Cvičenie 6

Nepleťte si Bayesov klasifikátor a Naivný Bayesov klasifikátor.

Bayesov klasifikátor

Pri klasifikácii je našou úlohou zatriediť príklady na základe známych vstupných atribútov do jednotlivých tried $c_1, c_2, \dots, c_K$. Najmenšiu chybu klasifikácie by sme dosiahli vtedy, ak by sme vedeli vypočítať podmienenú pravdepodobnosť $P(Y = c|X = \boldsymbol{x})$ pre každú triedu $y = c_1, c_2, \dots, c_K$ a príklad by sme zaradili do triedy, pre ktorú je $P(Y = c|X = \boldsymbol{x})$ najväčšia. Pravdepodobnosť $P(Y = c|X = \boldsymbol{x})$ však nepoznáme a k dispozícii máme iba ohraničenú množinu (výber) príkladov $(\boldsymbol{x}_i, y_i)$.

Podmienená pravdepodobnosť $P(Y = c|X = \boldsymbol{x})$ sa nedá dobre odhadnúť priamo na základe trénovacích dát, preto Bayesov klasifikátor používa na výpočet $P(Y = c|X = \boldsymbol{x})$ Bayesovo pravidlo:

$$P(Y = c|X = \boldsymbol{x}) = \frac{P(Y = c)P(X = \boldsymbol{x}|Y = c)}{P(X = \boldsymbol{x})}$$

Kde $P(Y = c|X = \boldsymbol{x})$ je podmienená pravdepodobnosť, že príklad patrí do triedy y ak má vstupné hodnoty $\boldsymbol{x}$, $P(Y = c)$ je pravdepodobnosť výskytu triedu y, $P(X = \boldsymbol{x})$ je pravdepodobnosť výskytu daného vektora vstupných hodnôt $\boldsymbol{x}$ (nezávisle od triedy) a $P(X = \boldsymbol{x}|Y = c)$ je pravdepodobnosť výskytu vstupných hodnôt $\boldsymbol{x}$ ak daný príklad patrí do triedy *c*.

Ak môže byť príklad zaradený iba do jednej triedy, pre pravdepodobnosť $P(X = \boldsymbol{x})$ platí:

$$P(X = \boldsymbol{x}) = \sum_{k=1}^{K} P(Y = c_k)P(X = \boldsymbol{x}|Y = c_k)$$

Keďže pri klasifikácii zaradíme príklad iba na základe porovnania $P(Y = c|X = \boldsymbol{x})$ a keďže pravdepodobnosť $P(X = \boldsymbol{x})$ nie je závislá od triedy, $P(X = \boldsymbol{x})$ nemusíme počítať a môžeme sa rozhodovať iba na základe výrazu:

$$P(Y = c)P(X = \boldsymbol{x}|Y = c)$$

Príklad s hodnotami $\boldsymbol{x}$ zaradíme do triedy y, pre ktorú bude výraz $P(Y = c)P(X = \boldsymbol{x}|Y = c)$ najväčší.

Pravdepodobnosť triedy $P(Y = c)$ vieme odhadnúť z dát priamo ako počet príkladov z triedy c / celkový počet príkladov. Pre $P(X = \boldsymbol{x}|Y = c)$ si musíme zvoliť nejaké rozdelenie pravdepodobnosti o ktorom predpokladáme, že bude dobre modelovať závislosť medzi priradením do tried a hodnotami $\boldsymbol{x}$. Ak sú všetky vstupné atribúty číselné, môžeme predpokladať napr. viacrozmerné normálne rozdelenie:

$$P(X = \boldsymbol{x}|Y = c) = f(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^M|\boldsymbol{C}|}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

Kde $\boldsymbol{C}$ je (symetrická, pozitívne definitná) kovariančná matica a $\boldsymbol{\mu}$ je vektor stredných hodnôt pre jednotlivé vstupné atribúty X_1, ..., X_M. Maticu $\boldsymbol{C}$ a vektor $\boldsymbol{\mu}$ potom odhadneme priamo z dát pre jednotlivé triedy.

Ako príklad si uvedieme postup pre klasifikáciu do dvoch tried (iris setosa a iris versicolor). Načítame si dáta a vytvoríme si podmnožiny pre jednotlivé triedy:

```
data(iris)
setosa = iris[iris$Species == "setosa",]
versicolor = iris[iris$Species == "versicolor",]

setosa = as.matrix(setosa[1:4])
versicolor = as.matrix(versicolor[1:4])
```

Vypočítame si $P(Y = y)$ pre obe triedy:

```
p_setosa = nrow(setosa)/(nrow(setosa)+nrow(versicolor))
p_versicolor =
nrow(versicolor)/(nrow(setosa)+nrow(versicolor))
```

Vypočítame si kovariančné matice a vektory stredov pre jednotlivé triedy:

```
C_setosa = cov(setosa)
C_versicolor = cov(versicolor)

mean_setosa = colMeans(setosa)
mean_versicolor = colMeans(versicolor)
```

Pre nový príklad s hodnotami *x* = (5.7, 2.8, 4.1, 1.3) vypočítame pravdepodobnosť $P(X = \boldsymbol{x}|Y = c)$ dosadením kovariančnej matice a vektora stredov pre jednotlivé triedy. Pre viachodnotové normálne rozdelenie môžeme pravdepodobnosť $P(X = \boldsymbol{x}|Y = c)$ vypočítať pomocou funkcie "dvnorm" z balíka "emdbook":

```
install.packages("emdbook")
library(emdbook)

x = c(5.7, 2.8, 4.1, 1.3)
p_x_setosa = dmvnorm(x, mean_setosa, C_setosa)
p_x_versicolor = dmvnorm(x, mean_versicolor, C_versicolor)

p_setosa * p_x_setosa
p_versicolor * p_x_versicolor
```

Keďže hodnota pre celkový výraz $P(Y = c)P(X = \boldsymbol{x}|Y = c)$ je oveľa väčšia pre triedu versicolor, príklad s hodnotami *x* = (5.7, 2.8, 4.1, 1.3) by sme zaradili do triedy versicolor.

Úlohy na cvičení

1. Načítajte si dáta iris a vypočítajte parametre Bayesovho klasifikátora pre všetky tri triedy setosa, versicolor a virginica
2. Klasifikujte všetky príklady, tzn. vypočítajte pre každý príklad z množiny iris pravdepodobnosti pre všetky tri triedy a určite do ktorej triedy by príklad patril. Vypočítajte celkovú presnosť klasifikácie (počet správne klasifikovaných príkladov / počet všetkých príkladov)
3. Nainštalujte si balík MASS a pomocou funkcie mvrnorm(n, means, C) si vygenerujte N = 500 príkladov s viachodnotovým normálnym rozdelením so stredmi means = c(2,3) a kovariančnou maticou C = matrix(c(9,6,6,16),2,2). Zobrazte vygenerované dáta na grafe. Postupne meňte hodnoty matice C, vygenerujte si nové dáta a pozorujte ako sa zmenia na grafe:
    a. Nastavte hodnoty mimo diagonály na 0.
    b. Nastavte hodnoty na diagonále na 10,10 a 2,2.
    c. Nastavte hodnoty mimo diagonály na -3
4. Pomocou funkcie mvrnorm si vygenerujte dve dátové množiny o veľkosti 100 a 200 príkladov pre triedy so stredmi (10,20) a (20,30) a kovariančnými maticami [(6,2), (2,8)] a [(6,-3), (-3,6)]. Vypočítajte parametre Bayesovho klasifikátora a klasifikujte príklad s hodnotami (15, 25).