UNIVERSITÀ DI ROMA TOR VERGATA

EEBL - Business Statistics

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*. Springer, Springer Series in Statistics, 2009. Dowloadable at http://www-bcf.usc.edu/ gareth/ISL/

- Review the slides on Model Evaluation and Selection.
- Linear model selection and regularization are dealt with in chapter 6. You can skip pages 226-227 on the Bayesian interpretation as well as the section on Partial Least Squares. Read also about Crossvalidation: section 5.1.
- Review the R script on Principal components regression.
- Review the slides on Classification, slides 1-19, 29-33.
- Classification is dealt with in chapter 4 of the textbook. We have covered sections 4.1-4.2, 4.3 (logistic regression), 4.4. (Discriminant analysis) up to page 146, then we did section 4.4.4.
- It is strictly not necessary, but if you want a more advance 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/.

  - The Lasso: section 3.4.2.-3.4.3
  - Principal Components Regression: section 3.5.1.
  - Loss functions for classifications: pages 20 - 22.
  - Classification: chapter 4, read the introduction, sec 4.1.
  - Linear regression for classification, sec. 4.2 (omit technical details).
  - 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 $X|G = 1 \sim N(70, 4)$ ($X|G = k \sim N(\mu_k, \sigma_k^2)$ signifies that

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

   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. (Exam question) Discuss the main characteristics of the shrinkage method known as Lasso (least absolute shrinkage and selection operator).

3. (Exam question) In the linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{y}$ is a vector of N standardised observations on the output, and $\mathbf{X}$ is an $N \times p$ matrix of measurements on p standardised inputs, the ridge estimator of the regression coefficients is

   $$\hat{\boldsymbol{\beta}}_r = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}.$$

   (a) Explain the rationale and the properties of this estimator, stressing the role of the shrinkage parameter λ.

   (b) How do we select the value of λ?

   This question refers to the principal components analysis of a data matrix consisting of the assessment of 7 attributes of an industrial supplier made by 50 buyers. The attributes are X1: Delivery speed; X2: Price level; X3: Price flexibility; X4: Manufacturer's image; X5: Service; X6: Salesforce's image; X7: Product quality.

   The parwise correlations are in the matrix

4.
```
> R = cor(X)
     X1   X2   X3   X4   X5   X6   X7
X1 1.00 0.93 0.88 0.57 0.71 0.67 0.93
X2 0.93 1.00 0.84 0.54 0.75 0.47 0.94
```

```
X3 0.88 0.84 1.00 0.70 0.64 0.64 0.85
X4 0.57 0.54 0.70 1.00 0.59 0.15 0.41
X5 0.71 0.75 0.64 0.59 1.00 0.39 0.57
X6 0.67 0.47 0.64 0.15 0.39 1.00 0.57
X7 0.93 0.94 0.85 0.41 0.57 0.57 1.00
```

The R function `eigen` returns the eigenvalues and the eigenvectors of the correlation matrix.

```
> eigen(R, symmetric=TRUE)
$values
[1] 5.035 0.934 0.498 0.421 0.081 0.020 0.011
$vectors
      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]
[1,] -0.43  0.11  0.08 -0.04  0.63  0.34  0.53
[2,] -0.42 -0.03  0.44  0.01 -0.00 -0.79  0.10
[3,] -0.42 -0.01 -0.20 -0.33 -0.70  0.16  0.40
[4,] -0.29 -0.67 -0.45 -0.30  0.26 -0.11 -0.30
[5,] -0.35 -0.29 -0.01  0.85 -0.17  0.20 -0.07
[6,] -0.29  0.64 -0.60  0.15  0.09 -0.24 -0.23
[7,] -0.41  0.20  0.43 -0.25 -0.05  0.37 -0.64
```

- How is the first principal component defined?
- What is the variance of the first principal component?
- Is the first principal component interpretable from the above results?
- Do you think that the data can be effectively summarised by the first principal component?
- What value do you get if you sum the squares of the loadings in the first eigenvector $\mathbf{a}_1$?

5. 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   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?
- How do you interpret the Null and Residual deviance reported in the table?

- Suppose that for an individual the value of the estimated logit, $\hat{\boldsymbol{\beta}}'\mathbf{x}_i$, equals 0.5. What is the corresponding estimated probability $\hat{p}_i$ of being a bad credit?
- What type of residuals are available for diagnostic checking and for goodness of fit assessment?

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

	$\hat{G}(X)$ (prediction outcome)	
G (actual value)	0	1
0	250	13
1	15	120

7. (*Exam question*)

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

8. (*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 $\mathrm{P}(X = \textit{long}|G = 0) = 0.7$ (i.e. 70% of the bad clients asked for *long* durations, and the remaining 30% for *short* durations), whereas $\mathrm{P}(X = \textit{long}|G = 1) = 0.6$ (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?