CITS5508 Machine Learning
Semester 1, 2024

Assignment 3

Assessed, worth 20%. Due: 8pm, Friday 24th May 2024

Discussion is encouraged, but all work must be done and submitted individually. This assignment has 10 assessed tasks, which total 70 marks.

In this assignment, you will explore the California Housing dataset using various techniques, including Principal Component Analysis (PCA), clustering, and supervised learning algorithms. The goal is to analyze and model housing prices in California based on different features.

1 Submission

Your submission consists of **two files**. The first file is a report describing your analysis/results. Your analysis should provide the requested plots, tables, and reflections on the results. Each deliverable task is indicated as **D** and a number. Your report should be submitted as a ".PDF" file. Name your file as `assig3_<student_id>.pdf` (where you should replace `<student_id>` with your student ID).

The second file is your Python notebook with the code supporting your analysis/results. Your code should be submitted as `assig3_<student_id>.ipynb`, the Jupyter Notebook extension.

Submit your files to LMS before the due date and time. You can submit them multiple times. Only the latest version will be marked. Your submission will follow the rules provided in LMS.

Important:

- You must submit the first part of your assignment as an electronic file in PDF format (do not send DOCX, ZIP or any other file format). Only PDF format is accepted, and any other file format will receive a zero mark.
- You should provide comments on your code.
- You must deliver parts one and two to have your assignment assessed. That is, your submission should contain your analysis and your Jupyter notebook with all coding, both with appropriate formatting.
- By submitting your assignment, you acknowledge you have read all instructions provided in this document and LMS.
- There is a general FAQ section and a section in your LMS, Assignments - Assignment 3 - Updates, where you will find updates or clarifications about the tasks when necessary. It is your responsibility to check this page regularly.

- You will be assessed on your thinking and process, not only on your results. A perfect performance without demonstrating understanding what you have done won't provide you marks.

- Your answer must be concise. A few sentences should be enough to answer most of the open questions. You will be graded on thoughtfulness. If you are writing long answers, rethink what you are doing. Probably, it is the wrong path.

- You can ask in the lab or during consultation if you need clarification about the assignment questions.

- You should be aware that some algorithms can take a while to run. A good approach to improving the Python speed is using the vectorised forms discussed in class. In this case, it is strongly recommended that you start your assignment soon to accommodate the computational time.

- For the functions and tasks that require a random procedure (e.g. splitting the data into 80% training and 20% validation set), you should set the seed of the random generator to the value "5508" or the one(s) specified in the question.

2 Dataset:

We will use a modified version of the California housing dataset, which contains data from the 1990 California census and includes information about housing prices and various factors affecting housing values across different districts in California. The dataset comprises features such as median income, housing median age, etc., and the target variable is the median house value for California districts. The dataset unit (that is, each row in the dataset) corresponds to a district. A district represents the smallest geographical unit for which the U.S. Census Bureau publishes sample data. Typically, a district contains a population ranging from 600 to 3,000 individuals.

The variables of the dataset are:

- longitude: district group longitude.

- latitude: district group latitude.

- housingMedianAge: median house age in the district.

- totalRooms: the total number of rooms in the district.

- totalBedrooms: the total number of bedrooms in the district

- population: district population.

- households: the total number of households in the district.

- medianIncome: median income in the district.

- oceanProximity: whether each district is near the ocean, near the Bay area, inland or on an island (categorical).

- medianHouseValue: the median house value (target variable).

In all asked implementations, you should set `random_state=5508` when necessary for results reproducibility.

3 Tasks

Reading the dataset:

Read the dataset using the provided file in LMS: housingCalifornia.csv.

The dataset has 20,640 rows and 10 columns (eight numeric variable, one categorical variable and one target variable).

D1

Exploratory data analysis and preprocessing.

(a) Plot the histograms of the non-categorical features and the target in a grid subplot using the `histplot` function from the `seaborn` with the default values. **2 marks**

(b) Compute the correlation matrix of all features (including the target features). Do not use the categorical variable (ocean proximity). Describe which features are more correlated (correlation coefficient higher than 0.8 in magnitude) and why you think this is the case. **2 marks**

(c) Present a scatter plot for each variable, displaying the corresponding variable on the x-axis and the target variable on the y-axis. **2 marks**

Analysing the impact of different data transformations. Create two versions of the dataset as described below. Remember to properly encode the categorical variable by creating $l - 1$ new binary/dummy variables, where l is the number of categories of the variable. Use "`<1H OCEAN`" as a reference (you should not add a dummy variable for this category).

- The original dataset (that is, with the new dummy variables, removing the categorical variable, and without any other changing the features). We will refer to this version as data1.
- The dataset with the transformation in the target variable: change the unit of the target variable to hundreds of thousands of dollars. For example, if the median house value of a district was 452600.0 before, in this dataset, it will be 4.526. Thus, you will keep all the original predictor variables (same features as data1) and update the target variable. This version will be referred to as data2.

D2

Split each of these two datasets into training and test sets, using 80% of the data for training (use the Python `train_test_split` function). Remember to set the random generator's state to the value "5508" for the splitting function. Fit two models in each dataset: a linear regression model and a Lasso regression model with $\alpha = 100$. Standardised both datasets appropriately (transforming the features to have zero mean and unit standard deviation). Thus, you provide results for the original and standardised version of data1) and for the original and standardised version of data2).

(a) In a table, report the RMSE for the training and test sets for the two models for each dataset. That is, your table should contain four rows with four values each. **8 marks**

(b) Discuss the RMSE values obtained results. Specifically, discuss if they have the same values and why and if they have different values and why. **3 marks**

D3

Create three new features (meanRooms, meanBedrooms, and meanOccupation) as follows:

`meanRooms = total rooms / households`. It represents the mean number of rooms per household.

`meanBedrooms = total bedrooms / households`. It represents the mean number of bedrooms per household.

`meanOcupation = population / households`. It represents the mean number of household members.

Create a new dataset (data3) by deleting the features total rooms, total bedrooms, households and population `and` by adding these new three features. The target variable should be as in data2, that is, expressed in hundreds of thousands of dollars. Split data3 into training and test sets, using 80% of the data for training and setting the random generator's state to the value "5508" for the splitting function. Fit a linear regression model and a Lasso regression model with $\alpha = 100$ to the data with the proper feature standardisation and without standardisation.

(a) Report the RMSE for the two models' training and test sets in a table. Your table should contain two rows with four values each. **4 marks**

(b) Discuss and justify the obtained values of RMSE. **2 marks**

(c) Report the estimated parameter values with the corresponding variable names for all models (12 in total, eight from D2 and four from D3). **4 marks**

(d) Discuss the obtained results. Are there similarities with the parameters' values from each model? Justify your answer. **2 marks**

Analysing the impact of different models.

D4

Consider data3, using the 80%-20% splitting of the data and the appropriate standardisation. Train a Lasso Regression using a 10-fold cross-validation and `Grid-SearchCV` to fine-tune the regularisation parameter α. Pay attention if you need to set the random state of your cross-validation procedure. In your grid, consider ten different values for α:
$\alpha : [0.0000001, 0.000001, 0.00001, 0.001, 0.001, 0.01, 0.1, 1, 10, 100]$.

(a) Report: **4 marks**

- The optimal α value according to the Grid-Search.
- The RMSE on the training set.
- The RMSE on the test set.
- The estimated parameter values with the corresponding variable names.

D5

Similar to **D4**, consider data3, using the 80%-20% splitting of the data and the appropriate standardisation. Train a Ridge Regression using a 10-fold cross-validation and `Grid-SearchCV` to fine-tune the regularisation parameter α. In your grid, consider ten different values for α:
$\alpha : [0.0000001, 0.000001, 0.00001, 0.001, 0.001, 0.01, 0.1, 1, 10, 100]$.

(a) Report: **2 marks**

- The optimal α value according to the Grid-Search.
- The RMSE on the training set.
- The RMSE on the test set.
- The estimated parameter values with the corresponding variable names.

(b) Compare the estimated values of the parameters and the α value from Ridge Regression and Lasso Regression (**D4**). **2 marks**

D6

Repeat the same process as in **D4** and **D5**, but now use a Decision Tree Regression. Remember to set the random generator's state of the class to the value "5508" for the splitting function. Consider data3, using the 80%-20% splitting of the data and the appropriate standardisation. Train a Decision Tree using a 10-fold cross-validation and `Grid-SearchCV` to fine-tune the regularisation parameters `max_depth`. In your grid, consider `max_depth:range(3,15,1)` and do not forget to set the random state to "5508".

(a) Report: **2 marks**

- The optimal `max_depth` value according to the Grid-Search.
- The RMSE on the training set.
- The RMSE on the test set.

D7

Consider the models you developed in **D4**, **D5**, and **D6**.

(a) Discuss their respective RMSE on the test set; which is the best model? Why? **1 mark**

(b) Considering the EDA analysis performed on **D1**, briefly discuss how the predictive capacity of the modes could be improved. **1 mark**

D8

Consider data3, and use only the numerical features. Using the 80%-20% splitting of the data, apply PCA in the training set using the standardised features. Pay attention if you need to set the random state of your PCA procedure.

(a) Plot the cumulative explained variance ratio as a function of the number of principal components. **2 marks**

(b) Determine the number of principal components necessary to preserve at least 90% of the variance. **1 mark**

(c) Train a `Linear Regression` using the selected number of principal components. Present the RSME for the training and test data. **2 marks**

(d) Use `GridSearchCV` to find the optimal number of principal components according to a 10-fold cross-validation and use a `Linear Regression` as the base model. Pay attention if you need to set the random state of your cross-validation procedure. Report the obtained optimal number of principal components and the RMSE for the training and test sets. **3 marks**

(e) Discuss the obtained results and compare them with the ones you obtained in **D7**. **2 marks**

Clustering analysis.

D9

Consider data3 without any categorical variable. Create the following clustering:

(a) Using this data, perform a hierarchical clustering with average linkage and Euclidean distance to cluster the districts. Cut the dendrogram at a height that results in four distinct clusters. Present the mean of the variables for each cluster and briefly summarise the characteristics of the districts in the four groups, including the size of each cluster. **3 marks**

(b) Using standardised features, perform a hierarchical clustering with average linkage and Euclidean distance to cluster the districts. Cut the dendrogram at a height that results in four distinct clusters. Did the groups change? What effect does scaling the variables have on the hierarchical clustering obtained? **2 marks**

(c) Using standardised features, apply the k-means clustering (with k=4) with Euclidean distance. Set the initial centroids of the k-means as the group means obtained from the hierarchical clustering in part (b). Compare the results with the hierarchical clustering from part (b). Which one do you think provides a better result? **2 marks**

(d) Perform PCA on the scaled data. Perform hierarchical clustering with average linkage and Euclidean distance on the first two principal component scores. Cut the dendrogram at a height that results in four distinct clusters. Present the scatterplot of the first two principal components using different colours for the instances on each cluster (four colours for four clusters). Compare the group characteristics to the group characteristics obtained in the previous tasks. **3 marks**

(e) Perform PCA on the scaled data. Apply the k-means clustering (with k=4) with Euclidean distance on the first two principal components scores, setting the random state to “5508”. Present the scatterplot of the first two principal components using different colours for the instances on each cluster (four colours for four clusters). Discuss the results. **2 marks**

D10

Consider data3 without any categorical variable. Compute the silhouette score by applying k-means on this dataset after scaling the features to have zero mean and unit standard deviation. Use values for k in `range(2,20,1)`. Remember to set `random_state=5508` for the `KMeans` class.

(a) Plot the silhouette scores for the different k values. According to this score, what was the optimal value of clustering? **2 marks**

(b) Considering the optimal k value obtained in the previous item, plot the k groups (using different colours for the instances in each group) on the first two principal component scores of the same data. In a side plot, plot the first principal component scores in which the instance colours represent the values of the categorical value you discarded for this part of the assignment. Comment on the relationship between the groups in these two plots. **3 marks**

(c) With clustering analysis and your findings from EDA, what are your conclusions about the data that may be impacting your models? **2 marks**