

UNIVERSITÀ DI ROMA TOR VERGATA

EEBL - Statistical Learning

Revision - Week 5

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*. Springer, Springer Series in Statistics.

- Splines: chapter 7. Regression splines are dealt with in sections 7.1–7.4; smoothing splines in section 7.5.
- Local polynomial regression: section 7.6. Section 7.8 deals with R examples.

You may also want to refer to: 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/>.

- Splines: chapter 5, sections 5.1, 5.2 up to page 146. Section 5.4. up to page 154.
- Local polynomial regression: chapter 6, sections 6.1-6.3, 6.6.

2 Solved exercises

1. (Bayes theorem) From our past experience we have estimated that the probability that a person entering our store buys an item is $\pi_1 = 0.4$, whereas the probability that s/he leaves the store without buying anything is $\pi_0 = 0.6$. Our clients are segmented in two groups, according to their age (X variable): the *young* and the *old*. We further know that $P(X = \text{young} | G = 1) = 0.6$ (i.e. 60% of our clients are *young*, and the remaining 40% are *old*), whereas $P(X = \text{young} | G = 0) = 0.5$ (i.e. 50% of the non-shoppers are *young*).

A young person enters our store shop. Is s/he more likely to buy or leave the store without buying?

You should compute $\pi_0 P(X = \text{young} | G = 0) = 0.6 \times 0.5$ and $\pi_1 P(X = \text{young} | G = 1) = 0.4 \times 0.6$. This implies that $P(G = 0 | X = \text{young}) = \frac{0.6 \times 0.5}{0.6 \times 0.5 + 0.4 \times 0.6} > P(G = 1 | X = \text{young}) = \frac{0.4 \times 0.6}{0.6 \times 0.5 + 0.4 \times 0.6}$ and thus s/he is more likely to leave the store without buying.

2. 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:

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	2.916	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

- What variables are significant at the 5% level?

*All those whose associated p-values is less than 0.05, that is (Intercept), Duration, I(CreditAmount²), I(Duration * CreditAmount).*

- How do you interpret the Null and Residual deviance reported in the table?

As the sum of the squares of the deviance residuals. The Null Deviance is minus twice the log-likelihood of the logistic regression model containing only the intercept ($\beta_1 = \dots = \beta_p = 0$); the Residual deviance is minus twice the log-likelihood of the estimated model with 5 explanatory variables.

3. Suppose that for an individual the value of the estimated logit, $\hat{\beta}' \mathbf{x}_i$, equals 0.5. What is the corresponding estimated probability \hat{p}_i of being a bad credit?

$$\hat{p}_i = \frac{e^{0.5}}{1+e^{0.5}} > 1/2$$

4. What type of residuals are available for diagnostic checking and for goodness of fit assessment?

The Pearson's residuals and the Deviance residuals.

5. Compute the true positive rate and the false positive rate from the following confusion matrix:

G (actual value)	$\hat{G}(X)$ (prediction outcome)	
	0	1
0	300	20
1	10	200

True Positive Rate: 200/210; False positive rate: 20/320.

3 Exam questions

1. A cubic smoothing spline is the function $f(x)$ minimising the following penalised least square objective function:

$$\sum_{i=1}^N [y_i - f(x_i)]^2 + \lambda \int [f''(x)]^2 dx.$$

- (a) What solution for $f(x)$ is obtained if we let the smoothing parameter λ go to infinity and what is the corresponding number of degrees of freedom of the fit?
- (b) What solution is obtained for $\lambda = 0$, instead?

2. Consider the regression model with a single input X , $Y = f(X) + \epsilon$, where $f(X) = E(Y|X)$ is an unknown conditional mean function and ϵ is the error term, such that $E(\epsilon|X) = 0$ and $\text{Var}(\epsilon|X) = 0$.

Discuss the approximation of $f(X)$ by a spline function, touching upon the following issues:

- Provide the definition of a polynomial spline function and discuss its representation as a regression model.
- Discuss (in words) what is a natural spline and why it is entertained more often than a polynomial spline.
- What model selection problems are posed by splines?

- A cubic smoothing spline is the function $f(x)$ minimising the following penalised least square objective function:

$$\sum_{i=1}^N [y_i - f(x_i)]^2 + \lambda \int [f''(x)]^2 dx.$$

Discuss the nature of the objective function and the role of the parameter λ .

3. 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

- - What is the interpretation of the Null and Residual deviance reported in the table?
 - Is the variable CreditAmount significant at the 10% level?
4. Consider the regression model with a single input X , $Y = f(X) + \epsilon$, where $f(X) = E(Y|X)$ is an unknown conditional mean function and ϵ is the error term, such that $E(\epsilon|X) = 0$ and $\text{Var}(\epsilon|X) = 0$.
 - Discuss the approximation of $f(X)$ by a local polynomial, explaining in words what is the rational of local polynomial fitting.
 - Explain in words what is a kernel function.
 - Discuss the problem of selecting the kernel, the bandwidth and the order of the polynomial.
 5. Classification is the second supervised learning problem that we considered.
 State with your words Bayes' classification rule and provide a list of the models methods that have been considered during the course for this supervised learning problem.