Domanda **1**

Risposta non ancora data

Punteggio max.: 1,00

Suppose that we wish to predict the area of survey (“Industrial site” or ”Seaside”) from the measurement of the concentration level X of a polluting agent.

We examine a large number of daily measurements and discover that the mean value of X for the industrial site is $\bar{x} = 17$, with a variance of 144, while the mean for the seaside area was $\bar{x} = 11$, with a variance of 196. Finally, the 70% of sampled measurements refer to the industrial area.

Assuming that X follows a normal distribution, predict the probability that a concentration level of 15 was taken at the industrial site.

Answer by coping and paste the code you will type in R.

Domanda **2**

Risposta salvata

Punteggio max.: 1,00

A group of pregnant women joined a clinical study. For each patient the gestation week [X_1], the age, assumption of misoprostol (a medication used to prevent and treat stomach ulcers but that can also cause miscarriage) [X_2, 0="no assumption", 1="assumption"], and glucose level [X_3] were measured.

A logistic regression model was fitted to quantify the effect of such features on the pregnancy end (0="regular pregnancy", 1="miscarriage"). The estimated coefficients are the following:

$$\hat{\beta}_0 = -2.4 \quad \hat{\beta}_1 = -0.9 \quad \hat{\beta}_2 = 1.4 \quad \hat{\beta}_3 = 0.06$$

Estimate the probability that a woman at the 9th week of pregnancy, that has not assumed misoprostol and with a glucose of 140 will have a miscarriage.

Answer by coping and paste the code you will type in R and the corresponding result.

Domanda **3**

Risposta salvata

Punteggio max.: 1,00

On average, what fraction of women with an odds of 0.26 of having twins will in fact have twins?

Scegli un'alternativa:

- There is not enough information
- 0.260
- 0.206
- 0.351
- 0.740

Domanda **4**

Risposta salvata

Punteggio max.: 1,00

Why additional measures (other than the misclassification error rate) may be needed to evaluate classification accuracy? Describe them briefly.

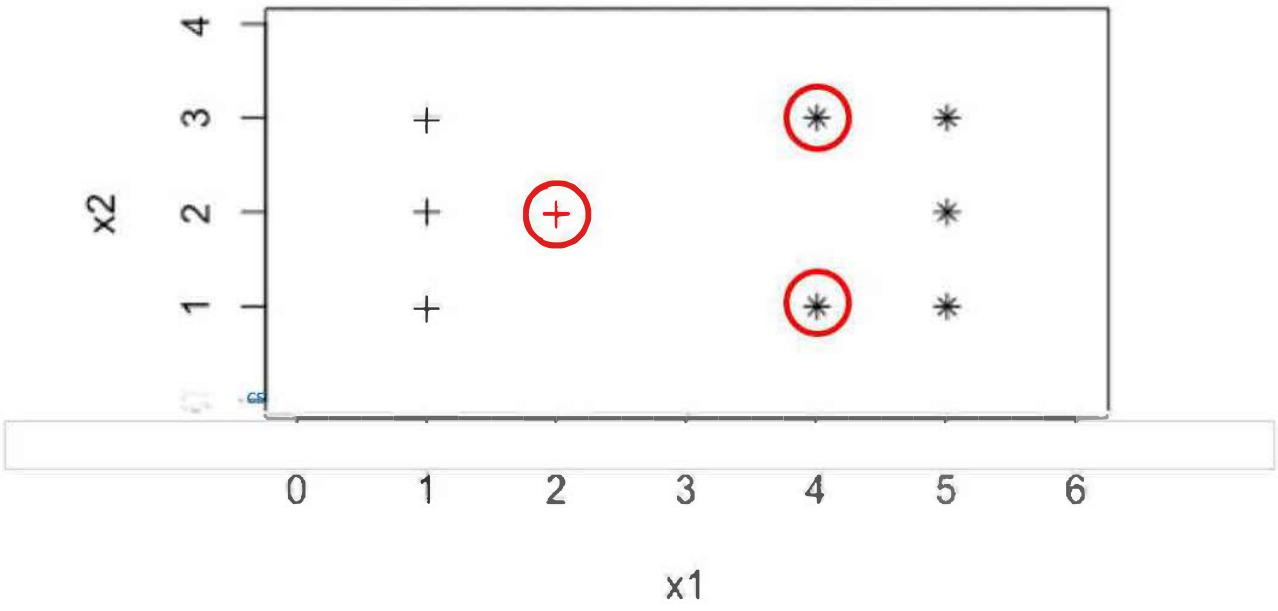
Please answer below or attach the file with the response.

Domanda **5**

Risposta salvata

Punteggio max.: 1,00

Suppose you are using a Linear SVM classifier with 2 class classification problem. Now you have been given the following data in which some points are circled red that are representing support vectors:



If you remove the following any one red points from the data, does the decision boundary will change?

Scegli un'alternativa:

- Yes
- No
- There is not enough information

Domanda **6**

Risposta non ancora data

Punteggio max.: 1,00

Load the workspace **`May31st sitting.RData`** into your R environment.

Using command **`ls()`** you will see objects: **`abal21.tr`** and **`abal21.te`**.

The aim is to predict the age of abalone from physical measurements. Specifically, the age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope -- a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age.

The dataset has 7 features and contains 410 observations in the training set (**`abal21.tr`**) and 410 in the test set (**`abal21.te`**). The response variable is **`Rings`**.

1. Perform variable selection via best subsets; use the BIC as criterion to choose the best model size. Which predictors are included in the model? Estimate the test error.
2. Could R^2 have been employed as criterion to choose the best model size? Justify your answer.
3. Use Lasso to perform variable selection. Tune the parameter lambda via cross-validation. Which predictors are retained? Estimate the test error.
4. Describe the role of the lambda parameter of the Lasso. Why would you need cross-validation to tune it?
5. Grow a regression tree to predict the response variable. Which feature appears to be more important? Estimate the test error.
6. Which of the previous approaches would you choose to predict **`Rings`**? Explain why.
7. Create a new variable `class` by coding `Rings` as "0" if `Rings<9` and "1" otherwise (e.g. `abal21.tr$class=ifelse(abal21.tr$Rings<9,0,1)` and similarly for the test set). Perform bagging on the training set (**exclude** `Rings` from the set of available features). Estimate the classification error of the test set. Which variable appears to be the most important feature?

When solving the exercise you may want to set a seed before any function that requires non-deterministic steps.

Please submit your output as an HTML file, by compiling a Rmd file wit your name on it.