Eksploracja danych

Piotr Lipiński

Eksploracja danych

- Rozpatrzmy pewne zjawisko opisywane przez informację uzyskaną przez zmierzenie i zarejestrowanie pewnych cech rozpatrywanego zjawiska.
- Pomiarów wykonuje się wiele uzyskując wiele **obserwacji** (rekordów). Każda obserwacja opisana jest przez ustaloną liczbę **cech** (atrybutów).
- Przykład: pomiar danych biometrycznych osób wchodzących do budynku
  - zjawiskiem jest wchodzenie człowieka do budynku (interesuje nas więc wiedza o ludziach wchodzących do tego budynku)
  - rejestrujemy atrybuty **numeryczne**, jak wzrost i waga, **kategoryczne** (zwane też **nominalnymi**), jak kolor oczu oraz **złożone** (formalnie numeryczne, ale zazwyczaj przetwarzane inaczej), jak odcisk palca i skan siatkówki

ID	wzrost [cm]	waga [kg]	kolor oczu	odcisk palca	skan siatkówki
Osoba1	164	47.50	szary	<BLOB>	<BLOB>
Osoba2	178	59.20	niebieski	<BLOB>	<BLOB>
Osoba3	192	98.30	zielony	<BLOB>	<BLOB>
Osoba4	187	55.90	czerwony	<BLOB>	<BLOB>

Eksploracja danych

- Niech $D = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ będzie zbiorem danych złożonym z N obserwacji $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$. Każda obserwacja $\mathbf{x}_i$ opisana jest przez d cech $x_{1i}, x_{2i}, \ldots, x_{di}$.
- Zbiór danych D można reprezentować w formie tabeli.
- Jeśli wszystkie atrybuty są numeryczne, to zbiór danych D można reprezentować w formie macierzy $\mathbf{X} \in R^{d \times N}$, zaś każdą obserwację w formie wektora $\mathbf{x} \in R^d$.

ID	wzrost [cm]	waga [kg]	kolor oczu	odcisk palca	skan siatkówki
Osoba1	164	47.50	szary	<BLOB>	<BLOB>
Osoba2	178	59.20	niebieski	<BLOB>	<BLOB>
Osoba3	192	98.30	zielony	<BLOB>	<BLOB>
Osoba4	187	55.90	czerwony	<BLOB>	<BLOB>

Niepewność danych

- Zbiór danych jest zazwyczaj obarczony zaburzeniem (m.in. błędem pomiarowym).
- Przykład:
  - zarejestrowany wzrost pewnej osoby, to nie jest jej rzeczywisty wzrost, a zazwyczaj jedynie wynik pomiaru tego wzrostu
    - wynik pomiaru może być obarczony błędem pomiarowym urządzenia
    - osoba może mieć zmienny wzrost (inny rano, kiedy jest wypoczęta, inny wieczorem, kiedy jest zmęczona)
    - mogą wystąpić błędy numeryczne związane z zapisem danych (reprezentacja binarna liczb rzeczywistych), transmisją danych (błędy transmisji), kompresją danych (dopuszczającą straty dokładności), itp.
  - kilka pomiarów wzrostu tej samej osoby nie musi więc prowadzić do uzyskania tych samych wartości
- Jak więc przetwarzać takie niepewne dane ?

Niepewność danych

- Można traktować rejestrowane wartości jako wartości przybliżone z pewną ustaloną dokładnością
  - na przykład: zarejestrowany wzrost to rzeczywisty wzrost ±1 cm
  - wada 1: jest to niewygodne (kumulacja błędów przy przetwarzaniu danych)
  - wada 2: nie rozwiązuje to problemów z innymi błędami niż pomiarowe
- Można traktować rejestrowane wartości jako obserwacje zmiennych losowych
  - przyjmujemy więc, że rezultat pomiaru jest wartością losową
  - nie znaczy to, że mierząc wzrost wybranego człowieka spodziewamy się rezultatu pomiaru z przedziału [0, 250] cm z jednakowym prawdopodobieństwem (oznaczałoby to kompletną nieprzydatność urządzenia pomiarowego i całej procedury pomiaru)
  - spodziewamy się raczej, że

    ZarejestrowanyWzrost = RzeczywistyWzrost + Zaburzenie
  - gdzie
    - ZarejestrowanyWzrost jest wynikiem pomiaru
    - RzeczywistyWzrost jest pewną liczbą (nielosową, stałą dla każdego człowieka, ale nieznaną i przypuszczalnie niemożliwą do dokładnego zmierzenia)
    - Zaburzenie jest składnikiem losowym (związanym m.in. z urządzeniem pomiarowym, procedurą pomiaru, warunkami pomiarów, itp., którego charakter losowości możemy starać się określić)

Niepewność danych

- Istotne jest ustalenie charakterystyki zaburzenia losowego.
- Przykład 1: Można byłoby założyć, że zaburzenie ma rozkład jednostajny U([-1, 1]) na odcinku [-1, 1].

Parę słów o rozkładzie jednostajnym

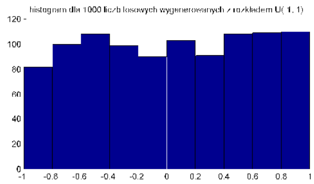
- gęstość rozkładu jednostajnego U(α, β) na odcinku [α, β] to

$$f(x;\alpha,\beta)=\begin{cases}\frac{1}{\beta-\alpha}, & \text{dla } x\in[\alpha,\beta]\\ 0, & \text{dla } x\notin[\alpha,\beta]\end{cases}$$

- wykresem przykładowej gęstości jest poniższa krzywa

Parę słów o rozkładzie jednostajnym

- generując liczby losowe z rozkładem U(α, β) powinniśmy dostawać histogramy zgodne z odpowiednią krzywą gęstości
- przykład: histogram dla 1000 liczb losowych wygenerowanych z rozkładem U(α, β)



Niepewność danych

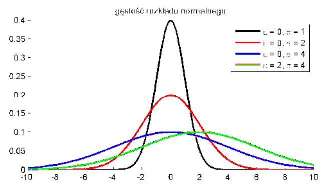
- Istotne jest ustalenie charakterystyki zaburzenia losowego.
- Przykład 1: Można byłoby założyć, że zaburzenie ma rozkład jednostajny U([-1, 1]) na odcinku [-1, 1].
  - podejście robi się podobne do podejścia traktującego dane jako wartości przybliżone z pewną ustaloną dokładnością
  - wada 1: jest to niewygodne (kumulacja zaburzeń przy przetwarzaniu danych)
  - wada 2: nie rozwiązuje to problemów z innymi błędami niż pomiarowe
  - wada 3: założenie jest trochę niepraktyczne, bo czasami w rzeczywistości mogą zdarzyć się bardzo duże błędy (np. przy wykonywaniu pomiarów podczas trzęsienia ziemii) – oczywiście prawdopodobieństwo takich bardzo dużych błędów jest bardzo małe
- Przykład 2: Można byłoby założyć, że zaburzenie ma rozkład normalny N(0, 1) o wartości oczekiwanej 0 i wariancji 1.

Parę słów o rozkładzie normalnym

- gęstość rozkładu normalnego N(μ, σ^2) o wartości oczekiwanej μ i wariancji σ^2 to

$$f(x;\mu,\sigma)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

- wykresem gęstości jest krzywa Gaussa, wartość oczekiwana wpływa na przesunięcie krzywej, a wariancja na jej kształt



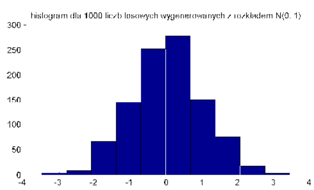
Parę słów o rozkładzie normalnym

- prawdopodobieństwo, że zmienna losowa X ~ N(μ, σ^2) przyjmie wartość z pewnego przedziału [a, b], to pole powierzchni pod wykresem krzywej Gaussa na odcinku [a, b]
- prawdopodobieństwa takie można wyliczyć numerycznie lub odczytać z tablic statystycznych
- można więc precyzyjnie odpowiadać na pytania:
  - jakie jest prawdopodobieństwo, że wartości zmiennej losowej X będą w przedziale [a, b] ?
  - jakie jest prawdopodobieństwo, że wartości zmiennej losowej X przekroczą zadany próg a ?
  - w jakim przedziale będą wartości X z prawdopodobieństwem 95% ?
- zakładając więc, że zaburzenie analizowanych danych ma określony rozkład, można precyzyjnie szacować dokładność obliczeń podczas dalszego przetwarzania danych

Parę słów o rozkładzie normalnym

- generując liczby losowe z rozkładem N(μ, σ^2) powinniśmy dostawać histogramy zgodne z odpowiednią krzywą Gaussa

  UWAGA: Jest kilka popularnych algorytmów generowania liczb pseudolosowych z rozkładem normalnym (zazwyczaj są one już zaimplementowane w popularnych narzędziach programistycznych), m.in. algorytm Boxa-Mullera czy algorytm Ziggurata.
- przykład: histogram dla 1000 liczb losowych wygenerowanych z rozkładem N(0, 1)

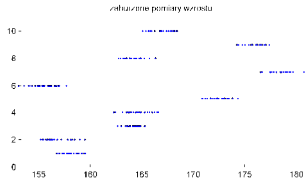


Niepewność danych

- Istotne jest ustalenie charakterystyki zaburzenia losowego.
- Przykład 2: Można byłoby założyć, że zaburzenie ma rozkład normalny N(0, 1) o wartości oczekiwanej 0 i wariancji 1.
  - podejście jest wygodniejsze (kumulacja zaburzeń to operacje na rozkładach normalnych, dość wygodne obliczeniowo)
  - zaburzenie może być dowolnie duże (choć większe zaburzenia są mniej prawdopodobne), więc model jest dość praktyczny
  - oczywiście trzeba dobrać parametry rozkładu: wartość oczekiwaną μ i wariancję σ^2
  - dodatkowa zaleta: wygodne modelowanie zależności między wartościami różnych cech (przyda się w dalszej części wykładu)

Niepewność danych

- Przykład: mierząc wzrost 10 osób, i dla każdej osoby wykonując 25 pomiarów (niezbyt precyzyjnym przyrządem) możemy otrzymać dane przedstawione na poniższym wykresie



  - Co można powiedzieć o zaburzeniu losowym?
  - Jak wyglądałby wykres, gdyby zaburzenie miało rozkład jednostajny?
  - Co zrobić z takimi danymi? Co wpisać do bazy danych?

Niepewność danych

- Zamiast skupiać się na INFORMACJI (pomiarach wzrostu 10 osób), należałoby skupić się na WIEDZY, która z tej informacji wynika.
- Wiedzę o wzroście każdej osoby można byłoby reprezentować przez rozkład prawdopodobieństwa opisujący zarejestrowany wzrost danej osoby.
  - rozkład prawdopodobieństwa dla wzrostu osoby A powinien opisywać zdarzenie losowe polegające na pomiarze wzrostu osoby A, czyli zarejestrowane 25 wartości pomiarów powinno pochodzić właśnie z tego rozkładu.
  - jak określono wcześniej

    ZarejestrowanyWzrost = RzeczywistyWzrost + Zaburzenie,

    jeśli więc wartość oczekiwana zaburzenia jest równa 0, to rzeczywisty wzrost danej osoby jest równy wartości oczekiwanej rozkładu zmiennej losowej ZarejestrowanyWzrost

Niepewność danych

- Należy więc znaleźć rozkład prawdopodobieństwa, z którego pochodzą zarejestrowane wartości wzrostu.
- Dopasowanie rozkładu prawdopodobieństwa do danych często określa się przez funkcję wiarygodności

$$L(\theta) = P(X \mid \theta) = \prod_{x \in D} P(x \mid \theta)$$

gdzie θ to szukany wektor parametrów rozkładu prawdopodobieństwa, X to badana zmienna losowa, a D to zarejestrowana próbka danych

(ostatnia równość zachodzi przy założeniu niezależności danych w próbce D).

Przykład końcowy

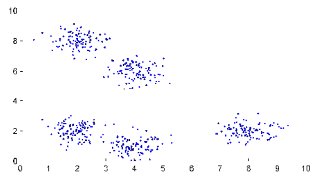
- Rozważamy znów dane o użytkownikach kart kredytowych (klientach pewnego banku). Interesują nas dwie sprawy:
  - Ile człowiek wydaje miesięcznie na zakupy spożywcze?
  - Ile człowiek wydaje miesięcznie na podróże?
- Cel: Bank chciałby coś zarobić
- Jaką wiedzę możemy pozyskać z powyższych danych?
  - zawsze możemy policzyć jakieś średnie
    - średnie wydatki na zakupy spożywcze
    - średnie wydatki na podróże
    - użyteczność takiej wiedzy jest jednak znikoma

Przykład końcowy

- Rozważamy znów dane o użytkownikach kart kredytowych (klientach pewnego banku). Interesują nas dwie sprawy:
  - Ile człowiek wydaje miesięcznie na zakupy spożywcze?
  - Ile człowiek wydaje miesięcznie na podróże?
- Cel: Bank chciałby coś zarobić
- Jaką wiedzę możemy pozyskać z powyższych danych?
  - można byłoby wyodrębnić charakterystyczne grupy ludzi i przygotować dla nich "ofertę specjalną":
    - osoby, które dużo wydają na zakupy spożywcze, a mało podróżują, można skłonić do oszczędzania na żywości i namówić na wycieczkę zagraniczną (prowizja od współpracujących biur podróży)
    - osoby, które dużo wydają na zakupy spożywcze, ale też dużo podróżują, można skłonić do oszczędzania na podróżach i namówić do kupna ekskluzywnej żywności (prowizja od współpracujących delikatesów)
    - …

Przykład końcowy

- Spójrzmy na dane: tabela o N = 500 wierszach (klienci) i o d = 3 kolumnach
  - X0 – identyfikator klienta
  - X1 – średnie miesięczne wydatki na zakupy spożywcze
  - X2 – średnie miesięczne wydatki na podróże



Przykład końcowy

- Czy X1 i X2 można traktować jako zmienne losowe?
  - przecież X1 i X2 to wartości precyzyjnie wyliczone z danych zgromadzonych w bankowej bazie danych
- Warto przypomnieć, że interesuje nas pewne ZJAWISKO, a zarejestrowane dane są jedynie jego OBSERWACJAMI.
  - zjawiskiem jest "styl życia" klientów, a obserwacjami ich zarejestrowane średnie miesięczne wydatki
  - jest całkiem prawdopodobne, że niektórzy klienci płacą też gotówką, więc przypisane im wartości X1 i X2 nie odpowiadają ich rzeczywistym wydatkom (są zaburzone)
- Na razie zignorujmy jednak zaburzenie wartości atrybutów X1 oraz X2 i przejdźmy do prostych algorytmów grupowania danych.