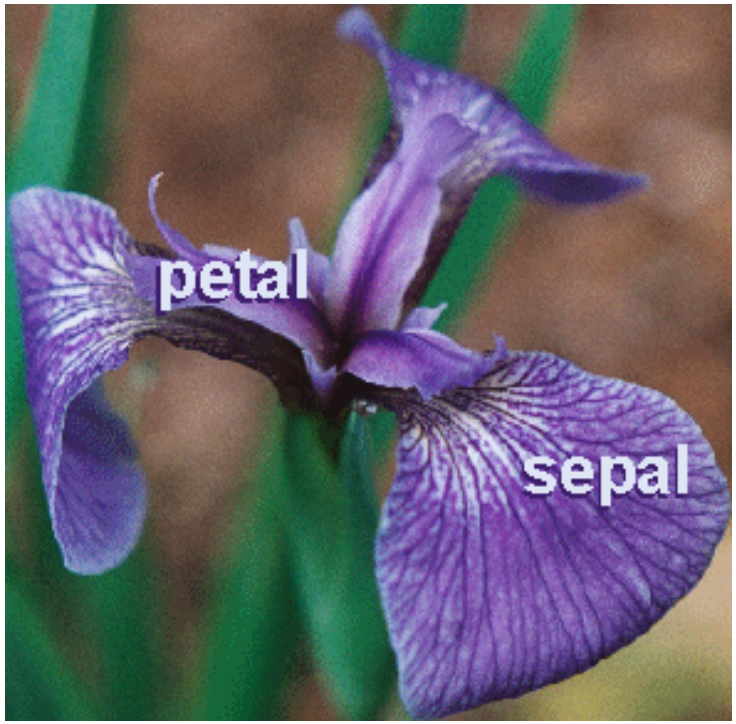
Vytěžování dat, cvičení 9: Strojové učení

Radomír Černoch, Michael Anděl, Miroslav Čepek

Evropský sociální fond
Praha & EU: Investujeme do vaší budoucnosti

Fakulta elektrotechnická, ČVUT

Iris data ještě jednou



Zmenšený dataset Iris

$sl =$ Sepal Length	$sw =$ Sepal Width	$pl =$ Petal Length	$pw =$ Petal Width	$c =$ Species
4.3	3.0	1.1	0.1	setosa
5.0	3.3	1.4	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.0	2.3	3.3	1.0	versicolor
5.7	2.8	4.1	1.3	versicolor
6.3	3.3	4.7	1.6	versicolor
4.9	2.5	4.5	1.7	virginica
6.2	2.8	4.8	1.8	virginica
6.8	3.2	5.9	2.3	virginica
7.7	3.0	6.1	2.3	virginica

Naïve Bayes

- NB klasifikace počítá pravděpodobnosti třídy, pokud známe hodnoty atributů. Např.
  $\mathrm{p}(c = \text{setoza} \mid sl = 5.5, sw = 3.3, pl = 4.8, pw = 9.9)$.
- Tuto pravděpodobnost neznáme přímo, pro její výpočet se používá Bayesovo pravidlo:

$$\mathrm{p}(c \mid sl, sw, pl, pw) = \frac{\mathrm{p}(c) \cdot \mathrm{p}(sl, sw, pl, pw \mid c)}{\mathrm{p}(sl, sw, pl, pw)} \tag{1}$$

- Protože distribuce $\mathrm{p}(sl, sw, pl, pw \mid c)$ má příliš mnoho parametrů, předpokládá se jejich nezávislost („naivita“ NB klasifikátoru):

$$\mathrm{p}(sl, sw, pl, pw \mid c) = \mathrm{p}(sl \mid c) \cdot \mathrm{p}(sw \mid c) \cdot \mathrm{p}(pl \mid c) \cdot \mathrm{p}(pw \mid c)\ . \tag{2}$$

Fáze 1: Učení

- Učení stanovuje parametry Naïve Bayes modelu = distribuce $\mathrm{p}(sl \mid c)$, $\mathrm{p}(sw \mid c)$, $\mathrm{p}(pl \mid c)$ a $\mathrm{p}(pw \mid c)$ a $\mathrm{p}(c)$.
- Distribuce $\mathrm{p}(c)$ říká, do kterého druhu patří libovolný květ kosatce, aniž bychom znali jakoukoli jeho vlastnost. Dá se říci, že tato distrubuce reprezentuje naše „předsudky“. **Vypočtěte** její parametry (předpokládejte *multinomialní rozdělení*).
- Ostatní určují rozložení jednotlivých atributů okvětních lístků pro každý druh kosatce zvlášť. Zachycuje tak vliv „měření“. **Vypočtěte** $\mathrm{p}(sl \mid c)$ (předpokládejte *normální rozdělení*).

Fáze 1: Učení

Správné výsledky

$$\mathrm{p}(sl \mid c) = \begin{cases} 3/10 & \text{pro } c = \text{setosa} \\ 3/10 & \text{pro } c = \text{versicolor} \\ 4/10 & \text{pro } c = \text{virginica} \end{cases} \tag{3}$$

$$\mathrm{p}(sl \mid c) = \begin{cases} \mathcal{N}(\mu = 5.0, \sigma = .70) & \text{pro } c = \text{setosa} \\ \mathcal{N}(\mu = 5.7, \sigma = .65) & \text{pro } c = \text{versicolor} \\ \mathcal{N}(\mu = 6.4, \sigma = 1.2) & \text{pro } c = \text{virginica} \end{cases} \tag{4}$$

Fáze 1: Učení

$$\mathrm{p}(sw \mid c) = \begin{cases} \mathcal{N}(\mu = 3.6, \sigma = .74) & \text{pro } c = \text{setosa} \\ \mathcal{N}(\mu = 2.8, \sigma = .50) & \text{pro } c = \text{versicolor} \\ \mathcal{N}(\mu = 2.9, \sigma = .30) & \text{pro } c = \text{virginica} \end{cases} \tag{5}$$

$$\mathrm{p}(pl \mid c) = \begin{cases} \mathcal{N}(\mu = 1.3, \sigma = .21) & \text{pro } c = \text{setosa} \\ \mathcal{N}(\mu = 4.0, \sigma = .70) & \text{pro } c = \text{versicolor} \\ \mathcal{N}(\mu = 5.3, \sigma = .79) & \text{pro } c = \text{virginica} \end{cases} \tag{6}$$

$$\mathrm{p}(pw \mid c) = \begin{cases} \mathcal{N}(\mu = .23, \sigma = .15) & \text{pro } c = \text{setosa} \\ \mathcal{N}(\mu = 1.3, \sigma = .30) & \text{pro } c = \text{versicolor} \\ \mathcal{N}(\mu = 2.0, \sigma = .32) & \text{pro } c = \text{virginica} \end{cases} \tag{7}$$

Fáze 2: Klasifikace

- Mějme okvětní lístek $sl = 5.5$, $sw = 2.4$, $pl = 3.7$ a $pw = 1.0$. O jaký druh kosatců se jedná? (Pro kontrolu: versicolor.)
- Výpočet je pouze zjednodušením vzorců z úvodu a dosazením: $\mathrm{p}(c \mid sl, sw, pl, pw) =$

$$= \frac{\mathrm{p}(c) \cdot \mathrm{p}(sl \mid c) \cdot \mathrm{p}(sw \mid c) \cdot \mathrm{p}(pl \mid c) \cdot \mathrm{p}(pw \mid c)}{\mathrm{p}(sl, sw, pl, pw)} = \quad (8)$$

$$= \frac{\psi(sl, sw, pl, pw, c)}{\mathrm{p}(sl, sw, pl, pw)} = \quad (9)$$

$$= \frac{\psi(sl, sw, pl, pw, c)}{\sum_c \mathrm{p}(c) \cdot \mathrm{p}(sl, sw, pl, pw \mid c)} = \quad (10)$$

$$= \underline{\underline{\frac{\psi(sl, sw, pl, pw, c)}{\sum_c \psi(sl, sw, pl, pw, c)}}} \quad (11)$$

Fáze 2: Klasifikace

$$\psi(sl = 5.5, sw = 2.4, pl = 1.4, pw = 1.1, c = \text{setoza}) =$$

$$\begin{aligned} = \mathrm{p}(c = \text{set.}) \cdot & \\ & \cdot \mathrm{p}(sl = 5.5 \,|\, c = \text{set.}) \cdot \mathrm{p}(sw = 2.4 \,|\, c = \text{set.}) \cdot \\ & \cdot \mathrm{p}(pl = 1.4 \,|\, c = \text{set.}) \cdot \mathrm{p}(pw = 1.1 \,|\, c = \text{set.}) = \end{aligned} \tag{12}$$

$$\boxed{\mathrm{p}(x \,|\, \mu, \sigma) = \frac{1}{\sqrt{2\,\pi\,\sigma^2}} \exp\left(\frac{x - \mu}{\sqrt{2\,\sigma^2}}\right)^2} \tag{13}$$

$$= \frac{3}{10} \cdot 0.44 \cdot 0.15 \cdot 1.8 \cdot 2.6 \cdot 10^{-7} = \underline{\underline{3.1 \cdot 10^{-8}}} \tag{14}$$

Fáze 2: Klasifikace

$$\psi(..., c = \text{versicolor}) = 5.5 \cdot 10^{-5};\ \psi(..., c = \text{virginica}) = 1.8 \cdot 10^{-9}$$

$$\begin{aligned} \text{p}(c = \text{setoza} \,|\, ...) &= \\ &= \frac{\psi(..., c = \text{versicolor})}{\psi(..., c = \text{set.}) + \psi(..., c = \text{ver.}) + \psi(..., c = \text{vir.})} = \\ &= 0.056\% \end{aligned} \tag{15}$$

$$\text{p}(c = \text{versicolor} \,|\, ...) = 99.94\% \tag{16}$$

$$\text{p}(c = \text{virginica} \,|\, ...) = 0.004\% \tag{17}$$

Bingo! Versicolor vyhrává.

Zadání úlohy

1. V prostření Matlab implementujte k-NN klasifikátor a otestujte jej na datasetu „Breast Cancer“ 2-fold krosvalidace[1]. Alternativně můžete validovat mode pomocí testovací sady (cca 1-3), kterou z data na začátku vydělíte.
2. Vzorky s chybějícími hodnotami (označené "?") můžete výjimečně zahodit.
3. Vyčíslete trénovací i tesovací chybi klasifikátoru pro hodnoty k od 1 do velikosti datasetu. Oba průběhy vyneste do grafu v závislosti na k.
4. Průběh grafu interpretujte s přihlédnutím k principu fungování algoritmu k-NN.

Zdrojové kódy odevzdejte do upload systému samostatně vedle PDF protokolu.

[1] viz. Wikipedie