


BI-21-11

Lineární regrese, regularizace pomocou
hrebenej regresie

Líneárna regresia

- predpokladáme lin. závislosť vysvetľovanej premennej na príznakoch

$$Y = w_1 x_1 + \dots + w_p x_p + \varepsilon$$

↑
neznané koeficienty

↑
náhodná veličina
(nevysvetliteľná časť)

$$Y = w_0 + w_1 x_1 + \dots + w_p x_p + \varepsilon = w^T \bar{x} + \varepsilon$$

↑
intercept

$(w_0 \ w_1 \ \dots \ w_p) \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_p \end{pmatrix}$

$$E\varepsilon = 0 \Rightarrow EY = w^T x$$

Chyba modelu

ztrátová funkce : $L: \mathbb{R}^2 \rightarrow \mathbb{R}$ (loss function)

$$L(y, \hat{y}) = (y - \hat{y})^2$$

↑ ↑
skutočná hodnota predikcia

$$RSS(w) = \sum_{i=1}^N L(y_i, w^T x_i) = \sum_{i=1}^N (y_i - w^T x_i)^2$$

- minimalizácia súčtu chyb
residual sum of squares

Gradient

- ukazuje smer maximalného rústu fce

Bud $f: \mathbb{R}^d \rightarrow \mathbb{R}$ fce viacero promennych, kt. má v bode $a \in \mathbb{R}^d$ konečné všetky parciálne derivácie. Gradient fce f v bode a definujeme ako vektor

$$\nabla f(a) = \left(\frac{\partial f}{\partial x_1}(a), \dots, \frac{\partial f}{\partial x_d}(a) \right)$$

∇f - zobrazenie, kt. každému bodu, kde to lze, priradí gradient v tomto bode

Hessova matice

$$f: \mathbb{R}^d \rightarrow \mathbb{R}$$

Hessova matice fce f v bode $a \in \mathbb{R}^d$ definujeme ako

$$H_f(a) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_i^2}(a) & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_d}(a) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_d \partial x_1}(a) & \cdots & \frac{\partial^2 f}{\partial x_d^2}(a) \end{pmatrix}$$

$$\text{kde } \frac{\partial^2 f}{\partial x_i \partial x_j}(a) = \left(\frac{\partial}{\partial x_i} \left(\frac{\partial f}{\partial x_j} \right) \right)(a)$$

značí druhú parciálnu deriváciu po x_i, x_j

Budú f: $\mathbb{R}^d \rightarrow \mathbb{R}$ fce d premenlivých a bod $x^* \in \mathbb{R}^d$ taký, že $\nabla f(x^*) = 0$

- Ak $s^T H_f(x^*) s > 0$ pre všetky $s \in \mathbb{R}^d$, $s \neq 0$

nabíja fce f v bode x^* ostrého lok. minima

- Ak pre každé x z nejakého okolia x^*

$$s^T H_f(x^*) s \geq 0 \quad \text{pre všetky } s \in \mathbb{R}^d$$

nabíja fce f v bode $x^* \in \mathbb{R}^d$ neostrého lokálneho minima

↪ Pozitívny definitnosť / semi-definitnosť Hessovy matice H_f
 v bode x^* resp. x

$$X = \begin{pmatrix} X_1^T \\ \vdots \\ X_n^T \end{pmatrix} = \begin{pmatrix} 1 & X_{1,1} & X_{1,2} & \cdots & X_{1,p} \\ \vdots & \vdots & & & \vdots \\ 1 & X_{N,1} & X_{N,2} & \cdots & X_{N,p} \end{pmatrix}$$

$$\begin{aligned} w^T X_i &= (w_1, w_2, \dots, w_j, \dots, w_n) \begin{pmatrix} X_{i,1} \\ X_{i,2} \\ \vdots \\ X_{i,j} \\ \vdots \\ X_{i,n} \end{pmatrix} \\ &= w_1 x_{i,1} + w_2 x_{i,2} + \dots + w_j x_{i,j} + \dots + w_n x_{i,n} \end{aligned}$$

$$\begin{aligned} &= \cancel{w_1 x_{i,1}} + \cancel{w_2 x_{i,2}} + \cancel{w_j x_{i,j}} + \dots \\ &- \end{aligned}$$

$$\begin{aligned} &\dots - \cancel{w_n x_{i,n}} \\ &+ \cancel{w_1 x_{j,1}} + \cancel{w_2 x_{j,2}} - \dots \end{aligned}$$

$$RSS(w) = \sum_{i=1}^n (Y_i - w^T X_i)^2 = \|Y - Xw\|^2$$

$$\frac{\partial RSS}{\partial w_j} = \sum_{i=1}^n 2(Y_i - w^T X_i)(-x_{i,j}) \quad \left| \begin{array}{l} \nabla RSS = -\sum_{i=1}^n 2(Y_i - w^T X_i)x_i = -2X^T(Y - Xw) \\ \downarrow \nabla RSS = 0 \end{array} \right.$$

$$X^T Y - X^T X w = 0$$

$$\frac{\partial^2 f}{\partial w_k \partial w_j} = \sum_{i=1}^n 2(-x_{i,k})(-x_{i,j})$$

$$H_{RSS}(w) = 2X^T X$$

$$\forall s \in \mathbb{R}^{p+1} : s^T (X^T X) s = (X_s)^T (X_s) = \|X_s\|^2 \geq 0$$

$\Rightarrow H_{RSS}(w)$ je vždy pozitívne semi-definitné

Neostrieč lok. minimum nastáva v ktoromkoľvek bode w , kt.

$$\text{splňuje } X^T y - X^T X w = 0$$

ak $X^T X$ je regularná

$$\exists_1 w : \hat{w}_{OLS} = (X^T X)^{-1} X^T y$$

$$\hat{y} = \hat{w}_{OLS} X = X^T w_{OLS} = X^T (X^T X)^{-1} X^T y$$

Gradientný sestup

začнемe s $w^{(0)}$ a pomocou

$$\begin{aligned} w^{(i+1)} &= w^{(i)} - \alpha \cdot \nabla \text{RSS}(w^{(i)}) = \\ &= w^{(i)} - \alpha \cdot 2X^T(Y - Xw^{(i)}) \end{aligned}$$

postupne konštrukujeme postupnosť vektorov, o kt. dôfame, že konverguje k skutočnému riešeniu \hat{w}_{OLS}

α je learning-rate

Regularita vs. lineárna nezávislosť

- normální rovnice má jednoznačné riešenie, pokiaľ $X^T X$ je regul.

$X_{\cdot i}$ je i -ty stĺpec matice X

$$XS = S_0 X_{\cdot 0} + S_1 X_{\cdot 1} + \dots + S_p X_{\cdot p}$$

$$\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{pmatrix} \begin{pmatrix} s_1 \\ \vdots \\ s_n \end{pmatrix} = S_1 \begin{pmatrix} x_{11} \\ \vdots \\ x_{m1} \end{pmatrix} + S_2 \begin{pmatrix} x_{12} \\ \vdots \\ x_{m2} \end{pmatrix} + \dots + S_p \begin{pmatrix} x_{1p} \\ \vdots \\ x_{mp} \end{pmatrix}$$

$$= \begin{pmatrix} x_{11} s_1 + x_{12} s_2 + \dots \\ x_{m1} s_1 + x_{m2} s_2 + \dots \\ \vdots \\ \vdots \end{pmatrix}$$

$$XS = 0 \Rightarrow X^T X S = 0 \Rightarrow S^T X^T X S = 0 \Rightarrow \|X S\|^2 = 0 \Rightarrow X S = 0$$

$\boxed{X^T X \text{ je regulárne} \Leftrightarrow \text{sú stĺpce matice } X \text{ lin. nezávislé}}$

Problém ak $N < p+1 - v \mathbb{R}^N$ neexistuje $p+1$ LN vektorov
 $N \geq p+1 - \text{nemusí byť LN}$

$X^T Y - X^T X w = 0$ - sústava $p+1$ lin. rovnic s $p+1$ neznáimi
 - vždy aspoň 1 riešenie, ak sú stĺpce nezávislé,
 tak práve jedno $w_{OLS} = (X^T X)^{-1} X^T Y$
 v opačnom prípade sa riešení $X(w-w') = 0$

Ako hľať riešenie (nejaké), keď nemôžeme invertovať matice $X^T X$?

Nájsť taký vektor \hat{w} , ktorý rieši normálnu rovnice a zároveň má najmenšiu normu $\|\hat{w}\|$.

$$\hat{w} = (X^T X)^+ X^T Y$$



Moorova-Penroseova pseudoinverzna matice
 $k X^T X$

-problémom sú kolineárne príznaky („skoro“ lineárne závisle“)

$$\|X_u\| \gg \|X_v\| = 0 \text{ pre nejaké } \|u\| = \|v\| = 1$$

Odhad \hat{w}_{OLS} je citlivý na malej nevhodnej zmeny y
pri opakovacom učení sa môže odhad radikálne zmeniť

$\Rightarrow \hat{w}_{OLS}$ má veľký rozptyl

Regularizace lin.r.

— prígenerovať ďalšie data alebo odobrat existujúce a dodať, že sú to výrieši

znižiť počet príznakov

zmeneť funkciu, kt. minimalizujeme
- pridať regularizačný člen

Hrebienková regresia

$$RSS_{\lambda}(w) = \|y - Xw\|^2 + \lambda \sum_{i=1}^P w_i^2$$

intercept nepenalizujeme

$$\lambda = 0 \Rightarrow RSS_0(w) = RSS(w)$$

$\lambda > 0 \Rightarrow$ minimum sa bude cíliť na také vektory w , ktoré majú čo najmenšie zložky

$$\underline{I} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{p+1, p+1}$$

$$RSS_{\lambda}(w) = \|y - Xw\|^2 + \lambda \sum_{i=1}^p w_i^2 = \|y - Xw\|^2 + \lambda w^T \underline{I}' w$$

$$\nabla RSS_{\lambda}(w) = -2X^T(y - Xw) + 2\lambda \underline{I}' w$$

$$H_{RSS_{\lambda}}(w) = 2X^T X + 2\lambda \underline{I}'$$

$$\text{norm. r. } = X^T y - X^T X w - \lambda \underline{I}' w = 0$$

$\forall s \in \mathbb{R}^{p+1}, s \neq 0 \wedge \lambda > 0$ platí

$$s^T (X^T X + \lambda \underline{I}') s = (X_s)^T (X_s) + \lambda s^T \underline{I}' s = \|X_s\|^2 + \lambda \sum_{i=1}^p s_i^2 > 0$$

pre λ existuje vždy 1 riešenie

$$\hat{w}_{\lambda} = (X^T X + \lambda \underline{I}')^{-1} X^T y$$

$$\hat{y} = X \hat{w}_{\lambda}$$

Očekávaná chyba modelu

$$EL(\gamma, \hat{\gamma}) = E(\gamma - \hat{\gamma})^2 = E(\gamma - E\gamma)^2 + E(\hat{\gamma} - E\gamma)^2$$

$$\text{var } \gamma = \text{var } \varepsilon = \sigma^2$$

$$EL(\gamma, \hat{\gamma}) = \underbrace{\sigma^2}_{\substack{\text{heads tráni telina} \\ \text{chybov}}} + E(\hat{\gamma} - E\gamma)^2$$

$$\begin{aligned} \text{MSE}(\hat{\gamma}) &= E(\hat{\gamma} - E\gamma)^2 = \dots = (E\hat{\gamma} - E\gamma)^2 + \text{var } \hat{\gamma} = \\ &= (\text{bias } \hat{\gamma})^2 + \text{var } \hat{\gamma} \end{aligned}$$

bias $\hat{\gamma} = E\hat{\gamma} - E\gamma$ značí význam odhadu

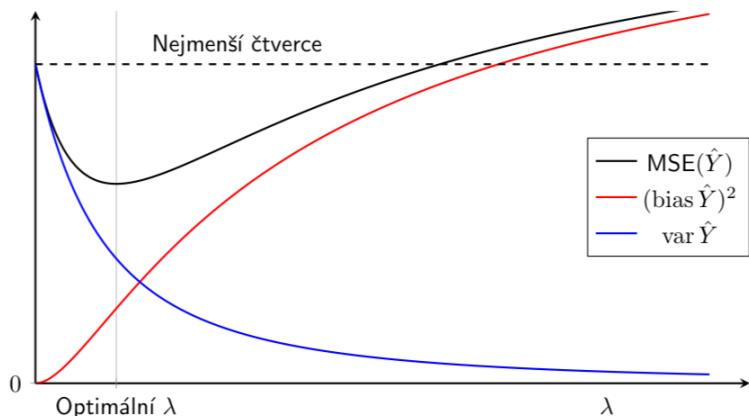
$$EL(\gamma, \hat{\gamma}) = \sigma^2 + (\text{bias } \hat{\gamma})^2 + \text{var } \hat{\gamma}$$

Bias-variance trade off

U hřebenové regrese lze ukázat, že (hodně zjednodušeně) platí

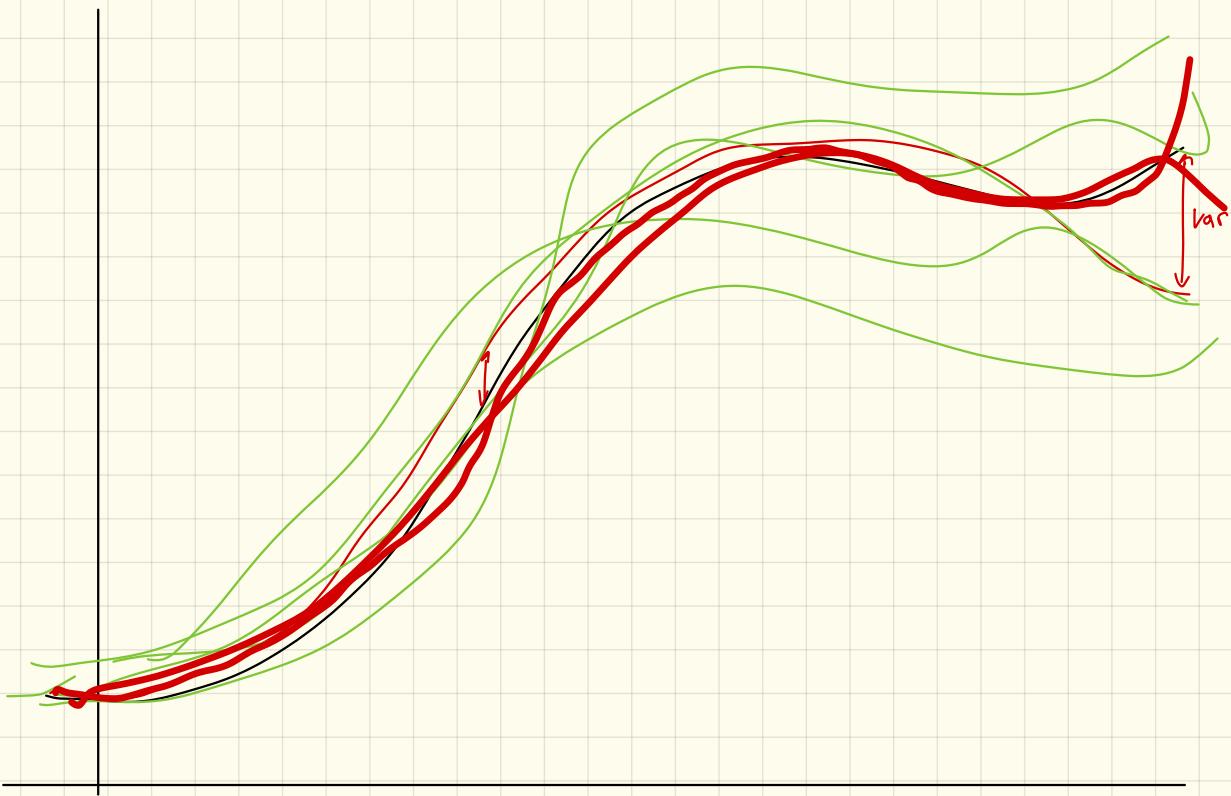
$$(\text{bias } \hat{Y})^2 \sim \left(1 - \frac{1}{1+\lambda}\right)^2 \quad \text{a} \quad \text{var } \hat{Y} \sim \left(\frac{1}{1+\lambda}\right)^2.$$

To znamená, že **s rostoucím λ vychýlení roste a rozptyl klesá**. Takovéto chování v závislosti na hyperparametrech modelu je typické a nazývá se *bias-variance tradeoff*.



$$MSE = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{X}_i^\top \hat{w}_\lambda)^2$$

-pred učením príznaky standardizovať, aby boli penalizované všetky rovnako



B1-21-12

Metoda nejbližších sousedů, metriky, metody
Shlukové analýzy (k-means, hierarchické shluk.)

Shlukováni

- nesupervizované učení (učení bez učitele)
- nemáme uprostredovací premenné

Clustering:

- blízke body budú v rovnakom žhluku
- vzdialé body budú v inom žhluku

Vzdialenosť (metrika):

dátoré body

ha množine X je funkcia $d: X \times X \rightarrow [0, +\infty)$

$\forall x, y \in X$ platí:

- $d(x, y) \geq 0$; $d(x, y) = 0 \Leftrightarrow x = y$ - pozitívna definítosť
- $d(x, y) = d(y, x)$ - symetria
- $d(x, y) \leq d(x, z) + d(z, y)$ - trojuholníková nerovnosť

Dvojice (X, d) - metrický prostor

$$d_2(x, y) = \sqrt{\sum_{i=1}^p (x_i - y_i)^2} - \text{Euklidova}$$

$$d_1(x, y) = \sum_{i=1}^p |x_i - y_i| - \text{Manhattanská}$$

$$d_\infty(x, y) = \max_i |x_i - y_i| - \text{Čebyševova}$$

Vstup:

- metrický priestor X so vzdialenosťou d
- množina dat $D \subset X$
- väčšinou aj počet zhlukov

Výstup:

- rozklad dat na zhluky $C = (C_1, \dots, C_k)$ $C_i \subset D$
 $C_i \cap C_j = \emptyset \quad i \neq j$

$$D = \bigcup_{i=1}^k C_i$$

Hierarchické zhlukovanie

- každý bod = zhluk

1. Nájdeme 2 zhluky, kt. majú k sebe najbližšie

2. Tieto 2 zhluky spojíme

Skončíme po $N-1$ opakovaniach s 1 zhlukom, kde sú všetky body

Vzdálenost žhlukov

Metoda nejbližšího souseda (single linkage) - dlhé řetazce

$$D(A, B) = \min_{x \in A, y \in B} d(x, y)$$

Metoda nejrizialenejšího souseda (complete linkage) - kompaktné žhluky

$$D(A, B) = \max_{x \in A, y \in B} d(x, y)$$

Párová vzdálenost (average linkage) - kompromis

$$D(A, B) = \frac{1}{|A||B|} \sum_{x \in A, y \in B} d(x, y)$$

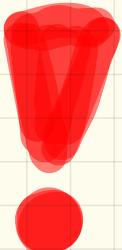
Kardiová metoda v \mathbb{R}^P - minimalizuje nárost vnitorného rozptylu

$$D(A, B) = \sum_{x \in A \cup B} \|x - \bar{x}_{A \cup B}\|^2 - \sum_{x \in A} \|x - \bar{x}_A\|^2 - \sum_{x \in B} \|x - \bar{x}_B\|^2$$

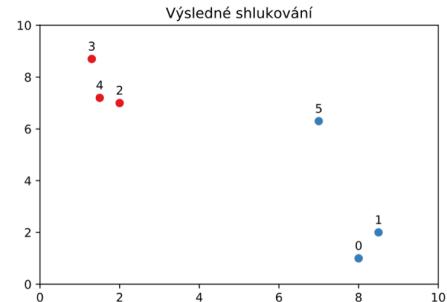
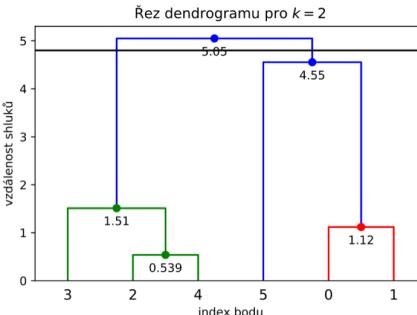
Ukončovacie podmienky:

- vzdialosť zhľukov
- počet zhľukov

- hypotetne háročné



Dendrogram



K-Means

- zhľukovanie ako optimalizačná úloha

účelová funkcia - snažíme sa minimalizovať

$$G(C) = \sum_{i=1}^k \frac{1}{|C_i|} \sum_{x_j \in C_i} d(x_j, y)^2 =$$

$\sqrt{(x_1 - y_1)^2 + \dots + (x_n - y_n)^2}$ Euklidova norma $\| \cdot \|_2$

$$= \sum_{i=1}^k \frac{1}{|C_i|} \sum_{x_j \in C_i} \|x_j - y\|^2$$

Hľadáme rozklad $C = (C_1, \dots, C_k)$ na priestore $X = \mathbb{R}^p$
vzbavenom Euklidovskou vzdialenosťou

$$\frac{1}{2|A|} \sum_{x_j \in A} \|x_j - y\|^2 = \sum_{x \in A} \|x - \bar{x}\|^2 = \min_{\mu \in \mathbb{R}^p} \sum_{x \in A} \|x - \mu\|^2$$

$$\bar{x} = \frac{1}{|A|} \sum_{x \in A} x \quad \text{-geometrický stred}$$

Algoritmus K-Means

Užitočná fce - $G(C) = \sum_{i=1}^k \sum_{x \in C_i} \|x - \bar{x}_i\|^2$

1. zafixujeme $\mu_i = \bar{x}_i$.

2. Vytvoríme nové zhluky $\bar{C} = \{\bar{C}_1, \dots, \bar{C}_K\}$ tak, že bude x presunuté do takého zhluku \bar{C}_i , v ktorom je vzdialosť $\|x - \mu_i\|^2$ najmenšia

$$\sum_{i=1}^k \sum_{x \in \bar{C}_i} \|x - \mu_i\|^2 \leq \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2$$

$$G(\bar{C}) = \sum_{i=1}^k \sum_{x \in \bar{C}_i} \|x - \bar{x}_i\|^2 \leq \sum_{i=1}^k \sum_{x \in C_i} \|x - \mu_i\|^2$$

$$G(\bar{C}) \leq G(C)$$

KNN

- učenie jednoduché, predikovanie je náročnejšie

trénovacie dátá sú naučený model

n-neighbours - počet susedov
metric - použitá metrika

weights - váhy najbližších susedov

hladanie susedov, určenie
indexovaním na hľadanie "učenia"

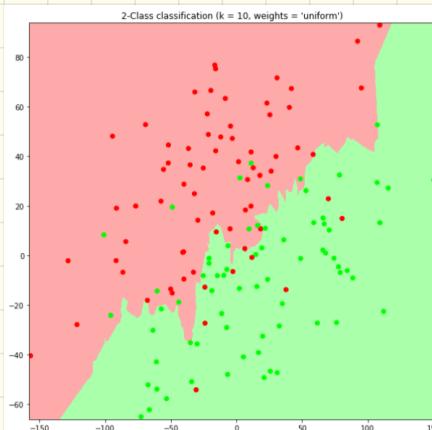
Minkovského k-metriky (L_k)

$$\|x - y\|_k = d_k(x, y) = \sqrt[k]{\sum_{i=0}^{p-1} |x_i - y_i|^k}$$

Normalizácia dat

do intervalu $[0, 1]$

$$x_i \leftarrow \frac{x_i - \min_x}{\max_x - \min_x}$$



BI - ŽI - 13

Naini Bayesuv klasifikátor, modely podmienených
pravdepodobností

Najvý Bayes

Klasifikácia na základe podmienenej pravdepodobnosti

$$\mathbf{X} = (X_1, X_2, \dots, X_p)^T$$

na základe train množiny odhadneme $P(Y=y | \mathbf{X}=\mathbf{x})$ pre $\mathbf{x} \in X, y \in Y$

predikcia $\hat{y} = \arg \max_{y \in Y} P(Y=y | \mathbf{X}=\mathbf{x})$
 \rightarrow maximum a posteriori (MAP)

Využitie Bayesovej vety

-ako odhadnúť $P(Y=y | \mathbf{X}=\mathbf{x})$?

$$P(Y=y | \mathbf{X}=\mathbf{x}) = \frac{P(\mathbf{X}=\mathbf{x} | Y=y) P(Y=y)}{P(\mathbf{X}=\mathbf{x})}$$

$$P(\mathbf{X}=\mathbf{x}) = \sum_{y \in Y} P(\mathbf{X}=\mathbf{x} | Y=y) P(Y=y)$$

(hľadáme maximum menovateľ môžeme zahodiť)

$$\Rightarrow P(Y=y | X=x) \propto P(X=x | Y=y) P(Y=y)$$

$$\Rightarrow \hat{Y} = \arg \max_{y \in Y} P(X=x | Y=y) P(Y=y)$$

Naive Bayes klasifikátor

predpoklad: Za podmienky $y=y$ sú všetky príznaky nezávislé

tj. $\forall y \in Y$ a $x = (x_1, \dots, x_p)^T \in X$ platí:

$$P(X=x | Y=y) = P(X_1=x_1 | Y=y) \cdot \dots \cdot P(X_n=x_n | Y=y)$$

$$\hat{Y} = \arg \max_{y \in Y} \prod_{i=1}^p P(X_i=x_i | Y=y)$$

- rezistentný proti problémom s dimenzionalitou

Bernoulli NB

$$\hat{P}_y = \frac{N_{1,y}}{N_{1,y} + N_{0,y}}$$

$N_{1,y}$ značí počet dát pre $X=1$ a $y=y$

\Rightarrow MLE odhad $\text{Be}(p)$

Bayesovský prístup

- pridáme apriórne rozdelenie

zmena nášho uvažovania na základe pozorovania

„expertné posúdenie situácie“

$$\mathbf{x} = (x_1, \dots, x_n)^T$$

$P(\mathbf{X}=\mathbf{x} | p)$ - pravd. že napozorujeme

$$f_p(p | \mathbf{x}) = \frac{P(\mathbf{X}=\mathbf{x} | p) f_p(p)}{P(\mathbf{X}=\mathbf{x})}$$

$\mathbf{X}=\mathbf{x}$, pokud p je správny parameter

$$P(\mathbf{X}=\mathbf{x}) = \int_p P(\mathbf{X}=\mathbf{x} | p) f_p(p) dp$$

vystriedaná P , že napozorujeme $\mathbf{X}=\mathbf{x}$

$$f_p(p) \propto p^{\alpha-1} (1-p)^{\beta-1} \quad -\text{Beta rozdelenie}$$

$\alpha = \beta = 1$ - rovnomeerne rozdelenie

$$\Rightarrow \hat{p}_y = \frac{N_{1,y} + 1}{N_{1,y} + N_{0,y} + 2} \quad \text{add-one smoothing / Laplace's rule of succession}$$

Kategorické rozdelenie

X nabýva k rôznych hodnôt c_1, \dots, c_k

$$\text{Cat}(p_y) \quad p_y = (p_{1,y}, \dots, p_{k,y})^T$$

$$P(X=c_j | Y=y) = p_{j,y}$$

$$\hat{p}_y = (\hat{p}_{1,y}, \dots, \hat{p}_{k,y})^T$$

$$\hat{p}_{j,y} = \frac{N_{j,y}}{N_{1,y} + \dots + N_{k,y}}$$

// j,y - počet dat pre $X=c_j$ a $y=y$

Bay. prístup

$$\hat{p}_{j,y} = \frac{N_{j,y} + 1}{N_{1,y} + \dots + N_{k,y} + k}$$

Spojité rozdelenie

$$\hat{y} = \arg \max_{y \in Y} \prod_{i=1}^l P(X=x_i | y=y) \prod_{i=l+1}^p f_{x_i | y}(x_i) P(y=y)$$

X_1, \dots, X_l sú diskrétné príznaky

X_{l+1}, \dots, X_p sú spojité

Lag-sum-exp trick

$$P(y=y | X=x) = \frac{P(X=x | y=y) P(y=y)}{\sum_{y \in Y} P(X=x | y=y) P(y=y)}$$

- numerické podtečenie

$P(X=x | y=y)$ - nízke hodnoty

$$\log P(y=y | X=x) = \log P(X=x | y=y) + \log P(y=y) - \log \sum_{y \in Y} P(X=x | y=y) P(y=y)$$

$$\log \sum_{\gamma \in \Gamma} P(X=x | \gamma = \gamma) P(\gamma = \gamma) = \log \sum_{\gamma \in \Gamma} e^{b_\gamma}$$

$$b_\gamma = \log P(X=x | \gamma = \gamma) + \log P(\gamma = \gamma)$$

$$\log \sum_{\gamma \in \Gamma} e^{b_\gamma} = \log \sum_j e^{B - b_j} = B + \log \sum_{j \in \Gamma} e^{b_j - B}$$

$$B = \max_{\gamma \in \Gamma} b_\gamma$$

Klasifikace textu

slovník - \mathcal{D}

dokument má $D = |\mathcal{D}|$ príznakov X_1, \dots, X_D

X_j - # výskytov j-teho slova $\in \mathcal{D}$

$$P(X=x | \gamma = \gamma) = \frac{n!}{\prod_{j=1}^D x_j!} \prod_{j=1}^D P_{j|\gamma}^{x_j}$$

$P_{j|\gamma}$ - P_j je náhodne vzaté slovo $\in \mathcal{D}$
tričky γ bude práve j-te slovo $\in \mathcal{D}$

$$\hat{P}_{j|\gamma} = \frac{N_{j|\gamma}}{N_\gamma} = \frac{\sum_{i=1}^n x_{i,j}}{\sum_{i=1}^n N_{i|\gamma}} \leftarrow \begin{array}{l} \text{počet výskytov j-th slova} \\ \text{v i-th dokumentu} \end{array}$$

$n = \prod_j x_j$ je počet slov v dokumentu

NLP

Jazyk - semanticko-syntaktické kontinuum

Syntax - spojovanie jazykových znakov (prípony, slova, vety) a vzťahy medzi nimi. Syntax je súčasť gramatiky. Napr. ako sú slová usporadávané vo vete.

Sémantika - významová stránka jazykových znakov.

Pragmatika - vzťah k ľudskejmu činiteľu a celej komunikatívnej situácii.
Napr. je rozdiel, keď na niekoho kričíme alebo mu píšeme list.

- v rámci ML sa sotva darí zachytiať syntaktickú stránku jazykov

Korpus

- prosté spojenie mnoho textu staženého z internetu alebo pečlivé vybraných textov, kt. sú doplnené rôznymi, klarne syntaktickými informáciami

ÚFAL (Ústav formální a aplikované lingvistiky) MFF UK

Český národný korpus

Slovenský projekt Araheo

Bag-of-words

- súbor dokumentov $D = \{d_1, d_2, \dots\}$

- každý dokument len ako množina slov, bez poradia

Korpus D reprezentujeme ako maticu:

- jeden riadok odpovedá jednému dokumentu

- stĺpcov je toľko, kolko je rôznych slov v celom korpusse

- na i-tom riadku (dokument d_i) a j-tom stĺpco (slovo w_j) je počet výskytov slova w_j v dokumente d_i .

d_1 : Toto je snad jediná věta ?!

d_2 : Je toto také jediná věta? To je věc.

	toto	je	snad	jediná	věta	také	to	věc
d_1	1	1	1	1	1	0	0	0
d_2	1	2	0	1	1	1	1	1

Problém: 1) Obrovská matice aj pre malé súbory textu

2) Stop slova, netradičné slova môžu byť silne podreprezentované

3) Matice a vektory repr. dokumenty sú veľmi riedke

Stop words

- byť, mať, spojky, predložky, barliky, --

- zoznamy stop words, odstrániť tie, kt. sa vyskytujú vo viac ako x% dokumentov

nevhodné pri clusterovaní dokumentov

Lemmatizace

- n flektívnych jazyku
- prevod slov do základného tvára
- otágované korpusy promocií promoce NNFSG -- 1, - -
 ↓ ↓ ↓
 žr sg. c.p

Tf - idf

term frequency - inverse document frequency

- akú vahu má slovo v danom dokumente

tf - miera ako často sa slovo vyskytuje v dokumente

w - slovo
d - dokument

$$tf(w, d) = \begin{cases} 1 & \text{ak } w \in d \\ 0 & \text{inak} \end{cases}$$

$$tf(w, d) = f_{w,d} = \# \text{výskytov } w \text{ v } d$$

$$tf(w, d) = \log(1 + f_{w,d})$$

$$tf(w, d) = \frac{f_{w,d}}{\sum_{w \in d} f_{w,d}} = \frac{f_{w,d}}{\text{počet slov v korpusu}}$$

$$tf-idf(w, d) = tf(w, d) \cdot idf(w, d)$$

idf - miera ako je slovo v korpusu obyklé

|D| - počet dokumentov

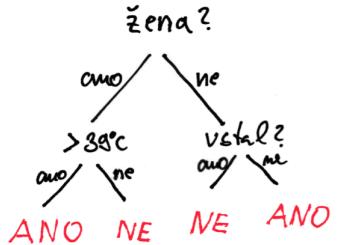
$$idf(w, d) = \log \frac{|D|}{|\{d \in D : w \in d\}|}$$

$$idf(w, d) = \log \frac{|D|}{1 + |\{d \in D : w \in d\}|}$$

B1-Z1-1L

Rozhodovací stromy, algoritmus konstrukce,
hyperparametry, náhodné stromy

Rozhodovacie stromy



rýmička	pohlaví	> 39°C	vstal(a)?	co říká strom
ano	muž	ne	ne	ano
ne	žena	ano	ano	ano
ne	muž	ne	ano	ne
ano	žena	ano	ne	ano



rýmička	pohlaví	> 39°C	vstal(a)?	co říká strom
ano	muž	ne	ne	ano
ne	žena	ano	ano	ne
ne	muž	ne	ano	ne
ano	žena	ano	ne	ano

Algoritmus ID3:

Množina 0 a 1 - 2

Změrař usporiadanosť - Entropie

$$\begin{aligned}
 H(\mathcal{D}) &= -P_0 \log P_0 - P_1 \log P_1 \\
 &= -\sum_{i=0}^{k-1} P_i \log P_i \\
 &= \frac{\#0}{\#1} - \text{pomer } 0 : 1
 \end{aligned}$$

Checeme vybrať príznak, kt. najviac zníži neusporiadanosť

Informačný zisk:

$$IG(\mathcal{D}, X_i) = H(\mathcal{D}) - t_0 H(\mathcal{D}_0) - t_1 H(\mathcal{D}_1)$$

$\mathcal{D}_0, \mathcal{D}_1$ - podmnožiny \mathcal{D}

$$t_i = \frac{\#\mathcal{D}_i}{\#\mathcal{D}}$$

Gini Index:

$$GI(\mathcal{D}) = 1 - \sum_{i=0}^{k-1} P_i^2 - \text{miera, že novopridaný prvok bude ťažko klasifikovať}$$

One-hot encoding

pôvodný príznak	→ dummy príznaky →	vstal	nevstal	spadl
vstal	→	1	0	0
nevstal	→	0	1	0
spadl	→	0	0	1

dummy variables

kategóriálne príznaky ↗ nominalné - miesto narodenia, fakulta, pohlavie
ordinalné - vzdelanie (základné < stredné...)

Regressný strom:

- priemer z hodnôt rovýskenu naom liste

$$MSE(\bar{y}) = \frac{1}{N} \sum_{i=0}^{N-1} (y_i - \bar{y})^2$$

CART

$$MSE(\bar{y}) = \bar{\epsilon}_L MSE(\bar{y}_L) + \bar{\epsilon}_R MSE(\bar{y}_R)$$

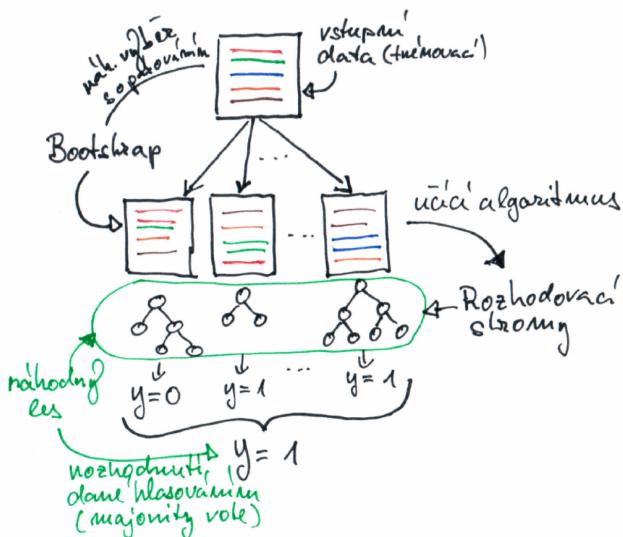
Ensemble metody

Bagging - bootstrap aggregating

Boosting

Bootstrap - výber s opakováním

Random forest klasifikátor (bagging)



n-estimators - počet stromov
max_depth - max hloubka stromu

- robustnost - odolné vůči přečlenům
- weak learners

sample_weight

t_L a t_R v $IG(\bar{d})$ sa prepožítať ako $\frac{\text{"suma väč v } \bar{d}_L\text{"}}{\text{"suma väč v } \bar{d}\text{"}}$

- strom sa učí tak, aby predikoval správne hlavne dátové body s vyššou vähou

AdaBoost (boosting)

- pri konštrukcii n -teho stromu je zvýšená väčšina týchto bodov, ktorí $(n-1)$ -ty strom klasifikoval zle

n -estimators - počet stromov

Algoritmus:

1. nastavia sa rovnomerne vähy $w_i = \frac{1}{N}, m=1$
 2. if $m \leq n$.estimators, nauč strom T^m na dátach \bar{d} s vähami w_i
 3. Do e^m ulož súčet väč $\geq \bar{d}$, kt. boli ťažké klasifikované stromom T^m
 4. If $e^m = 0$ alebo $e^m \geq \frac{1}{2}$, skonči
 5. pre stromom T^m ťažké klasifikované body nastav nové vähy $w_i \leftarrow \frac{1-e^m}{e^m} w_i$
 6. znormalizuj vähy, tak aby $\sum w_i = 1$
 7. $m++$, späť do bodu 2.
- $0 < e^m < \frac{1}{2}$, takže vähy
sú zvýšené!

Predikcia

1. každému stromu T^m priradíme váhu

$$w_T^m = \text{learning_rate} \cdot \log \frac{1 - e^m}{e^m}$$

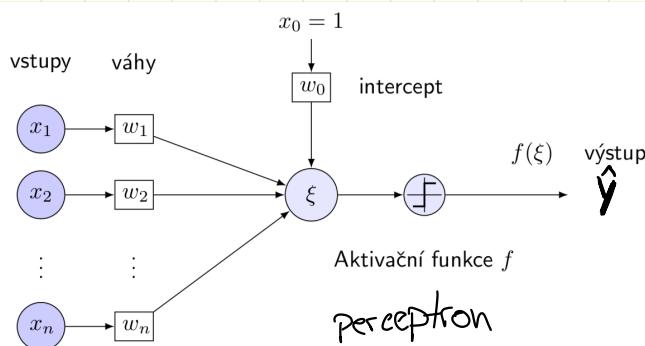
2. sečti váhy w_T^m , kt. predikujú $y=1$ a $y=0$

3. Vyber možnosť, kde je súčet väčší

B1-21-15

Neurónové siete, struktura perceptronovej
siete, výpočet výstupu neuronu, učení perc. sítie

Neuronové siete



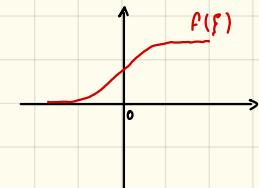
Vnitřní potenciál:

$$\xi = w_0 + \sum_{i=1}^N w_i x_i = w^T x + w_0$$

$$f(\xi) = \begin{cases} 1 & \text{kde } f(\xi) \geq 0 \\ 0 & \text{kde } f(\xi) < 0 \end{cases}$$

$$f(\xi) = 1 \Leftrightarrow \sum_{i=1}^N w_i x_i \geq -w_0$$

prahová hodnota



Update váh:

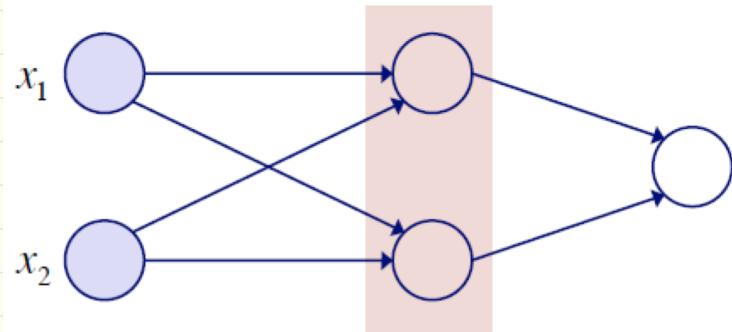
$$\delta w_i = \eta (\gamma - \hat{y}) x_i \quad \hat{y} = f(w^T x)$$

$$w_i \leftarrow w_i + \delta w_i$$

Výstup j-tého neuronu v i-tej vrstvě:

$$g_j^{(i)}: \mathbb{R}^{n_{i-1}} \rightarrow \mathbb{R}$$

n_{i-1} - počet neuronov v $(i-1)$ -tej vrstvě



Vícevrstvá síť s jednou skrytou vrstvou (růžově).

- NN s jednou skrytou vrstvou dokáže approximovať ľubovoľnú fce spojiteľ s kompaktným nosičom v $\mathbb{R}^n \rightarrow$ praxi nepoužiteľné
- používame radšej hlbšie siete, ktoré vytrárajú sofistikovanejšie príznaky

Aktivačné funkcie:

- zaručujú diferencovateľnosť

Logistická fce (sigmoida) - výstup je pravdepodobnosť

$$f(\xi) = \frac{1}{1 + e^{-\xi}}$$

Hyperbolický tangens

$$f(\xi) = \tanh(\xi) = \frac{e^\xi - e^{-\xi}}{e^\xi + e^{-\xi}}$$

Oriģinálna fce (ReLU)

$$f(\xi) = \max(0, \xi) = \begin{cases} x & \text{pre } \xi \geq 0 \\ 0 & \text{pre } \xi < 0 \end{cases}$$

Učenie gradientným sestupom

$$\text{minimalizujeme: } J(w) = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 = \frac{1}{N} \sum_{i=1}^N (y_i - g(x_i))^2$$

Batch training

- Máme N s parametry $w = (w_1, \dots, w_m)^T$ a trénovacie data $(y_1, x_1), \dots, (y_N, x_N)$
- inicializujeme všetky vähy w ako male náhodné čísla

Opakujeme, pokiaľ nie sú splnené zastavovacie kritéria:

1. položime $J(w) = 0$

2. pre trénovacie dvojice (y_i, x_i) :

i) spočítajme \hat{y}_i v bode x_i

ii) provedeme prepočet celkovej chyby

$$J(w) \leftarrow J(w) + \frac{1}{N} (y_i - \hat{y}_i)^2$$

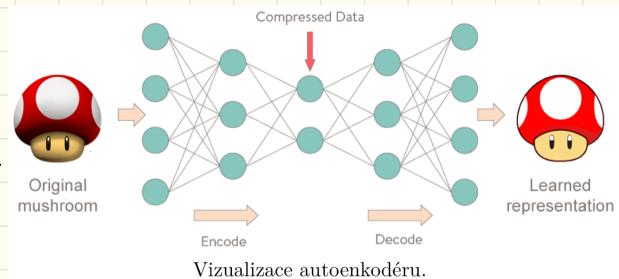
iii) spočítame gradient

$$\nabla_w J = \left(\frac{\partial J}{\partial w_1}, \dots, \frac{\partial J}{\partial w_m} \right)^T$$

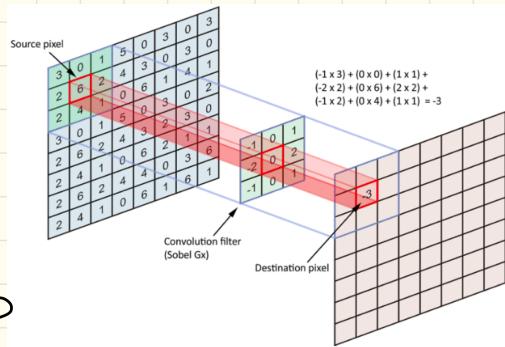
iv) provedeme prepočet väz $w \leftarrow w - \eta \nabla_w J$

Autoenkodéry

- enkodér, dekodér
- komprezia dat, redukcia dimenziality,
- detekcia odlehlých hodnot



| - cieľom je vytvoriť model, ktorý v časti enkodéra vytvorí abstrakčné príznaky (kód), ktoré sa v dekodériu dokážu vrátiť späť do pôvodného priestoru príznakov a replikovať čo najvernejšie vstupné dátá



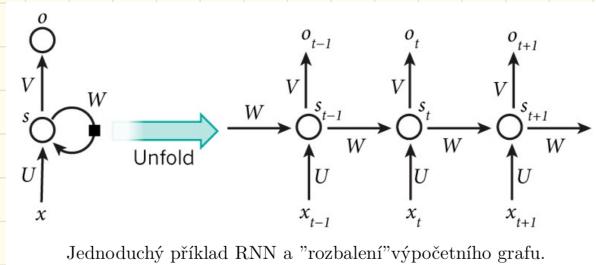
Príklad výpočtu konvolúcie medzi susednými vrstvami CNN.

Konvolučné NN (CNN)

- klasifikácia obrázkov
- diskrétné konvolúcie vstupov - NN sa sama učí vhodné konvolúcie

Rekurentné NN (RNN)

- pri výpočte výstupa sa okrem aktuálnych vstupov používajú aj predchádzajúce
- umožňuje pochytia časovej závislosti: : spracovanie textu, časové radov, skladanie hudby



BI-21-1C

Logistická regrese

Logistická regrese

- klasifikace

- diskrétny → spojity

$$Y=1 \rightarrow P(Y=1)$$

$$Y=0 \rightarrow P(Y=0)$$



$$\epsilon \{0, 1\}$$



$$\epsilon [0, 1]$$

$$P(Y=1|x, w)$$

$$x = (x_1, x_2, x_3)$$

$$w = (w_0, w_1, w_2, w_3)$$

$$P(Y=0|x, w) = 1 - P(Y=1|x, w)$$

Cislo $w_0 + w_1 x_1 + \dots + w_p x_p$ musíme dosadiť do fce, ktorej obor hodnôt je podmnožinou intervalu $[0, 1]$

sigmoida

$$f(x) = \frac{e^x}{1+e^x}$$

$$f(0) = \frac{1}{2}$$

$$D_f = \mathbb{R}$$

$$H_f = (0, 1)$$

$$f(x) = P(Y=1|x, w) =$$

$$= \frac{e^{wx}}{1+e^{wx}}$$

AK $P(Y=1|x_i, w) > \frac{1}{2} \Rightarrow Y=1$
 $< \frac{1}{2} \Rightarrow Y=0$

$P(Y=1|x_i, w) = \frac{1}{2}$ hranice rozhodnutia
 \Downarrow

$$w_0 + w_1 x_1 + \dots + w_p x_p = 0$$

MLE odhad w

$$P_1 = P(Y=1|x_i, w) = \frac{e^{w^T x}}{1+e^{w^T x}}$$

$P_{Y_i}(x_i, w)$ - i-ty datový bod s hodnotou vysčítanou proměnné
 y_i a hodnotami $x_i = (x_0, x_1, \dots, x_p)$

$$L(w) = \prod_{i=1}^N P_{Y_i}(x_i, w)$$

maximalizujeme

II - predpokladáme
 nezávislost datových
 bodov

$$l(w) \approx \ln L(w) = \sum_{i=1}^n \ln p_{y_i}(x_i, w) =$$

$$= \sum_{i=1}^n \left(Y_i \ln \left(\frac{e^{w^T x}}{1+e^{w^T x}} \right) + (1-Y_i) \ln \left(\frac{1}{1+e^{w^T x}} \right) \right)$$

$$\ln \frac{a}{b} = \ln a - \ln b$$

$$\ln 0 = 1$$

$$= \sum_{i=1}^n \left(Y_i w^T x - \ln (1+e^{w^T x}) \right)$$

$$Y_i = \begin{cases} 0 & \rightarrow (1-0) \ln \left(\frac{1}{1+e^{w^T x}} \right) \\ 1 & \rightarrow 1 \cdot \ln \frac{e^{w^T x}}{1+e^{w^T x}} \end{cases}$$

$$Y_i \cdot e^{w^T x} - \ln (1+e^{w^T x})$$

$$\frac{\partial l}{\partial w_j}(w) = \sum_{i=1}^n (X_{i,j} (Y_i - P_1(x_i, w))) \quad j = 0, 1, \dots, P$$

$$\nabla l(w) = X^T(Y - P)$$

$$P = (P_1(x_1, w), P_1(x_2, w), \dots, P_1(x_n, w))^T$$

$$w^T x = w_1 x_{11} + w_2 x_{12} + \dots + w_j x_{ij} + \dots + w_n x_{nn} +$$

$$+ w_1 x_{21} + w_2 x_{22} + \dots + w_j x_{2j} + \dots + w_n x_{2n} +$$

$$+ \dots$$

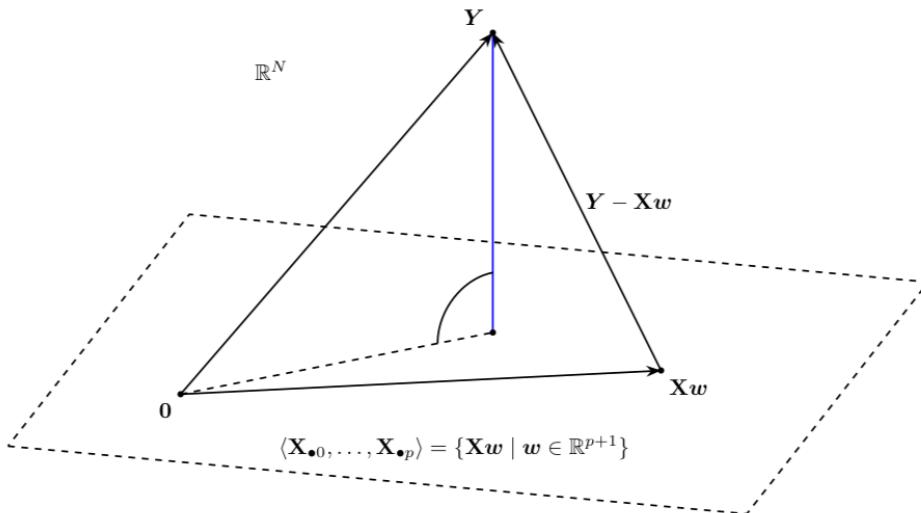
$$+ w_1 x_{j1} + w_2 x_{j2} + \dots + w_j x_{jj} + \dots + w_n x_{jn} +$$

$$+ w_1 x_{n1} + \dots + w_j x_{nj} + \dots + w_n x_{nn}$$

$$\frac{\partial l}{\partial w_j} = \sum_{i=1}^n x_{ij}$$

$$\ln \left(\frac{e^{w^T x_i}}{1+e^{w^T x_i}} \right) = \frac{e^{w^T x_i}}{1+e^{w^T x_i}} = P_1(x_i, w)$$

Geometrická interpretace metody najm. štvorců



Geometrická interpretace metody nejmenších čtverců (3/3)

- Bod $\mathbf{X}\mathbf{w}$ je k bodu \mathbf{Y} nejbližší, jestliže je vektor $\mathbf{Y} - \mathbf{X}\mathbf{w}$ na ten podprostor kolmý.
- To znamená, že je kolmý na všechny vektory $\mathbf{X}_{\bullet 0}, \dots, \mathbf{X}_{\bullet p}$, které ho generují:

$$(\mathbf{X}_{\bullet i})^T (\mathbf{Y} - \mathbf{X}\mathbf{w}) = 0 \quad \text{pro všechny } i = 0, \dots, p.$$

- To lze maticově zapsat jako

$$\mathbf{X}^T (\mathbf{Y} - \mathbf{X}\mathbf{w}) = \mathbf{0} \quad \text{a tedy} \quad \mathbf{X}^T \mathbf{Y} - \mathbf{X}^T \mathbf{X}\mathbf{w} = \mathbf{0}.$$

