Úvod od strojového učení
Ilustrace k přednášce

Barbora Hladká, Martin Holub
MFF UK Praha, 2015
http://ufal.mff.cuni.cz/course/npfl054

Chi-kvadrát test nezávislosti

Princip: Jsou-li dvě kategoriální proměnné statisticky nezávislé, pak hodnoty v kontingenční tabulce mají multinomické rozdělení, takže střední hodnota počtu výskytů dvojice (x,y) je rovna p(x)*p(y)*N, kde N je celkový počet pozorování.

Příklad

Máme 100 pozorování dvou diskrétních proměnných X a Y (soubor xy.100.csv).

- Jsou tyto proměnné statisticky nezávislé?
- Testujte pro hladinu spolehlivosti 90%, 95% a 99%.
- Vypočítejte hodnotu chí-kvadrát statistiky, kritické hodnoty a p-hodnotu.

Vzorové řešení v R

```
*************************
*** Independence test ***
*************************

* Task:
  There are 100 observations of variables X and Y.
  Are the variables statistically independent?
  Use the chi-square independence test.

---------------------------------------------------------------

# reading the data
> observations = read.table("xy.100.csv", head=T)
> str(observations)
'data.frame':    100 obs. of  2 variables:
 $ x: Factor w/ 3 levels "A","B","C": 1 3 3 2 2 1 3 2 3 3 ...
 $ y: Factor w/ 2 levels "No","Yes": 1 2 1 1 1 2 2 1 2 2 ...

N = nrow(observations)

# contingency table
> observed = table(observations$y, observations$x)
> observed

       A  B  C
  No  11 39  8
  Yes  5 21 16
```

```
# marginal distributions
> p.x = table(observations$x)/N
> p.y = table(observations$y)/N

> p.x
   A    B    C
0.16 0.60 0.24

> p.y
  No  Yes
0.58 0.42


# expected values (on assumption that X and Y are independent)
> expected = p.y %*% t(p.x) * N
> expected

         A    B     C
  No  9.28 34.8 13.92
  Yes 6.72 25.2 10.08


# chi-square statistic
> sum((observed - expected)^2 / expected)
[1] 7.960454


# p-value
> 1 - pchisq(7.960454, df=2)
[1] 0.0186814


# chi-square critical values
> qchisq(0.99, df=2)
[1] 9.21034
> qchisq(0.95, df=2)
[1] 5.991465
> qchisq(0.90, df=2)
[1] 4.60517



# the same by the built-in chisq.test()
> chisq.test(observed)
      Pearson's Chi-squared test

data:  observed
X-squared = 7.9605, df = 2, p-value = 0.01868


* Conclusion:
  The null hypothesis (that the variables are independent) cannot
  be rejected only at significance level 1%.
```

```
---------------------------------------------------------------

* Remark:
  In fact the data used in this exercise was randomly (and independently)
  generated by the following commands:
  > x = sample(c('A', 'B', 'C'), 100, replace = T, prob = c(20,50,30))
  > table(x)
      x
       A  B  C
      16 60 24

  > y = sample(c('Yes', 'No'), 100, replace = T, prob = c(40,60))
  > table(y)
      y
       No Yes
       58  42

  So, the result of the independence test is not surprising.

  In addition, we can also test if the generated samples are in line with
  the required distributions. Now, the variables X and Y will be tested
  separately.

*****************************
*** Goodness-of-fit tests ***
*****************************

> table(observations$x)
 A  B  C
16 60 24

> chisq.test(table(observations$x), p=c(0.2, 0.5, 0.3))
      Chi-squared test for given probabilities

data:  table(observations$x)
X-squared = 4, df = 2, p-value = 0.1353


> table(observations$y)
 No Yes
 58  42

> chisq.test(table(observations$y), p=c(0.6, 0.4))
      Chi-squared test for given probabilities

data:  table(observations$y)
X-squared = 0.1667, df = 1, p-value = 0.6831
```