STAT 8178/7178

Instructions:

This assignment covers from weeks 1 to 8

1. Due on 5th May 2023
2. For all the questions please provide the relevant mathematical derivations, the computer programs (only using R software) and the plots.
3. Please submit on iLearn a single PDF file containing all your work (code, computations, plots, etc.). Other file formats (e.g. Word, html) will NOT be accepted.
4. Try to use Rmarkdown through Rstudio. But it is not compulsory to use Rmarkdown even if facilitate to reproduce results. Only upload the pdf file.

Question 1: 10 marks

Consider m samples $\{(x_1, y_2), \ldots, (x_m, y_m)\}$ with $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$.

(a) [2 marks] Show that the following function $g(w)$ is convex:

$$g(w) = \frac{1}{2}\sum_{i=1}^{m}(y_i - w^T x_i)^2,$$

where $w \in \mathbb{R}^d$.

(b) [2 marks] Show that the following function $h(w)$ is convex:

$$h(w) = \frac{1}{2}\gamma||w||_2^2,$$

where $\gamma > 0$

(c) [2 marks] Using results from (a) and (b), show that $f(w) = g(w) + h(w)$ is convex.

(d) [4 marks] Solve the optimization problem

$$\min_{w \in \mathbb{R}^d} f(w)$$

by expressing the minimizer w in terms of the data matrix of x and the vector y.

Question 2: 21 marks

In this question, we consider breast cancer prediction where the label, or outcome variable `diagnosis` has been coded as "M" in case of malignant lumps and "B" in case of benign lumps. A popular dataset in this context is called the Wisconsin Breast Cancer Dataset and is based on clinical data released in the early 1990's. The feature vector x is composed of continuous variables such as `radius_mean`, `texture_mean`, etc., each potentially affecting the probability of malignancy. We use the 80-20 splitting strategy where we split it randomly between training data and testing data. We want to build a prediction model to predict malignant lumps based on main important features.

1. 1 mark Write down the logistic model for this task.
2. 1 mark Load the following file `Breast.Rdata` using

   ```
   load("Breast.Rdata")
   ```

   The above code provides two data frames: `train` and `test`. For both data sets, the first column (named "diagnosis") is the categorical outcome. How many attributes are available to predict the outcome? How many samples are included in the two data frames. What is the distribution of the outcome variable in the two data sets ?
3. 1 mark Run a logistic model (previously defined) using all attributes (Hint: use the glm function and specify the family argument). You will estimate your model using the train dataset.
4. 2 marks Provide the confusion matrix for the train set of your classifier using a threshold of 0.5 and provide the accuracy of the model for the train set.

5. 2 marks Provide the confusion matrix for the test set of your classifier using a threshold of 0.5 and provide the accuracy of the model for the test set.

6. 2 marks Why the accuracy for the test set is lower than the one for the training set ? Which accuracy to report for assessing the performance of your classifier?

7. 2 marks We want to get a parsimonious model, meaning that we want to keep the most relevant features. One way to tackle this challenge is to run a penalized regression model. One scientist is struggling to choose between a ridge and a lasso regression model. Give some justification to choose between the two strategies.

8. 2 marks We want to run a penalized logistic regression using a lasso penalty. To do it you will use the glmnet R package using `cv.glmnet` and `glmnet` functions (Hint: do not forget to use "family=binomial"). Choose the best tuning parameter lambda using a K-fold cross-validation strategy (Hint: use the argument type.measure="class" for choosing lambda to get the smallest miss-classification error). Plot the cross-validation error according to the log of lambda.

9. 2 marks Run the penalized logistic regression for the lambda you have chosen at the previous step. How many attributes are still in the model?

10. 2 marks For this model and a threshold at 0.5, define your classifier (Hint: use the function with argument type="response"). Report the confusion matrix on the test set and the accuracy of the model.

11. 2 marks Present on the same plot the ROC curves for the logistic model and the penalized logistic model.

12. 2 marks Report the two AUC (Area Under the ROC curve). What is your preference between the two models ?

Question 3: 7 marks

Let $X_1, ..., X_n$ be a random sample from a population with the following Bernoulli distribution

$$P(X = x) = \theta^x (1-\theta)^{1-x}, \quad x = 0 \text{ or } 1, \ 0 \leq \theta \leq 0.5.$$

We know that the maximum likelihood estimator (MLE) for θ is,

$$\hat{\theta} = \min\{\bar{X}, 0.5\},$$

and the method of moments estimator (MOM) for θ is given by,

$$\tilde{\theta} = \bar{X}.$$

1. 1 mark Generate a random sample with 500 observations from the above distribution function when $\theta = 0.45$.

2. 2 marks Compute bootstrap estimates of θ using the MLE and MOM estimates. Let's assume 1000 replications.

3. 2 marks Compare the performance of the $\hat{\theta}$ and $\tilde{\theta}$ using your bootstrap samples by computing the bias and standard error.

4. 2 marks Find Bootstrap percentile intervals of the $\hat{\theta}$ and $\tilde{\theta}$ using your bootstrap samples. Compare the results. Use the significance level $\alpha = 0.05$.

Question 4: 7 marks

The `remiss.csv` data set contains the remission times for 42 leukemia patients in weeks. Some of the patients were treated with the drug called 6-mercaptopurine ($group = 0$), and the rest were part of the control group ($group = 1$).

1. 1 mark Create a box plot for the remission times for two treatment and control groups. Compare the remission times of the two groups.

2. 1 mark Use a normal probability plot to check whether the distribution of the remission times for each group is Normal.

3. 1 mark Write the null hypothesis and alternative hypothesis to test for the equality of means between the two groups.

4. 4 mark Perform the above hypothesis test using Monte Carlo simulation to get the critical values. Estimate the p-value. Do you reject H_0 or H_1? You can assume the distribution of the remission times for each group is Normal and their population variances are equal. Use the significance level $\alpha = 0.05$.