12. Lineární klasifikace do více tříd, způsoby učení, softmax.

Lineární klasifikace do více tříd

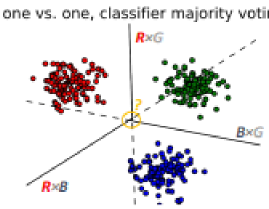
Lineární klasifikace do více než dvou tříd

- A: s využitím trénování **1 vs 1** a majoritním hlasováním
- B: s využitím trénování **1 vs REST (ALL) a majoritním hlasováním**
- C: s využitím trénování **1 vs REST (ALL) a výběrem dle maximálního skóre**
- D: s využitím funkce **SOFTMAX**

Způsoby učení

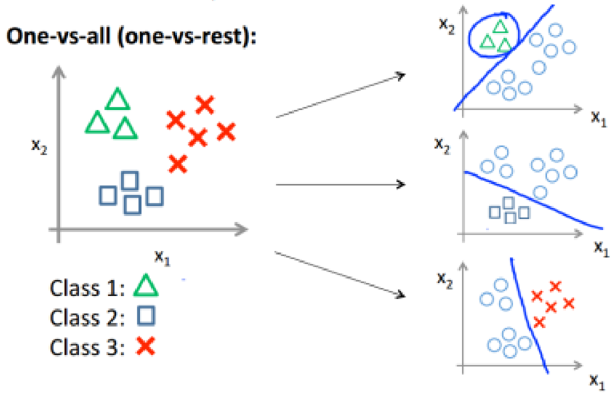
A: trénování 1 vs 1 a klasifikace většin. hlasováním

- Pro každé dvě třídy z celkem C tříd se postupně natrénuje příslušný binární klasifikátor
  - Celkem je třeba natrénovat $C(C-1)/2$ klasifikátorů = nevýhoda
- Klasifikace se pak provede pro každý z vytvořených klasifikátorů a bod $\boldsymbol{x}$ je klasifikován na základě nejvyššího počtu „hlasů“ od klasifikátorů → majority voting



B: Princip trénování (1 vs ALL)

Celkem C bin. klasifikátorů je natrénováno na úplné sadě dat – jedna třída reprezentována daty třídy, pro kterou je klasifikátor trénován, a druhá třída daty všech ostatních tříd:



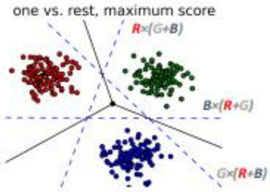
Možnost C

- V tomto případě jsou vytvářeny nové separující nadroviny
- Separující nadrovina mezi třídami i a j je dána rovnicí

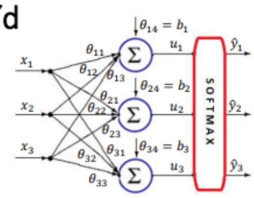
$$f_i(\boldsymbol{x}) = f_j(\boldsymbol{x})$$

- Bod $\boldsymbol{x}$ je klasifikován na základě nejvyššího hodnoty diskriminační funkce, tj. $\max\limits_k f_k(\boldsymbol{x})$

one vs. rest, maximum score



Lineární klasifikace do více tříd



- Elegantnější způsob klasifikace do více tříd spočívá ve vytvoření jednoho paralelního modelu
  - Model má C výstupů, kde C je počet tříd
  - Každý výstup je dán skalárním součinem vstupní vektoru a příslušného vektoru vah, na který je posléze aplikována funkce SOFTMAX
- Model pak nemá jeden vektor parametrů $\boldsymbol{\theta}$, ale matici parametrů $\boldsymbol{\Theta}$
- Příklad:
  - Vektor **x** má dimenzi 3 (máme tři příznaky)
  - Klasifikujeme shodou okolností také do 3 tříd

$$\boldsymbol{\Theta}^T = \begin{bmatrix} [\theta_{11}, \theta_{12}, \theta_{13}, \theta_{14} = b_1] \\ [\theta_{21}, \theta_{22}, \theta_{23}, \theta_{24} = b_2] \\ [\theta_{31}, \theta_{32}, \theta_{33}, \theta_{34} = b_3] \end{bmatrix}$$

Softmax

- Funkce SOFTMAX má C vstupů a C výstupů

Platí, že výstup $\hat{y}_c = SMAX(\boldsymbol{u}) = \dfrac{e^{u_c}}{\sum_{d=1}^{C} e^{u_d}}$

- Všechny výstupy jsou kladná čísla
- Součet všech výstupů dohromady je roven číslu 1

- Výsledkem klasifikace je třída, pro kterou je hodnota fce SOFTMAX nejvyšší
- Funkce má C výstupů a C vstupů
- Díky sumě ve jmenovateli každý výstup závisí na všech vstupech
- Celkem C výstupů tedy můžeme derivovat podle C různých vstupů
- Vznikne čtvercová matice parciálních derivací = Jacobiho matice = Jacobián

Zavedení pravděpodobnostnímu modelu

- K dispozici máme soubor trénovacích dat $\boldsymbol{X}$, celkem N vektorů
- Pro každý vektor $\boldsymbol{x}_i$ známe jeho příslušnost ke třídám: vektor $\boldsymbol{y}_i$
  - $\boldsymbol{y}$ obsahuje samé nuly a pouze jednu hodnot 1
- Zavedeme funkci $\pi_c(\boldsymbol{y}_i)$, která nabývá hodnoty jedna pouze pokud výstup $\boldsymbol{y}_i$ odpovídá třídě c
- Pak platí

$$\mathrm{P}(\boldsymbol{y}_i|\boldsymbol{\Theta}, \boldsymbol{x}_i) = \prod_{c=1}^{C} \hat{y}_{i,c}^{\pi_c(\boldsymbol{y}_i)}$$

- Protože $\boldsymbol{y}_i$ má jen jednu nenulovou hodnotu, uplatní se ze součinu pouze jeden člen
  - Ostatní jsou umocněny na hodnotu nula a jsou tedy rovny hodnotě jedna

- $\pi_1(\boldsymbol{y}_i) = 1$ pro $\boldsymbol{y}_i = [1,0,0]$, jinak 0 (=> správná třída je první třída)

- K nalezení parametrů modelu použijeme metodou MLE
- Všechny vektory $\boldsymbol{y}_i$ jsou nezávislé a ze stejného rozdělení:
  - Jejich sdružené rozdělení pravděpodobnosti je dáno součinem dílčích rozdělení:

$$\mathrm{P}(\boldsymbol{Y}|\boldsymbol{\Theta}, \boldsymbol{X}) = \prod_{i=1}^{N} \prod_{c=1}^{C} \hat{y}_{i,c}^{\pi_c(\boldsymbol{y}_i)}$$

- Logaritmus této pravděpodobnosti lze vyjádřit jako

$$log\mathrm{P}(\boldsymbol{Y}|\boldsymbol{\Theta}, \boldsymbol{X}) = \sum_{i=1}^{N} \sum_{c=1}^{C} \pi_c(\boldsymbol{y}_i) log(\hat{y}_{i,c}) = J(\boldsymbol{\Theta}, \boldsymbol{X})$$

Funkce $J(\boldsymbol{\Theta}, \boldsymbol{X}) = \sum_{i=1}^{N} \sum_{c=1}^{C} \pi_c(\boldsymbol{y}_i) log(\hat{y}_{i,c})$ představuje křížovou entropii (cross entropy)

- Stejně jako u logistické regrese

Minimalizovat křížovou entropii znamená minimalizovat rozdíl mezi dvěma rozděleními pravděpodobnosti

- Jedno reprezentuje skutečné rozdělení $\mathbf{y}_i$ pro trénovací data
- Druhé hodnoty $\hat{\boldsymbol{y}}_i$ predikované modelem s parametry $\boldsymbol{\Theta}$

Derivaci položíme rovnu nule a pro celý soubor dat dostaneme rovnici:

$$\sum_{i=1}^{N} \boldsymbol{x}_i(\pi_{a,i} - \hat{y}_{a,i}) = \sum_{i=1}^{N} \boldsymbol{x}_i(\pi_{a,i} - \hat{y}_{a,i}) = 0$$

Tato rovnice nemá pro $\boldsymbol{\theta}_a$ analytické řešení, $\hat{y}_{a,i}$ je složitá nelineární funkce $\boldsymbol{\theta}_a$

Maximum věrohodnosti lze nalézt pouze numericky, např. metodou SGD

Budeme hledat minimum záporné kriteriální funkce = maximu věrohodnosti

$$-J(\boldsymbol{\Theta}, \boldsymbol{X}) = -\sum_{i=1}^{N} \sum_{c=1}^{C} \pi_c(\boldsymbol{y}_i) log(\hat{y}_{i,c})$$

Porovnání dosud odvozených řešení pomocí SGD

- Lineární regrese

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha \sum_{i=1}^{N} \boldsymbol{x}_i (\mathbf{x}_i{}^T \boldsymbol{\theta} - y_i) \qquad y_i \in R$$

- Binární logistická regrese (= lineární klasifikace na dvě třídy):

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha \sum_{i=1}^{N} \boldsymbol{x}_i (\sigma(\mathbf{x}_i{}^T \boldsymbol{\theta}) - y_i) \qquad y_i \in \{0, 1\}$$

- Lineární klasifikace do více tříd s využitím funkce SOFTMAX:

$$\boldsymbol{\Theta}_{t+1} = \boldsymbol{\Theta}_t - \alpha \sum_{i=1}^{N} \boldsymbol{x}_i \left(\hat{\boldsymbol{y}}_i^T - \boldsymbol{\pi}_i{}^T\right) = \boldsymbol{\Theta}_t - \alpha \sum_{i=1}^{N} \boldsymbol{x}_i (SMAX(\mathbf{x}_i{}^T \boldsymbol{\Theta})^T - \boldsymbol{y}_i{}^T) \qquad \boldsymbol{y}_i \in [\dots 0; 0; 1; 0; \dots]$$