Istraživanje podataka

Vežbe 2

26. Februar 2020

Outline

Outline

Bliskost-sličnost i različitost

- Sličnost
  - Numerička mera koliko su dva objekta slična
  - Što dva objekta više liče jedan na drugi sličnost im je veča
  - Često se meri vrednostima u intervalu $[0, 1]$
- Različitost
  - Numerička mera koliko su dva objekta različita
  - Što dva objekta više liče jedan na drugi različitost im je manja
  - Najmanja različitost je često 0; gornja granica varira
  - Kao sinonim koristi se i termin rastojanje
- Blizina (eng. proximity) označava ili sličnost ili različitost

Sličnost i različitost za jedan atribut

Attribute Type	Dissimilarity	Similarity
Nominal	$d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$	$s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$
Ordinal	$d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to $n-1$, where n is the number of values)	$s = 1 - \frac{\lvert p-q \rvert}{n-1}$
Interval or Ratio	$d = \lvert p - q \rvert$	$s = -d$, $s = \frac{1}{1+d}$ or $s = 1 - \frac{d - min\_d}{max\_d - min\_d}$

Različitost između objekata podataka

Rastojanje Minkovskog:

$$dist = (\sum_{k=1}^{n} |p_k - q_k|^r)^{\frac{1}{r}}$$

gde je

- r parametar
- n broj dimenzija (atributa)
- p_k i q_k su vrednosti k. atributa objekata p i q

Rastojanje Minkovskog

- $r = 1$ Menhetn (L1 norma) rastojanje
  - Hamingovo rastojanje
- $r = 2$ Euklidsko rastojanje
- $r \to \infty$. "supremum" (Lmax norma) rastojanje
  - Predstavlja maksimum razlike između odgovarajućih komponenti vektora
- standardizacija
- normalizacija

Mera sličnosti za binarne podatke

- p i q - binarni vektori
  - M_{01} broj atributa koji su 0 u p i 1 u q
  - M_{10} broj atributa koji su 1 u p i 0 u q
  - M_{00} broj atributa koji su 0 u p i 0 u q
  - M_{11} broj atributa koji su 1 u p i 1 u q

Mera sličnosti za binarne podatke

- Jednostavno uparivanje koeficijenata (eng. SMC)

$$SMC = \frac{broju\_uparenih}{broj\_atributa} = \frac{M_{11}+M_{00}}{M_{01}+M_{10}+M_{11}+M_{00}}$$

- Žakardovi (Jaccard) koeficijenti
  - asimetrični binarni atributi

$$J = \frac{broj\_parova\_11}{broj\_ne\_oba-su-nula\_vrednosti\_atributa} = \frac{M_{11}}{M_{01}+M_{10}+M_{11}}$$

Mera sličnosti

- Kosinusna sličnost
  $cos(p, q) = \frac{p \bullet q}{\|p\|\|q\|}$
  - p i q - dva vektora
  - $\bullet$ označava skalarni proizvod vektora
  - $\|d\|$ označava dužinu vektora d
  - asimetrični podaci
  - najčešća mera sličnosti dokumenata

Mera sličnosti

- Korelacija

$$r = \frac{kovarijansa(x,y)}{standardna\_devijacija(x) * standardna\_devijacija(y)}$$

  - x i y - dva vektora
  - Korelacija dva objekta koji imaju binarne ili neprekidne atribute je mera linearnog odnosa između njihovih atributa

Mera sličnosti

$$r = \frac{\sum (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum (x_i - \bar{X})^2 \sum (y_i - \bar{Y})^2}}$$

kovarijansa(x,y)= $s_{xy} = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})(y_k - \bar{y})$

standardna devijacija(z)= $s_z = \sqrt{\frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^2}$

srednja vrednost od z= $\bar{z} = \frac{1}{n} \sum_{k=1}^{n} z_k$

Zadaci

1. Sledeći atributi su korišćeni za opis članova krda azijskih slonova: težina, visina, dužina kljove, površina uveta. Koju meru bliskosti treba koristiti za poređenje ili grupisanje slonova?

Zadaci

1. Sledeći atributi su korišćeni za opis članova krda azijskih slonova: težina, visina, dužina kljove, površina uveta. Koju meru bliskosti treba koristiti za poređenje ili grupisanje slonova?
   Svi atributi su numerički, ali mogu imati različit opseg vrednosti (zavisno od skale na kojoj su mereni). Nisu asimetrični i veličina atributa je važna. Euklidsko rastojanje, pri čemu se vrši standardizacija da sredina bude 0 i standardna devijacija 1.

Zadaci

2 Data je dokument-term matrica u kojoj je tf_{ij} frekvencija i-te reči (terma) u j-tom dokumentu i m je broj dokumenata. Ako je data transformacija nad promeljivom
$tf_{ij}^{'} = tf_{ij} * \log(\frac{m}{df_i})$
gde je df_i broj dokumenata u kojima se term i pojavljuje (dokument frekvencija terma). Ova transformacija je poznata kao inverzna dokument frekvencija.

Zadaci

- Šta je rezultat ove transformacije ako se reč pojavljuje u jednom dokumentu? U svakom dokumentu?

Zadaci

- Šta je rezultat ove transformacije ako se reč pojavljuje u jednom dokumentu? U svakom dokumentu?
  Ako se reč pojavljuje u svakom dokumentu ima težinu 0, a ako se pojavljuje u jednom dokumentu ima težinu $\log m$.

Zadaci

- Šta je rezultat ove transformacije ako se reč pojavljuje u jednom dokumentu? U svakom dokumentu?
  Ako se reč pojavljuje u svakom dokumentu ima težinu 0, a ako se pojavljuje u jednom dokumentu ima težinu $\log m$.
- Koji je cilj ove transformacije?

Zadaci

- Šta je rezultat ove transformacije ako se reč pojavljuje u jednom dokumentu? U svakom dokumentu?
  Ako se reč pojavljuje u svakom dokumentu ima težinu 0, a ako se pojavljuje u jednom dokumentu ima težinu $\log m$.
- Koji je cilj ove transformacije?
  Razlikovanje dokumenta po rečima koja se retko pojavljuju.

Zadaci

3 Upoređivanje mera sličnosti i razlika

- Izračunati Hamingovo rastojanje i Žakardov koeficijent za vektore
  x=0101010001
  y=0100011000

Zadaci

3 Upoređivanje mera sličnosti i razlika

- Izračunati Hamingovo rastojanje i Žakardov koeficijent za vektore
  x=0101010001
  y=0100011000
  Hamingovo rastojanje = broj različitih bitova=3
  J= broj parova 11/ broj ne oba-su-nula vrednosti atributa=2/5=0.4

Zadaci

- Ako se poredi koliko su slična dva organizma različitih vrsta preko broja gena koji dele, koju meru treba koristiti, Hamingovo rastojanje ili Žakardov koeficijenat radi poređenja genetskog sklopa dva organizma? (Svaki organizam je predstavljen kao binarni vektor, gde je svaki atribut 1 ako organizam sadrži određeni gen, a u suprotnom je 0).

Zadaci

- Ako se poredi koliko su slična dva organizma različitih vrsta preko broja gena koji dele, koju meru treba koristiti, Hamingovo rastojanje ili Žakardov koeficijenat radi poređenja genetskog sklopa dva organizma? (Svaki organizam je predstavljen kao binarni vektor, gde je svaki atribut 1 ako organizam sadrži određeni gen, a u suprotnom je 0).
  Žakardov koeficijent je bolji za poređenje genetskog sklopa dva organizma, jer se dobija podatak koliko gena dele.

Zadaci

- Ako se porede dva organizma iste vrste (npr. dva čoveka), da li je bolje koristiti Hamingonovo rastojanje ili Žakardov koeficijent? Dva čoveka imaju preko 99,9% istih gena.

Zadaci

- Ako se porede dva organizma iste vrste (npr. dva čoveka), da li je bolje koristiti Hamingonovo rastojanje ili Žakardov koeficijent? Dva čoveka imaju preko 99,9% istih gena.
  Hamingonovo rastojanje, jer nas zanimaju razlike

Zadaci

4 Za vektore x i y izračunati navedene mere sličnosti ili razlike:

- x=(1,1,1,1), y=(2,2,2,2) kosinusna sličnost, korelacija, Euklidsko rastojanje

Zadaci

4 Za vektore x i y izračunati navedene mere sličnosti ili razlike:

- x=(1,1,1,1), y=(2,2,2,2) kosinusna sličnost, korelacija, Euklidsko rastojanje
  cos(x,y)=1, corr(x,y)=0/0, Euklidsko(x,y)=2

Zadaci

4 Za vektore x i y izračunati navedene mere sličnosti ili razlike:

- x=(1,1,1,1), y=(2,2,2,2) kosinusna sličnost, korelacija, Euklidsko rastojanje
  cos(x,y)=1, corr(x,y)=0/0, Euklidsko(x,y)=2
- x=(0,1,0,1), y=(1,0,1,0) kosinusna sličnost, korelacija, Euklidsko rastojanje, Žakardov koeficijent

Zadaci

4 Za vektore x i y izračunati navedene mere sličnosti ili razlike:

- x=(1,1,1,1), y=(2,2,2,2) kosinusna sličnost, korelacija, Euklidsko rastojanje
  cos(x,y)=1, corr(x,y)=0/0, Euklidsko(x,y)=2
- x=(0,1,0,1), y=(1,0,1,0) kosinusna sličnost, korelacija, Euklidsko rastojanje, Žakardov koeficijent
  cos(x,y)=0, corr(x,y)=-1, Euklidsko(x,y)=2, Žakard(x,y)=0

Zadaci

- x=(0,-1,0,1), y=(1,0,-1,0) kosinusna sličnost, korelacija, Euklidsko rastojanje

Zadaci

- x=(0,-1,0,1), y=(1,0,-1,0) kosinusna sličnost, korelacija, Euklidsko rastojanje
  cos(x,y)=0, corr(x,y)=0, Euklidsko(x,y)=2

Zadaci

4 Ako mera sličnosti ima vrednosti u intervalu $[0, 1]$, kako biste transformisali vrednost sličnost u vrednost različitosti u intervalu $[0, \infty]$?

Zadaci

4 Ako mera sličnosti ima vrednosti u intervalu $[0, 1]$, kako biste transformisali vrednost sličnost u vrednost različitosti u intervalu $[0, \infty]$?
$d = -\log s$

Zadaci

5 Bliskost je obično definisana između para objekata.

- Kako se može izračunati razlika između dva skupa tačaka u Euklidskom prostoru?

Zadaci

5 Bliskost je obično definisana između para objekata.

- Kako se može izračunati razlika između dva skupa tačaka u Euklidskom prostoru?
  Npr. računanjem centroida između skupa tačaka

Zadaci

5 Bliskost je obično definisana između para objekata.

- Kako se može izračunati razlika između dva skupa tačaka u Euklidskom prostoru?
  Npr. računanjem centroida između skupa tačaka
- Kako se može definisati bliskost između dva skupa objekata?

Zadaci

5 Bliskost je obično definisana između para objekata.

- Kako se može izračunati razlika između dva skupa tačaka u Euklidskom prostoru?
  Npr. računanjem centroida između skupa tačaka
- Kako se može definisati bliskost između dva skupa objekata?
  Prosečna vrednost bliskosti parova iz različitih grupa, ili najmanja ili najveća bliskost parova iz različitih grupa.

Priprema podataka

- **diskretizacija** - transformacija neprekidnog atributa u kategorički atribut
- **binarizacija** - transformacija atributa u jedan ili više binarnih atributa

Outline

Praktičan zadatak

Na skupu bank.csv izvršiti sledeće promene:

- Eliminisati instance koje imaju negativnu vrednost u atributu *srednje_god_stanje_eur*.
- Vrednosti atributa *starost* podeliti u 5 kategorija jednake širine.
- Promeniti kategorije atributa *bracno_stanje* u *u_braku* i *nije_u_braku*.
- Standardizovati numeričke atribute.