UNIVERSITÀ DI ROMA TOR VERGATA

EEBL - Statistical Learning

Revision - week 1

1. Read the following parts of the textbook "An Introduction to Statistical Learning", by James et al. (2021, 2nd ed.)

   - Chapter 1: Introduction. In particular, study the matrix introduction in pages 9-12 and review the slides on vector and matrices.
   - Chapter 2: Read pages 15-17, for an introduction to Regression. As for Section 2.1.1. try to read as much as you can. We will discuss this section next week.
   - Read section 2.1.4 for Supervised Versus Unsupervised Learning. Some of the material in this section will be clearer at a later stage.
   - Read section 2.1.5 for Qualitative vs Quantitative variables and the distinction between Regression and Classification.
   - Section 2.3 is a nice introduction to R.

2. Download and install R and R-Studio. Download and install R (base) from The Comprehensive R Archive Network https://cran.r-project.org/. Download and install RStudio: select the free version at https://posit.co/download/rstudio-desktop/ (you can install R directly from there). A useful website is Quick R: http://www.statmethods.net/

3. Review the notion of variance, covariance and correlation. It is crucial that these notions are well understood at this stage.

   If $\mathbf{y}$ and $\mathbf{x}$ are vectors with N observations on variables Y and X, then $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ and $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ are the sample means of Y, and X, respectively, $s_y^2 = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})^2$ and $s_x^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$ are the sample variances; $s_y = \sqrt{(s_y^2)}$ and $s_x = \sqrt{(s_x^2)}$ are the standard deviations.

   $$s_{xy} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x})$$

   is the sample covariance, while the correlation is

   $$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

   and it takes values between -1 and 1.

   These elements can be arranged in the covariance matrix:

   $$\mathbf{S} = \begin{bmatrix} s_x^2 & s_{xy} \\ s_{xy} & s_y^2 \end{bmatrix}.$$

   This is a symmetric matrix, as $s_{xy} = s_{yx}$.

   In R you obtain all the pairwise correlations using the function `cor()`.

4. The following points refer to the dataset br2.csv (See the folder Lab1), containing 1080 observations on 11 variables referring to price and house characteristics for houses sold in Baton Rouge, Luisiana, 2005. The variables are:

```
price          sale price, dollars
sqft           total square feet
Bedrooms       number of bedrooms
Baths          number of full baths
Age            age in years
Owner          = 1 if owner occupied at sale; = 0 if vacant or tenant
Pool           = 1 if pool present
Traditional    = 1 if traditional style; = 0 if other.
Fireplace      = 1 if fireplace present
Waterfront     = 1 if on waterfront
    DOM            Days on the market
```

   (a) Using R, display the scatterplot diagram of $\{(x_i, y_i), i = 1, \ldots, n\}$ for: (X = sqft, Y = price); (X = Age, Y = price).

   (b) Are house prices correlated with Age?

   (c) Compute the median, mean, variance and standard deviation of the quantitative variables in the dataset.

   (d) How many houses have the Pool?

   (e) Find out how to obtain the covariance matrix and the correlation matrix of the quantitative variables in the dataset.

5. Using the numbers in the table below, compute the correlation coefficient between X and Y.

	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
	93.3	96.6	-9.9	-4.3	98.8	18.4	42.6
	106.3	100.8	3.1	-0.1	9.5	0.0	-0.4
	109.1	108.3	5.9	7.3	34.4	53.7	43.0
	96.2	92.9	-7.0	-8.0	49.7	64.6	56.6
	111.3	106.1	8.0	5.1	64.8	26.3	41.3
Mean	103.3	100.9	0.0	0.0	51.4	32.6	36.6

6. Consider the matrix and vectors

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix}, \quad \mathbf{i} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

   Compute $\mathbf{X}'\mathbf{X}$, $\mathbf{X}\mathbf{X}'$, $\mathbf{i}'\mathbf{y}$, $\mathbf{X}'\mathbf{i}$ and $\mathbf{X}'\mathbf{y}$. Are the vectors $\mathbf{i}$ and $\mathbf{y}$ orthogonal?

7. Consider the Table in Exercise 7, page 54 of the textbook. Compute the pairwise Euclidean distances between Observations 3, 4 and 5.