

Vytěžování dat, cvičení 9: Strojové učení

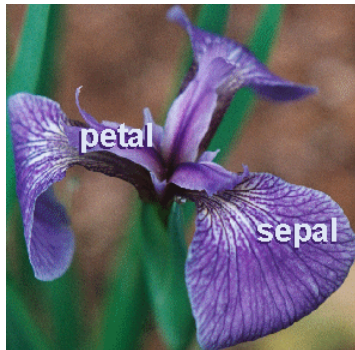
Michael Anděl, Miroslav Čepek, Radomír Černoch

November 18, 2014



Evropský sociální fond
Praha & EU: Investujeme do vaší budoucnosti

Fakulta elektrotechnická, ČVUT



Zmenšený dataset Iris

| <i>sl</i> = Sepal Length | <i>sw</i> = Sepal Width | <i>pl</i> = Petal Length | <i>pw</i> = Petal Width | <i>c</i> = 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 |

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

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

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

$$p(sl, sw, pl, pw \mid c) = p(sl \mid c) \cdot p(sw \mid c) \cdot p(pl \mid c) \cdot p(pw \mid c) \quad (2)$$

- ▶ Učení stanovuje parametry Naïve Bayes modelu = distribuce $p(s/l | c)$, $p(sw | c)$, $p(pl | c)$ a $p(pw | c)$ a $p(c)$.
- ▶ Distribuce $p(c)$ říká, do kterého druhu patří libovolný květ kosatce, aniž bychom znali jakoukoli jeho vlastnost. Dá se říci, že tato distribuce 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** $p(s/l | c)$ (předpokládejte *normální rozdělení*).

Správné výsledky

$$p(s|c) = \begin{cases} 3/10 & \text{pro } c = \text{setosa} \\ 3/10 & \text{pro } c = \text{versicolor} \\ 4/10 & \text{pro } c = \text{virginica} \end{cases} \quad (3)$$

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

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

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

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

- ▶ 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:
 $p(c | sl, sw, pl, pw) =$

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

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

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

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

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

$$= p(c = \text{set.}) \cdot$$

$$\cdot p(sl = 5.5 | c = \text{set.}) \cdot p(sw = 2.4 | c = \text{set.}) \cdot$$

$$\cdot p(pl = 1.4 | c = \text{set.}) \cdot p(pw = 1.1 | c = \text{set.}) = \quad (12)$$

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

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

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

$$p(c = \text{setosa} \mid \dots) =$$

$$= \frac{\psi(\dots, c = \text{versicolor})}{\psi(\dots, c = \text{set.}) + \psi(\dots, c = \text{ver.}) + \psi(\dots, c = \text{vir.})} = 0.056\% \quad (15)$$

$$p(c = \text{versicolor} \mid \dots) = 99.94\% \quad (16)$$

$$p(c = \text{virginica} \mid \dots) = 0.004\% \quad (17)$$

Bingo! Versicolor vyhrává.

1. V prostředí Matlab implementujte k -NN klasifikátor a otestujte jej na datasetu „Breast Cancer“ 2-fold krosvalidace¹.
2. Zjistěte chybu klasifikátoru zvláště na trénovací i testovací množině pro hodnoty k od 1 do velikosti datasetu. Oba průběhy vynesete do grafu v závislosti na k .
3. Průběh grafu interpretujte s přihlédnutím k principu fungování algoritmu k -NN.

Zdrojové kódy odevzdejte do upload systému samostatně vedle PDF protokolu.

¹viz. Wikipedie