UNIVERSITÀ DI ROMA TOR VERGATA

EEBL - Statistical Learning

Revision - week 4

1 Where to study

G James, D Witten, T Hastie, and R Tibshirani and J Friedman. *An Introduction to Statistical Learning with Applications in R.* 2nd ed. Springer, Springer Series in Statistics, 2021.

- Classification is dealt with in chapter 4 of the textbook. We have covered sections 4.1-4.2, 4.3 (logistic regression) up to page 139, sec. 4.4, up to page 154 (Discriminant analysis; the ROC curve will be discussed later on).

- It is not strictly needed, but if you want a more advanced treatment, see T Hastie, R Tibshirani and J Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition. Springer, Springer Series in Statistics, 2009.

  Website: http://www-stat.stanford.edu/ElemStatLearn/

  - Loss functions for classifications: pages 20 - 22.
  - Classification: chapter 4, read the introduction, sec 4.1.
  - Linear Discriminant Analysis: sec. 4.3 (up to page 112).
  - Logistic regression: section 4.1 and page 119.

2 Exercises and complements

1. Revise the use of discriminant analysis for classification. How do we use Bayes theorem for estimating the posterior probabilities? Review the illustration used during the lectures. The computations are in the script `SimulatedExample.R`.

   Let
   $$G = \begin{cases} 1 & \text{No Admission} \\ 0 & \text{Admission} \end{cases}$$

   Further, let X be the final mark obtained on an entry test.

   From a training sample you estimate the prior probabilities $\hat{\pi}_0 = 0.5, \hat{\pi}_1 = 0.5$, and $X|G = 0 \sim N(80, 4)$, whereas $X|G = 1 \sim N(70, 4)$.

   Note that $X|G = k \sim N(\mu_k, \sigma_k^2)$ signifies that

   $$f_k(x) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left\{-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right\}, k = 0, 1,$$

is the probability density function of X in group k.

How do you classify a student with an entry score $X = 78$?

Notice that in the script `dnorm()` computes the value of the density function, the second and third argument are the mean and the standard deviation. The function `pnorm()` computes $\Phi(x) = \int_{-\infty}^{x} f(u)\mathrm{d}u = P(X \leq x)$. Finally, the function `rnorm()` generates a random draw from the Gaussian distribution.

2. Compute the missclassification rate, the true positive rate and the false positive rate from the following confusion matrix:

	$\hat{G}(X)$	
G	0	1
0	250	13
1	15	120

3. (*Exam question*) According to our past experience the probability that a client is creditworthy is $\pi_1 = 0.8$ (the probability of a bad client is $\pi_0 = 0.2$). Our clients are segmented in two groups, according to the credit duration (X variable): *long* and the *short*. We further know that $\mathrm{P}(X = \mathit{long}|G = 0) = 0.7$ (i.e. 70% of the bad clients asked for *long* durations, and the remaining 30% for *short* durations), whereas $\mathrm{P}(X = \mathit{long}|G = 1) = 0.6$ (i.e. 60% of the good clients asked for *long* durations).

   A new client asks for credit with a *long* duration. Compute the posterior probability that he/she is a bad client?

4. (*Exam question*) Let $G = 1$ be the indicator of default ($G = 1$ is default, $G = 0$ is no default). From a training sample you estimate the prior probability $\pi_1 = P(G = 1)$ as $\hat{\pi}_1 = 0.2$. Let X denote the amount of the credit measured in thousands. You estimate that the probability of asking for a large credit amount, larger than 200 thousands, is larger for the group $G = 1$, and that in particular $\hat{P}(X > 200|G = 1) = 0.8$, whereas $\hat{P}(X > 200|G = 0) = 0.3$. Use Bayes Theorem to classify a customer who asks for a credit larger than 200.

5. (*Exam question*)

   In the logistic regression of a dichotomous random variable G taking two states, $\{0, 1\}$, on a set of input variables $X = (X_1, \dots, X_p)$, how is $P(G = 1|X = x)$ specified as a function of the values of the input x?

6. *This question is worth 12 points (7+(3+2))*
   Logistic regression is a fundamental parametric tool for the prediction of a nominal input variable G with two response categories.

   (a) Discuss in detail the specification of the model, stating what model is assumed for $P(G = 1|X = x)$.

(b) The following table presents the main estimation results for the logistic regression of the indicator of a bad credit on `Duration` and `CreditAmount`, their squares and interactions (1000 observations).

```
Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                -1.850e+00  2.827e-01  -6.544 5.99e-11
Duration                    7.075e-02  2.426e-02   ????? 0.003541
CreditAmount               -1.662e-04  9.195e-05  -1.807 0.070731
I(CreditAmount^2)           4.675e-08  1.021e-08   4.580 4.65e-06
I(Duration^2)               6.184e-04  5.602e-04   1.104 0.269640
I(Duration * CreditAmount) -1.345e-05  3.732e-06  -3.605 0.000312

    Null deviance: 1221.7  on 999  degrees of freedom
Residual deviance: 1143.8  on 994  degrees of freedom
AIC: 1155.8
```

- Compute the z value for the variable `Duration`.
- Suppose that for an individual the value of the estimated logit, $\hat{\boldsymbol{\beta}}'\mathbf{x}_i$, equals 0.2. What is the corresponding estimated probability $\hat{p}_i$ of being a bad credit?
- What is the interpretation of the Null and Residual deviance reported in the table?