przybliżonych Pawlaka zostały użyte w stworzeniu klasyfikatora plam słonecznych z prawdopodobieństwem błędnej klasyfikacji bardzo małym i akceptowalnym z punktu widzenia celów badań plam słonecznych.

Wsparcie diagnostyki w medycynie

Przykładowo, drzewa decyzyjne zostały użyte w klasyfikacji danych medycznych kobiet powyżej 20-go roku życia na:

- dane wskazujące cukrzycę,
- dane wskazujące nieobecność cukrzycy .

Rekord danych zawiera informacje, takie jak: wynik testu glukozowego, ciśnienie rozkurczowe krwi, indeks masy ciała, wiek i inne. Dane z 532 przypadków zostały użyte jako zbór treningowy. Jako dostępne testy przyjęto porównanie wartości wybranego atrybutu, np. wynik testu glukozowego ‹ 127.5, wiek ‹ 42 itp. Obliczone drzewo decyzyjne klasyfikowało pozostałe przypadki.

Bankowość i marketing

Weźmy przykład z badań marketingowych. Przypuśćmy, że ankieterzy zebrali ok. 1000 rekordów danych o klientach. Liczba atrybutów w rekordzie wynosi ok. stu. Przykładowo, przyjęto atrybuty, takie jak: ocena w skali od 1 do 10 wartości wiedzy, ocena w skali od 1 do 10 wartości siły fizycznej.. Psycholog razem z socjologiem sklasyfikowali 50 wybranych rekordów do kategorii, takich jak: „królowe disco polo", „gadżeciarze", „młode leniwe byczki" i jeszcze około 10 innych kategorii. Aby wydostać stosowną informację marketingową, agencja badań marketingowych potrzebuje pełnej klasyfikacji wszystkich rekordów. Po zastosowaniu jednej z metod eksploracji danych (np. indukcji reguł) otrzymano ogólny klasyfikator, który dokonał potrzebnej klasyfikacji. Oczywiście, aby ocenić wiarygodność otrzymanej klasyfikacji, poproszono obu panów psychologa i socjologa, by sklasyfikowali inne niż poprzednio 50 rekordów i porównano wyniki ich klasyfikacji z klasyfikacją automatyczną. Jeżeli zgodność była bardzo znacząca, można uznać wyniki automatycznej klasyfikacji jako podstawę do wyciągnięcia stosownej informacji marketingowej.

Klasyfikacja danych biologicznych

Bardzo często jednym z podstawowych urządzeń laboratoriów biochemicznych jest komputer o dużej mocy obliczeniowej analizujący automatycznie zbierane dane np. o łańcuchach białek. Stosowane są niemal wszystkie metody eksploracji danych, w tym lasy losowe.

LITERATURA

1. Cichosz P., *Systemy uczące się*, WNT, Warszawa 2000
2. Koronacki J., Cwik J., *Statystyczne systemy uczące się*, EXIT, Warszawa 2008
3. Koronacki J., Mielniczuk J., *Statystyka dla studentów kierunków technicznych i przyrodniczych*, WNT, Warszawa 2006
4. Lovet J. N., Trueblood R. P., *Zastosowanie języka SQL do analizy statystycznej i eksploracji danych*, MIKOM, Warszawa 2002
5. Poe V., Klauer P., Brobst S., *Tworzenie hurtowni danych*, WNT, Warszawa 2000

Omówienie literatury

Przykłady zbiorów treningowych i pojęć zostały wzięte z książki [1], algorytm indukcji drzewa decyzyjnego został przedstawiony również w oparciu o tę książkę.(Rozdział o drzewach decyzyjnych.).

Przykład eksploracji o naturze statystycznej został opracowany na podstawie przykładu prezentowanego w książce [4]. Można tam przeczytać intuicyjne wyjaśnienia wielu pojęć statystyki.

Formalne definicje zmiennej losowej, jej wartości średniej i odchylenia standardowego, jak też zagadnienia związane z estymatorami tych wielkości można studiować na podstawie książki [3].

Zagadnienie walidacji krzyżowej jest dobrze przedstawione w książce [2], str. 95 – 98.

W książce [5] sa przedstawione podstawowe informacje o specjalnych bazach danych zwanych hurtowniami danych.

W projekcie **Informatyka +**, poza wykładami i warsztatami, przewidziano następujące działania:

- 24-godzinne kursy dla uczniów w ramach modułów tematycznych
- 24-godzinne kursy metodyczne dla nauczycieli, przygotowujące do pracy z uczniem zdolnym
- nagrania 60 wykładów informatycznych, prowadzonych przez wybitnych specjalistów i nauczycieli akademickich
- konkursy dla uczniów, trzy w ciągu roku
- udział uczniów w pracach kół naukowych
- udział uczniów w konferencjach naukowych
- obozy wypoczynkowo-naukowe.

Szczegółowe informacje znajdują się na stronie projektu

www.informatykaplus.edu.pl