University of Bristol
ECONM0011 (Machine Learning for Economics)
Professor Vincent Han

Coursework

Deadline: 28 April 2025

- This coursework is a team project. The output will be evaluated as a team, that is, all team members will receive the same score. Therefore, cooperation among team members is extremely important. If there is a free-rider problem, try to figure out how to resolve it. If not successful, please report the problem to the instructor (Vincent Han). Please show work ethic to your team members.
- Please submit one pdf file per group with the names (and student ID numbers) of all the members. Please indicate the group number in the file name. Then, clearly state in the file what each member's completed tasks and contributions are. Again, each member should have significant contributions to the project.
- All the questions should be answered using Python. Report all the codes used in answering the questions. The code can be attached as an appendix.
- All (sub)questions are equally weighted. For Part III, groups with the top three accuracy rates will receive bonus points.

Part I: Classification Analysis

Consider the data set *Heart.csv*, which consists of a sample of patients. It contains information about whether he or she has heart disease (a variable called *AHD*). It also contains information about the possible predictors of heart disease, such as age, sex, cholesterol level, and other heart and lung function measurements. We want to explore this sample and construct a classifier that classifies patients into heart disease and no heart disease.

1. Describe the data set by reporting summary statistics.
2. Then, conduct necessary preprocessing: If there are any missing observations, remove them. For classification (and also many other machine learning methods), it is helpful to first normalize the explanatory variables. Except for the categorical variables, standardize all the other variables (i.e., $RestBP, Chol, MaxHR$) so that their means are zero and standard deviations are one. For all the questions in 1 and 2 below, we will use this preprocessed data set.

   * You want to quantify all the categorical variables before proceeding. Think about the most appropriate way of doing it. For Sex, the dataset doesn't contain information about whether 1 is female or male. For this missing information, assume that a male is more likely to have heart diseases than a female (which is a well-known fact), and conclude how to interpret the value of Sex.

(a) After all the preprocessing, report summary statistics.

3. Given this dataset, now the goal is to construct a classifier that achieves the largest accuracy possible.

   (a) Find three different classifiers (i.e., classification methods), one of which is your best classifier. For each classifier, choose necessary tuning parameters to be optimal. In doing so, describe in what sense such a choice of tuning parameter values is optimal.

   (b) Consider the three classification methods chosen in (b) (with optimal tuning parameter values). For each method, construct a confusion matrix (see Table 4.4 in ISLR). Compare the matrices across the three methods.

   (c) Compare the ROC curves across the three methods.

   (d) Finally, compare the accuracy of the three classifiers (with optimal tuning parameter values) and show that your choice yields the highest accuracy.

4. Using the best classifier you find in (d), classify whether the following patient has a heart disease or not: This patient is a 55-year-old woman who has a typical *ChestPain, Thal* is normal, and $RestBP = 130, Chol = 246, Fbs = 0, RestECG = 2, MaxHR = 150, ExAng = 1, Oldpeak = 1, Slope = 2, Ca = 0$.

Part II: Regression Analysis

Consider the data set *Credit.csv* which contains information about individual's credit scores and other characteristics. Using this dataset, we want to understand which characteristics are important in predicting average credit card debt (*balance*).

1. First, we want to consider the following (nonlinear) regression model:

$$y_i = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} + \sum_{j=1}^{J} \sum_{k=1}^{J} \gamma_{jk} x_{ij} x_{ik} + e_i$$

   where y_i is the balance and $\{x_{ij}\}_{j=1}^{J}$ are the continuous characteristics (*standardized*) and characteristic dummies. We want to make sure creating relevant dummy variables and dropping missing observation in *Credit.csv* before proceeding.

   (a) Report summary statistics.

   (b) Conduct the estimation of the model using lasso. In doing so, we want to choose the optimal λ using a 5-fold CV.

      i. Draw a figure that shows how the optimal λ value is chosen.

      ii. Report the coefficient plot as in the left panel of Figure 6.6 in the text book.

   (c) Using the estimated model in (b), predict the balance of a married 70-year-old Asian female whose income is 100, limit is 6000, rating is 500, has 3 cards, has 12 years of education, and is not a student.

2. Next, consider using tree-based methods for the same data set.

   (a) We want to fit a random tree with maximum depth of 3. Visualize this tree and interpret the result.

   (b) Now we want to use a random forest. Calculate the test MSEs with the maximum depth of 3 and the number of trees in $\{1, 5, 10, 50, 100, 200\}$. Then, calculate the test MSEs without specifying the maximum depth and with the number of trees in $\{1, 5, 10, 50, 100, 200\}$. Plot *all* the test MSEs (including the test MSE for the random tree in (a)) in one graph. Discuss the results.

   (c) Using the random forest that yields the lowest test MSE in (b), predict the balance of the person in part 1(c).

   (d) For this random forest, plot a graph that shows the importance of variables (i.e., similar to Figure 8.9 in the text book) and discuss the findings. Also, compare the findings with what you found in part 1(b)-ii.

Part III: Image Classification

Consider the dataset *Font_Images.zip*, which contains images of fonts. Using this dataset, we aim to construct a classifier that can classify font images in three steps:

1. Image preprocessing and construction of data:

   (a) First, take each image and randomly crop it to create square images using the provided *Crop.py* script.

   (b) Label these crops to indicate they belong to the same font. Essentially, this will create $\{X_i, Y_i\}_{i=1}^n$ where X_i represents the pixel values of crop i and Y_i is the font identity of crop i.

2. Training a classifier:

   (a) Using the training set portion of the data you constructed in Step 1, train a model of your choice.

      i. Report all details of the model and training (e.g., type of model, model architecture, tuning parameters, training duration, etc.).

      ii. When choosing tuning parameter values, consider using the validation set portion of the data.

3. Evaluate your classifier:

   (a) Using the test set portion of the data, evaluate your model by reporting the classification accuracy and the ROC curve.