Cvičenie 3

Lineárna regresia

Pri lineárnej regresii sa snažíme predpovedať hodnotu výstupného atribútu (tzv. závislej premennej) y na základe vstupných atribútov $\boldsymbol{x} = (x_0, x_1, \dots, x_M)$ (nezávislých premenných) pomocou lineárnej funkcie:

$$f(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_M x_M$$

kde $\beta_0, \beta_1, \dots, \beta_M$ sú číselné parametre/koeficienty (váhy), ktoré musíme vypočítať na základe trénovacích príkladov $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \dots, (\boldsymbol{x}_N, y_N)$ tak aby sa minimalizovala kvadratická chyba medzi skutočnou hodnotou y_i a predpovedanou hodnotou vypočítanou podľa $f(x_i)$:

$$RSS = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(\boldsymbol{x}_i))^2$$

Parametre $\beta_0, \beta_1, \dots, \beta_M$ môžeme vypočítať aj analyticky riešením sústavy rovníc, ktoré môžeme maticovo zapísať ako:

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{\beta}$$

V danej sústave je $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_M)^T$ vektor váh, $\boldsymbol{y} = (y_1, y_2, \dots, y_N)^T$ vektor s výstupnými hodnotami trénovacích príkladov a $\boldsymbol{X}$ je matica, kde prvý stĺpec sú samé 1, a ostatné stĺpce zodpovedajú vstupným atribútom trénovacích príkladov (tzn. každý riadok matice $\boldsymbol{X}$ bude zodpovedať jednému príkladu $(1, x_{i,0}, x_{i,1}, \dots, x_{i,M})$).

V R môžeme parametre lineárneho modelu vypočítať pomocou funkcie lm. Nasledujúci príklad načíta dátovú množinu Iris, vyfiltruje iba príklady z triedy setosa, a vypočíta lineárny model s jedným vstupným atribútom iris$Sepal.Length pre výstupnú hodnotu iris$Sepal.Width

```
data(iris)
setosa <- iris[iris$Species == "setosa",]
attach(setosa)
plot(Sepal.Length, Sepal.Width)

sepal_lm <- lm(Sepal.Width ~ Sepal.Length)
abline(sepal_lm)
```

V príklade je vykreslená regresná priamka, ktorá určuje predikovanú hodnotu $f(x)$. Ďalšie vstupné atribúty je možné pridať do modelu pomocou operátora +, napr. lm(Y ~ X1 + X2 + X3) je lineárny model, ktorý predikuje atribút Y na základe troch vstupných atribútov X1, X2 a X3.

Pomocou funkcie summary si môžeme vypísať základné štatistiky o rezíduách a parametroch modelu:

```
summary(sepal_lm)
```

Rozdiely medzi predikovanou hodnotu a skutočnou hodnotou y sa označujú ako rezíduá. V časti Residuals sú informácie o základných štatistikách rezíduí. Tieto rozdiely by mali byť v absolútnej hodnote čo najmenšie. Rezíduá pre všetky trénovacie príklady môžete získať pomocou funkcie residuals

```
residuals(sepal_lm)
```

Celkovú chybu (sumu štvorcov rezíduí – z čoho sa minimalizovanie kvadratickej chybovej funkcie označuje metóda najmenších štvorcov) môžete vypísať funkciou:

```
deviance(sepal_lm)
```

V prehľade viete ďalej zistiť hodnoty samotných parametrov v časti Coefficients. Parameter β_0 sa označuje ako Intercept, pretože určuje v ktorom bode bude regresná priamka pretínať os Y. Keďže ostatné parametre $\beta_1, \dots, \beta_M$ zodpovedajú jednotlivým vstupným atribútom, parametre sú označené názvom zodpovedajúceho atribútu, napr. Sepal.Length. To, ako je dôležitý atribút pre výslednú predikciu určuje absolútna veľkosť zodpovedajúceho parametra, ak sa parameter blíži k 0, atribút je možné z modelu vynechať bez toho aby sa výrazne zmenila predikcia. Významnosť atribútov sa testuje štatistickým testom ktorého hypotéza je, že sa zodpovedajúci parameter = 0. Čím menšia je táto pravdepodobnosť, tým dôležitejší je daný atribút (čo je označené vo výpise aj hviezdičkami, *** dôležitý atribút, ** stredne dôležitý, atď.)

Ak chcete vypočítať predikovanú hodnotu pre nové vstupné dáta, môžete použiť funkciu predict:

```
predict(sepal_lm, newdata = data.frame(Sepal.Length=6))
```

Pomocou predikt môžeme vypočítať aj hodnoty pre viac príkladov naraz.

Podľa variancie vstupných dát vieme pri lineárnej regresii odhadnúť aj varianciu parametrov, ktorá je uvedená ako štandardná odchýlka pri výpise summary pre každý parameter. Keďže parametre sú odhadované iba s určitou varianciou, aj samotná predikcia je len odhad pre ktorý môžeme určiť varianciu v akom rozsahu sa bude predikovaná hodnota najpravdepodobnejšie pohybovať. Tento interval si môžete vypísať pomocou funkcie predict:

```
predict(sepal_lm, newdata = data.frame(Sepal.Length=6),
interval="prediction")
```

Pomocou predict je potom možné zobraziť aj intervaly spoľahlivosti pre predikciu (v tomto intervale by sa mala s veľkou pravdepodobnosťou pohybovať skutočná hodnota). Najprv si vygenerujeme vstupné dáta (interval od 4 do 6 s krokom 0.25) a potom vypočítame predikciu aj s intervalom:

```
lengths <- seq(from=4, to=6, by=0.25)
```

```
predictions <- predict(sepal_lm, newdata =
data.frame(Sepal.Length=lengths), interval="prediction")
```

Hodnoty si zobrazíme v grafe ako interval pomocou horného a spodného ohraničenia:

```
plot(Sepal.Length, Sepal.Width)
lines(predictions[,1] ~ lengths, col=1)
lines(predictions[,2] ~ lengths, col=1, lty=2)
lines(predictions[,3] ~ lengths, col=1, lty=2)
```

Všimnite si, že sa pri krajných hodnotách 4 a 6 interval mierne rozširuje (tzn. pre krajné vstupné hodnoty môže byť predikovaná hodnota vypočítaná s väčšou varianciou).

Úlohy na cvičení

1. Načítajte dátovú množinu Iris, odfiltrujte príklady z triedy setosa a vytvorte lineárny model predikujúci atribút Petal.Width pre vstupný atribút Petal.Length. Vizualizujte dáta a regresnú priamku
2. Vypočítajte celkovú sumu kvadrátov rezíduí modelu z predošlého príkladu a korelačný koeficient medzi vstupným a výstupným atribútom. Zobrazte histogram rezíduí.
3. Vypočítajte predikciu a interval spoľahlivosti pre trénovacie dáta. Zobrazte graf s dátami, regresnou priamkou a intervalovými hranicami.
4. Z množiny iris odfiltrujte iba príklady z triedy versicolor, vypočítajte predikciu a intervalové odhady. Zobrazte graf s dátami, regresnou priamkou, a intervalovými hranicami.
5. Vytvorte rozšírený model, ktorý bude mať okrem Petal.Length na vstupe aj atribút Sepal.Width. Zistite, ktoré parametre modelu sú štatisticky významné, a porovnajte chybu rozšíreného modelu s predchádzajúcim príkladom.
6. Medzi vstupné atribúty môžete pridať aj faktory (nominálne atribúty). Ak má faktor napr. 3 hodnoty, setosa, versicolor a virginica, pri zahrnutí do modelu sa prvá hodnota setosa použije ako tzv. referenčná a pre zvyšné hodnoty versicolor a virginica sa vygenerujú binárne (tzv. kontrastné) atribúty, pre ktoré sa vypočítajú dodatočné parametre modelu. Načítajte celú množinu dát Iris a vytvorte model, ktorý bude predikovať Petal.Width na základe atribútov Petal.Length a Species. Zistite, ktoré z kontrastných hodnôt versicolor, alebo virginica je viac dôležitá pre predikciu.
7. Vygenerujte si náhodné vektory X1, X2 a X3 s 50 hodnotami podľa uniformného rozdelenia pravdepodobnosti z intervalu 1 až 10. Vytvorte vektor hodnôt Y = 0.5 + 2*X1 + 1.5 *X2 + X3. K vektoru Y pripočítajte náhodný šum s normálnym rozdelením s 0 strednou hodnotou a štandardnou odchýlkou 0.5. Vypočítajte parametre lineárneho modelu priamo analyticky z výrazu $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{\beta}$. Vypočítajte pre každý príklad predikciu a rezíduá.
8. Pridajte do predchádzajúceho modelu atribút X4, ktorého hodnoty sú vždy dvojnásobkom atribútu X1 (tzn. X4=2*X1). Vypočítajte znova analyticky parametre modelu (malo by dôjsť k chybe, pre maticu X by sa nemala dať vypočítať inverzná matica ak obsahuje lineárne závislé stĺpce). Vypočítajte korelačný koeficient medzi X1 a X4. Odstráňte atribút X1 a zvona vypočítajte parametre lineárneho modelu, porovnajte vypočítané parametre s predchádzajúcim príkladom.