Eksploracja Danych

Dane

(c) Marcin Sydow

Zawartość wykładu

- źródła i formaty danych
- tabela jako podstawowy format danych
- zmienne
- wskaźniki położenia
- wskaźniki rozproszenia
- rozkłady zmiennych
- podsumowanie

Źródła danych

- bazy danych
- hurtownie danych
- dokumenty i kolekcje dokumentów
  - tekst
  - html
  - xml
  - xls
- sieć WWW
- serwery WWW (np. logi)
- portale społecznościowe
  - fb
  - twitter, etc.
- pliki graficzne i multimedialne
- czujniki
- sieci sensorów
- usługi sieciowe, etc.

Formaty danych

Możliwych jest bardzo wiele formatów danych, np:

- csv
- text
- xls
- xml
- pdf
- binarne
- grafowe
- arff

Jakość danych źródłowych

Rola jakości danych wejściowych (reguła "GIGO").

Jeśli dane będą nieodpowiednie (złej jakości, niereprezentatywne, etc.) to niezależnie od poprawnych procedur i modeli wyniki też będą niskiej jakości.

Jako pierwszy etap pracy z danymi niezbędna jest więc faza oceny przydatności/jakości danych.

Ocena przydatności danych

Na tym etapie nie należy jeszcze modyfikować danych źródłowych.

Należy sprawdzić m.in.:

- jakie informacje można odczytać z danych źródłowych?
- czy na podstawie tych danych można uzyskać odpowiedzi na pytania postawione w ramach eksperymentu?
- jakie potencjalne problemy mogą wystąpić podczas próby realizacji eksperymentu

Tabela danych i inne formaty

Podstawowym formatem danych w eksploracji danych jest format **tabeli**

wiersze: obserwacje (przypadki)

kolumny: zmienne (atrybuty)

(przykłady)

(R: format taki nazywany jest “tidy”)

Zmienne

- jakościowe (kategoryczne)
  - porządkowe (np. rozmiar koszulki)
  - regularne (np. kolor auta)
- ilościowe (numeryczne)
  - ciągłe (np. ilość wydobytego gazu ziemnego)
  - dyskretne (np. liczba wyprodukowanych aut)

(przykłady)

Stałe mogą być usunięte (ale należy się upewnić, że są faktycznie stałe w populacji a nie tylko w próbce)

Wartości niepowtarzalne (o ile nie są kluczem), też mogą być usunięte.

Specjalny przypadek: zmienne monotoniczne (ale mogą np. oznaczać czas).

Zmienne anachroniczne

To takie, które zostały uzupełnione w późniejszym czasie, gdy znano już inne wydarzenia.

Pozornie wygladają na bardzo wartościowe (silna korelacja ze zmienną przewidywaną), ale faktycznie są bezwartościowe.

(przykład)

Brakujące dane

Należy odróżnić dwie sytuacje:

- nieznane wartości (brak pomiaru/informacji)
- nieistniejące wartości (wynika ze stanu faktycznego)

- czy występują w danych wartości NULL?
- czy zostały zidentyfikowane i zastąpione (np. flagą "missing")
- czy można uzupełnić braki?
- czy brak niesie wartościową informację?

Reprezentatywność danych

Generalnie im większa próbka, tym lepsza reprezentacja populacji, pod warunkiem, że próbe nie jest obciążona tzn. nie została wybrana przy pewnych specyficznych założeniach, które mogą ją warunkować.

Np. ankieta wykonana tylko przez telefony stacjonarne dotyczy specyficznej podpopulacji ludzi mających takie telefony.

Błędy pomiaru:

- błąd systematyczny (obciążenie)
- błąd przypadkowy (szum)

Statystyki opisowe

Zmienną można podsumować pojedynczą statystyką. Istnieją różne rodzaje statystyk:

- statystyki pozycyjne (wskaźniki położenia)
- statystyki rozproszenia (wskaźniki rozproszenia)
- statystyki symetrii rozkładu (np. skośność)
- statystyki spłaszczenia rozkładu (np. kurtoza)

Statystyki pozycyjne

Statystyki, które podsumowują bezwzględne położenie wartości:

- moda (dominanta): najczęstsza wartość danej zmiennej
- średnia
- mediana (wartość środkowa)
- kwartyle (są trzy: mediana jest drugim kwartylem)

Wartość średnia w próbie (ang. mean)

Zakładamy, że próba jest wektorem $x = (x_1, ..., x_n)$:

$$mean(x) = \frac{1}{n} \sum_{i=1}^{n} x_i$$

(jaka jest złożoność czasowa obliczenia średniej?)
(R: `mean(x)`)

Mediana (ang. median)

Mediana wyznacza wartość środkową w ciągu uporządkowanych wartości danej zmiennej. W przypadku parzystej liczności jest to średnia z 2 środkowych wartości.

Załóżmy, że próba została posortowana niemalejąco $(x_{(1)}, ..., x_{(n)})$:

$$median(x) = x_{(\frac{n+1}{2})}$$

(n nieparzyste)

$$median(x) = \frac{(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)})}{2}$$

(n parzyste)
(jaka jest złożoność czasowa obliczenia mediany?)
(R: `median(x)`)

Średnia vs mediana

Średnia jest podstawową statystyką pozycyjną i łatwą w policzeniu, ale jest mniej niż mediana odporna na wartości odstające.

Przykład: gdy x oznacza pensję w danej firmie, a nieliczna grupa zarabia dużo więcej niż inni (np. prezesi, rektorzy, etc.) wartość średnia może dawać nieco zmylający obraz zarobków. Mediana jest bardziej odporna na nieliczne wartości odstające.

(w przypadku rozkładów nieunimodalnych zarówno wartość średnia jak i mediana mogą być mało przydatne - lepiej wtedy pogrupować dane i odrębnie podsumować oddzielne grupy)

Średnia ucinana (ang. trimmed mean) i średnia winsorowska

Warianty średniej bardziej odporne na wartości odstające.

Średnia ucinana: (pomija k wartości najmniejszych i k największych)

$$mean_t(x) = \frac{1}{n-2k} \sum_{i=k+1}^{n-k} x_{(i)}$$

(R: `mean(x, trim=`α`)`, gdzie α to ułamek wartości do pominięcia)

Średnia winsorowska: (zastępuje te wartości najbliższym sąsiadem co do wartości)

$$winsor(x) = \frac{1}{n}[(k+1)x_{(k+1)} + \sum_{i=k+2}^{n-k-1} x_{(i)} + (k+1)x_{(n-k)}]$$

Inne przydatne statystyki

- liczba istniejących wartości
- liczba unikatowych wartości
- liczba brakujących wartości

Wskaźniki rozproszenia

Statystyki informujące o tym jak bardzo poszczególne wartości są rozproszone (zmienne):

- rozstęp próby
- wariancja i odchylenie standardowe
- rozstęp ćwiartkowy (IQR: inter-quartile range)

Wariancja w próbie

Zakładamy, że próba jest wektorem $x = (x_1, ..., x_n)$.
Wariancja:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - mean(x))^2$$

Jest miarą rozrzutu wartości danej zmiennej wokół wartości średniej.
Pewnym problemem technicznym jest fakt, że jednostką wariancji jest kwadrat jednostki badanej zmiennej.

Odchylenie standardowe:

$$s = \sqrt{s^2}$$

(odchylenie standardowe ma już te same jednostki co badana zmienna)
(R: `var(x)`, `sd(x)`)

Odchylenie przeciętne

Obecność kwadratów we wzorze na wariancję powiększa wpływ wartości znacznie odbiegających od średniej (i jednocześnie pomniejsza tych będących b.blisko)

Odchylenie przeciętne nie ma takiej wady:

$$d_1 = \frac{1}{n} \sum_{i=1}^{n} |x_i - mean(x)|$$

(jednak jest to wyrażenie trudniejsze analitycznie)

Kwartyle i IQR

Kwartyl może być rozumiany jako uogólnienie mediany.

Oznacza on wartość będącą w odpowiednio w:

- jednej czwartej (dolny kwartyl) (Q_1)
- połowie (mediana)
- trzech czwartych (górny kwartyl) (Q_3)

posortowanych niemalejąco wartości danej zmiennej.

Rozstęp ćwiartkowy (IQR):

$$IQR = Q_3 - Q_1$$

(miara rozproszenia)

Zauważmy, że pomiędzy dolnym i górnym kwartylem znajduje się dokładnie połowa obserwowanych wartości.

Wykres ramkowy

Jest to graficzna forma podsumowania kilku podstawowych wskaźników położenia i rozproszenia jednocześnie:

- Q1: dolny bok ramki
- mediana: pozioma linia w środku ramki
- Q3: górny bok ramki
- górny "wąs" czyli najwyższa obserwacja od mediany będąca nie dalej niż $Q_3 + 1,5 \times IQR$
- dolny "wąs" czyli najniższa obserwacje od mediany będąca nie dalej niż $Q_1 - 1,5 \times IQR$
- wartości odstające: zaznaczane jako kropki poza wąsami.

(R: `boxplot(x)`)

Rokład zmiennej (dyskretnej) i Histogram

Jak często zdarza się dana wartość.

Histogram: Graficzne narzędzie do wizualizacji rozkładu.

(R: `hist(x)`)
Uwaga: bardzo ważnym parametrem histogramu jest szerokość koszyka.

(przykład)

Rozkład:

- jednomodalny (istnieje jedno maksimum)
- bimodalny (dwa maksima)
- wielo-modalny (wiele maksimów)

Gęstość rozkładu zmiennej ciągłej

Jest to ciągła funkcja nieujemna, która dererminuje prawdopodobieństwo, że zmienna ta przybiera daną wartość.

- pole powierzchni pod wykresem gęstości wynosi 1.
- prawdopodobieństwo, że zmienna ciągła ma wartość pomiędzy a i b równa się polu powierzchni pod wykresem (całce z) gęstości pomiędzy wartościami a i b.

Histogram można uważać za pewne przybliżenie wykresu funkcji gęstości.

Dystrybuanta rozkładu zmiennej ciągłej

Jest to funkcja $\Phi(a)$ informująca jakie jest prawdopobieństwo, że zmienna o danym rozkładzie przyjmie wartości niewiększe niż a (pole pod krzywą gęstości od $-\infty$ do a).

Rozkład Normalny (Gaussa)

Funkcja gęstości rozkładu normalnego $N(\mu, \sigma)$ dana jest wzorem:

$$\phi_{\mu,\sigma} = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

(μ, σ to parametry będace odpowiednio wartością średnią rozkładu i jego odchyleniem standardowym)

Rozkład $N(0, 1)$ nazywamy standardowym rozkładem normalnym.

Rozkład normalny ma pewne bardzo specjalne własności powodujące, że jego przybliżenia są często spotykane (dzięki centralnemu twierdzeniu granicznemu).

(R: `dnorm`, example: `plot(dnorm(seq(-5,5,0.01)))`)
(R: `rnorm,qnorm,pnorm`: generator losowy, kwantyl, dystrybuanta)

Kwantyle

Kwantyl to uogólnienie kwartyla na przypadek ciągły.
Kwantyl rzędu $\alpha \in \{0, 1\}$.
(R: `quantile(x)`)

Przykładowe znane rozkłady

Rozkłady zmiennej:

- jednostajny $U(0,1)$ (każda wartość na przedziale $[0,1]$ jest jednakowo prawdopodobna)
- dwupunktowy (są tylko 2 możliwe wartości z prawdopodobieństwem p i 1-p)
- dwumianowy (liczba "1" w n próbach dwupunktowego)
- Poissona ($p(k;\lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$) (R: plot(dpois(seq(0,10),2))

Standaryzacja zmiennej

Przekształcenie liniowe zmiennej x tak, aby miała wartość oczekiwaną 0 i odchylenie standardowe 1:

$$\frac{x - mean(x)}{sd(x)}$$

(R: `sx <- (x-mean(x))/sd(x)`
sprawdzenie: `mean(sx)`, `sd(sx)`)

Współczynnik korelacji liniowej (Pearsona) dwóch zmiennych

Jest to jedna z miar współzależności dwóch zmiennych x i y:

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - mean(x)}{sd(x)}\right)\left(\frac{y_i - mean(y)}{sd(y)}\right)$$

Przyjmuje wartości od -1 (pełna negatywna zależność liniowa) przez 0 (brak zależności) do 1 (pełna pozytywna zależność liniowa)
(nieustandaryzowany wariant nazywany jest kowariancją)

(R: cor(x,y))

Istnieją też inne warianty tej miary (nieliniowe) np: kendalla, spearmana.
(R: cor(x, y, method = "kendall"), cor(x,y,method = "spearman")

Korelacje między zmiennymi

Eksploracja Danych

(c) Marcin Sydow

Wstęp

Tabela

Zmienne

Statystyki Opisowe

Wskaźniki rozproszenia

Rozkład zmiennej

Rozkład normalny

Zależności zmiennych

Momenty centralne

Podsumowanie

Korelacja nie oznacza przyczynowości

Przykłady przypadkowych korelacji:

`http://www.tylervigen.com/spurious-correlations`

Momenty centralne

Moment centralny rzędu k zmiennej x:

$$\mu_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - mean(x))^k$$

(zauważmy, że np. dla k=2, μ_k to wariancja[1])

Momenty wyższych rzędów używane są do innych miar rozkładu.

[1] Uwaga: w przypadku nieobciążonej estymacji momentów z próbki wzory są nieco zmodyfikowane, ale idea jest b.podobna

Skośność

Jest to miara symetrii rozkładu.

$$A = \frac{\mu_3}{\sigma^3}$$

- rozkład symetryczny (np. Gaussa)
- rozkład lewoskośny (ujemna skośność): dłuższy lewy ogon
- rozkład prawoskośny (dodatnia skośność): dłuższy prawy ogon

Uwaga: nieobciążony estymator skośności z próbki ma nieco bardziej skomplikowany wzór, ale podobna idea.

Kurtoza

Miara zagęszczenia rozkładu wokół wartości centralnej.

$$Kurt(x) = \frac{\mu_4}{\sigma^4}$$

- wartość zero (np. rozkład Gaussa)
- wartość ujemna (platykurtoza): spłaszczony
- wartość dodatnia (leptokurtoza): wyostrzony

Uwaga: nieobciążony estymator kurtozy z próbki ma nieco bardziej skomplikowany wzór, ale podobna idea.

Miara zróżnicowania dyskretnej zmiennej: Entropia informacji

Im wyższa entropia tym wyższe zróżnicowanie.

Mający dany rozkład $P(x = k)$ zmiennej x, jej entropia dana jest wzorem:

$$H(x) = -\sum_k P(x = k) log_2(P(x = k))$$

W termodynamice entropia (upraszczając) jest miarą chaosu (im wyższe uporządkowanie tym niższy chaos)

W teorii informacji jest ona miarą wartości informacji.
(R: `entropy(c(0.5,0.5))`)

Podsumowanie

- źródła i formaty danych
- tabela jako podstawowy format danych
- zmienne
- wskaźniki położenia
- wskaźniki rozproszenia
- rozkłady zmiennych

Przykładowe pytania/zadania/problemy

- rola fazy oceny przydatności danych. Jakie pytania można w tej fazie postawić?
- opisz format tabelaryczny. Co reprezentują wiersze, co reprezentują kolumny?
- wymień rodzaje zmiennych i dla podanej tabeli dokonaj klasyfikacji zmiennych
- wymienić, podać wzory i umieć obliczyć wskaźniki położenia zmiennej dla niewielkich próbek
- j.w. dla wskaźników rozproszenia
- co to jest rozkład zmiennej
- wzór i podstawowe własności gęstości rozkładu normalnego
- standaryzacja zmiennej
- momenty centralne i ich zastosowanie w podsumowaniach zmiennych

Dziękuję za uwagę.