1. ODKRIVANJE SKUPIN

16/1/20

Evklidova razdalja: $d(a,b)=\sqrt{\sum_{i=1}(a_i-b_i)^2}$

Manhattanska razdalja: $d(a,b)=\sum_{i=1}^{n}|a_i-b_i|$

Minkowska razdalja: $d(a,b)=\left(\sum_{i=1}^{n}|a_i-b_i|^r\right)^{\frac{1}{r}}$

Razdalje med skupinami:

Single linkage

Complete linkage

Average linkage

Wardova razdalja (centroidi):

$$d_w(C_u,C_v)=\sum_{x\in C_{uv}} d(x,R_{uv})^2-\left(\sum_{x\in C_u} d(x,R_u)^2+\sum_{x\in C_v} d(x,R_v)^2\right)$$

2. METODA VODITELJEV

Izbira k gruč → minimizacija: $\sum_{i=1}^{k}\sum_{x\in C_i} d(v^{(i)},x)$

↳ prevod ↳ SSE (sum of squared errors): $SSE=\sum_{i=1}^{k}\sum_{x\in C_i}(v^{(i)}-x)^2$

Kohezija: ~~oddaljenost~~
↳ podobnost primerov znotraj skupine

Ločljivost:
↳ oddaljenost primerov med različnimi skupinami

silhueta - mera ocenjevanja kvalitete razvrstitve

$s=\frac{b-a}{\max(b,a)}$ $[-1,1]\to 0$ je dobra silhueta točke

Silhuetni koeficient == silhueta razbitja (mera razbitja združuje)

Normalizacije: $\boxed{z_{ij}=\frac{x_{ij}-\mu_j}{\sigma_j}}$ $\sigma_j=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_{ij}-\mu_j)^2}$

3. RAZVRŠČANJE BESEDIL

16/1/20

inverse document frequency:

$$IDF(t) = \log \frac{|D|}{|\{d : t \in d\}| + 1}$$

$|D|$ → št. dokumentov

$|\{d : t \in d\}|$ → št. el., ki vsebujejo el. t (kjer je $tf(t,d) \neq 0$)

$$tf\text{-}idf(t,d) = tf(t,d) \times idf(t)$$

Kosinusna podobnost:

$$a \cdot b = \|a\| \cdot \|b\| \cdot \cos\theta$$

$$sim(x,y) = \cos(\theta) = \frac{xy}{\|x\| \cdot \|y\|}$$

Podobnost po jaccardu: $J(x,y) = \frac{|x \cap y|}{|x \cup y|}$

4. PROJEKCIJE IN ZMANJŠEVANJE DIMENZIONALNOSTI PODATKOV

središče projekcije na enotski vektor:

$$\bar{x} = \frac{1}{m} \sum_i x^{(i)}$$

Razpršenost podatkov na projekciji:

$$Var(u_1^T X^T) = \frac{1}{m} \sum_{i=1}^{m} (u_1^T x^{(i)} - u_1^T \bar{x})^2$$

↳ želimo maksimizirat, zbrat primaren enotski vektor $\vec{u}_1$

$u_1^T \cdot u_1 = 1$

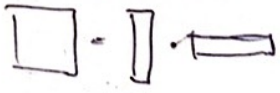
Kovariančna matrika:

$$S = \frac{1}{m} \sum_{i=1}^{m} (x^{(i)} - \bar{x})(x^{(i)} - \bar{x})^T$$

↳ n-atributov → $\mathbb{R}^{n \times n}$

$$Var(u_1^T X^T) = u_1^T S u_1$$

$$S = (x^i - \bar{x}) \cdot (x^i - \bar{x})^T$$



$$S u_1 = \lambda_1 \cdot u_1$$

$$Var(u_1^T x^T) = \lambda_1$$

(Numerična) potenčna metoda

↳ računanje prvih n lastnih vrednosti pri velikem št. atributov

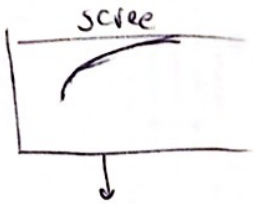
↳ 1 naenkrat

Večrazsežnostno lestvičenje: vizualizira podatke v obliki zemljevida (MDS) → pomanjkljivost: optimizira (tudi) oddaljene primere

Stohastična vložitev sosedov (SNE)

↳ izboljšava MDS, poudarek je na sosedih, skoncentrira se na lokalnost

↳ naključen začetek vložitve



16/1/20

5. LINEARNA REGRESIJA

Kriterijska funkcija: $J = \frac{1}{2m} \sum_{i=1}^{m} (f(x^{(i)} - y^{(i)})^2 = \frac{1}{2m} \sum_{i=1}^{m} \varepsilon^{(i)2}$

↳ jo minimiziramo

$$y = f(x) = ax + b$$

$$\frac{\partial J(a,b)}{\partial a} = \frac{1}{m} \sum_{i=1}^{m} \left((ax^{(i)} + b) - y^{(i)} \right) x^{(i)}$$

$$\frac{\partial J(a,b)}{\partial b} = \frac{1}{m} \sum_{i=1}^{m} \left((ax^{(i)} + b) - y^{(i)} \right)$$

Multivariantna lin. reg.

$h_\theta(x) = \theta_0 + \theta_1 x_1 + \ldots + \theta_n x_n$

$(y = ax + b)$

$= \theta_1 x + \theta_0$

$h(x) = \theta^T x$

↳ prvi element je 1

$$\min_\theta \sum_{i=0}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

kriterijska fja:

$$J(\theta) = \frac{1}{2m} \sum_{i=0}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

za pol" — da ni odvisno od št. podatkov

Popravek za 1 parameter:

$$\theta_i \leftarrow \theta_i - \frac{\alpha}{m} (h_\theta(x) - y) x_i$$

za vse primere pa:

$$\theta \leftarrow \theta - \frac{\alpha}{m} \sum_{j=1}^{m} \left(h_\theta(x^{(j)}) - y^{(j)} \right) x_i^{(j)}$$

Paketni pristop (za vse primere naenkrat):

$$\theta \leftarrow \theta - \alpha X^T (X\theta - \vec{y})$$

Stohastični pristop (v for loopu vsak primer naenkrat)

Normalizacija:

$$x_i \leftarrow \frac{x_i - \mu_i}{\sigma_i}$$

μ_i ← avg

σ_i ← varianca

6. REGULARIZACIJA

→ minimizira vsoto kvadratov thet

→ pomaga pri preprečevanju preveliko prileganja, nočem prevelikih koeficientov

16/1/20

$$RMSE(Y) = \sqrt{\frac{\sum_{i=1}^{k} y^{(i)} - \hat{y}^{(i)}}{k}}$$

testna množica

→ napaka neodvisna od domena odvisne spr.

$$R^2(\hat{y}) = 1 - \frac{\sum_k (y^{(i)} - \hat{y}^{(i)})^2}{\sum_k (y^{(i)} - \bar{y})^2}$$

delež razložene variance

avg

dober model ima to 0, torej je $R^2 \doteq 1$ → good :)

$$J = \frac{1}{2m} \sum_{i=1}^{m} (f(x^{(i)}) - y^{(i)})^2 + \eta \sum_{j=1}^{n} \theta_j^2$$

cenovna fun. za lin. reg.

želimo čim manjše (parametre)

η → stopnja regularitacije (0,1 ; 0,01 ali 0,001)

↳ naredi binaren nabor značilk (jih nastavi na 0)

L_2 nastavi nepomembne thete (parametre) na zelo malo in ne 0.

Gradient (odvod cenilne funkcije):

$$\frac{\partial}{\partial \theta_i} J(\theta) = \frac{1}{m}(h_\theta(x) - y)\, x_i + 2\eta\theta_i$$

$$\theta_j \leftarrow \theta_j \left(1 - \frac{\lambda}{m}\right) - \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right) x_j^{(i)}$$

7. LOGISTIČNA REGRESIJA

Log fun: $g(z) = \frac{1}{1+e^{-z}}$

↳ $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$

odvod: $\frac{dg(z)}{dz} = g(z)\cdot[1-g(z)] = \frac{1}{1+e^{-z}} \cdot \left(1 - \frac{1}{1+e^{-z}}\right)$

$$p(y|x;\theta) = (h_\theta(x))^y (1-h_\theta(x))^{1-y}$$

Verjetje L

Verjetnost $L(\theta) = p(\vec{y}|X;\theta) = \prod_{i=1}^{m} h_\theta(x^i)^{y^i} \cdot (1-h_\theta(x^i))^{1-y^i}$

iščemo parametre θ, da bo L največja

logaritem verjetja $l(\theta)$, ker je odvajanje funkcij z mnogo produkti, precej mučno

logaritem verjetja (log likelihood):

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{m} \left[y^{(i)} \log h_\theta(x^i) + (1-y^i) \log(1-h_\theta(x^i)) \right]$$

gradient log. verjetja → gd maksimizira

$$\frac{\partial}{\partial \theta_j} l(\theta) = \sum_{i=1}^{m} (y^i - h_\theta(x^i))\, x_j^i$$

optimizacija parametrov modela: $\theta_j \leftarrow \theta_j + \alpha \sum_{i=1}^{m} (y^i - h_\theta(x^i))\, x_j^i$

8. KLASIFIKACIJSKA DREVESA & GOZDOVI 17/1/20

Residualna entropija

$$H_{res} = H(C|X) = \sum_i p(x_i) H(C|x_i)$$

Information gain:

$$IG(x) = H(C) - H(C|X)$$

↓ cr. entr.

koliko informacije nam prispeva poznavanje vrednosti posameznega atributa.

Information gain ratio:

$$IGR(x) = \frac{IG(x)}{H(x)}$$ → uravnotežen informacijski prispevek

Mera nečistoče po Giniju:

$$Gini(C) = \sum_i p(C=c_i) \times (1 - p(C=C_i)) = 1 - \sum_i [p(C=c_i)]^2$$

pove: kako pogosto bi bil za naključno izbran primer napačno napovedan razred, ki bi ga uteženo naključno napovedali

→ nič je, ko vsi primeri pripadajo istemu razredu

Kategoritacija / diskretizacija atributa

Gozdovi:

Raw importance: $RI(X) = \frac{1}{k} \sum_{i=1}^{k} C_i - C_{i,x}$
(pomembnost atributa)

9. PRIPOROČILNI SISTEMI

17/1/20

napoved preferenčne ocene uporabnika u za izdelek i: $\rightarrow \hat{r}_{ui} = \frac{\sum s(u,u') \times r_{u'i}}{\sum s(u,u')}$

(utež = podobnosti med uporabnikoma)

(vsota uteži = normalizacija)

podobnost: (cos)

$s_c(u,u') = \frac{r_u r_{u'}}{|r_u||r_{u'}|}$

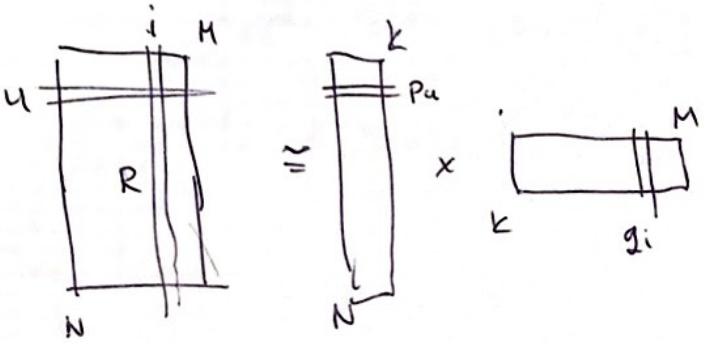
$P_{N\times k}$ in $Q_{k\times M} \rightarrow R \approx PQ$

$k \ll M, N$

k... stopnja razcepa

Latentni profil uporabnika p_u

L.p. stvari: q_i



približek ocene:

$\rightarrow \hat{r}_{ui} = p_u^T \cdot q_i$

$e_{ui}^2 = \frac{1}{2}\left(r_{ui} - \sum_{k=1}^{k} p_{uk} q_{ki}\right)^2$ $\xrightarrow{\text{odvod}}$ $\frac{\partial e_{ui}^2}{\partial p_{uk}} = -e_{ui} \cdot q_{ki}$

$\frac{\partial e_{ui}^2}{\partial q_{ki}} = -e_{ui} \cdot p_{uk}$

update:

$p_{uk} \leftarrow p_{uk} + \alpha e_{ui} q_{ki}$

$q_{ki} \leftarrow q_{ki} + \alpha e_{ui} \cdot p_{uk}$

vektorsko:

$p_u \leftarrow p_u + \alpha e_{ui} q_i$

$q_i \leftarrow q_i + \alpha e_{ui} p_u$

ISMF:

inkrementalna simultana matrična faktorizacija

(iskanje razcepa matrike na P in Q)

10. POVEZOVALNA PRAVILA

17/1/20

$\sigma(X)$ ... št. transakcij, ki jih vsebuje X

$s(X)$ ... delež podprtih transakcij ali delež transakcij, ki vsebujejo nabor X (↑ vseh)

$$s(X) = \frac{\sigma(X)}{|T|} = \frac{\sigma(X)}{N}$$

Teorem 1

$s(X) \geq minsupp \Rightarrow s(Y) \geq minsupp, \quad Y \subseteq X$

Teorem 2

$s(X) < minsupp \Rightarrow s(Y) < minsupp, \quad X \subseteq Y$

Teorem 3: podpora nabora nikoli ne presega (↑ podpore) njegove podmnožice

$$\forall X, Y \in \mathcal{L} : X \subseteq Y \Rightarrow s(Y) \leq s(X)$$

Apriori algoritem

Povezovalna pravila

Podpora (delež transakcij, kjer najdemo vse stvari iz povezovalnega pravila (lže poznamo)): $s(X \to Y) = \frac{\sigma(X \cup Y)}{N}$ — minsupp

Zaupanje (delež transakcij, ki vsebujejo desno stran pravila Y med transakcijami, ki vsebujejo levo stran pravila X): $c(X \to Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$ — minconf