3
(/progress/?course=unitar%2Fcourses%2F2023-05-itim5123-machine-learning&student=yapkokhoong)

36%

Assignment 1 (Part 1)

Assignment 1 (Part 1)

Lab 1: Introduction to Machine Learning

This part introduces some basic concepts of machine learning with Python. In this assignment, you will use the K-Nearest Neighbor (KNN) algorithm to classify the species of iris flowers, given measurements of flower characteristics.

By the completion of this part, you will:

1. Follow and understand a complete end-to-end machine learning process including data exploration, data preparation, modeling, and model evaluation.
2. Develop a basic understanding of the principles of machine learning and associated terminology.
3. Understand the basic process for evaluating machine learning models.

Lab Steps

1. Make sure that you have completed the setup requirements as described in the Lab Overview section.
2. Now, run jupyter notebook and open the “IntroductionToMachineLearning.ipynb” notebook.
3. Examine the notebook and answer the questions along the way.

This assignment consists of 35 Multiple Choice-Questions (MCQs) and Multi-Selection Questions (MSQ). It carries **20%** of your carry marks. This activity should be completed in **Lesson 3**. Do note that you will only have **one attempt**. Once you have answered and clicked the "**Submit**" button (at the end of the page), your answers will be submitted and you **cannot** re-do this assignment.

1. From the plot, which species are more separated than the others?

○ A. Setosa

○ B. Versicolor

○ C. Virginica

2. What is the accuracy printed?

○ A. 95.0
[✓ ○ B. 96.0]
○ C. 97.0

3. How many cases are mis-classified?

[✓ ○ A. 3]
○ B. 4
○ C. 5

Lab 2: Visualizing Data for Regression

There are **two** goals for data exploration and visualization. **First** to understand the relationships between the data columns. **Second** to identify features that may be useful for predicting labels in machine learning projects. Additionally, redundant, colinear features can be identified. Thus, visualization for data exploration is an essential data science skill. This process is also known as **exploratory data analysis**.

In this lab, your **first** goal is to **explore** a dataset that includes information about automobile pricing. In other labs, you will use what you learn through visualization to create a solution that predicts the price of an automobile based on its characteristics. This type of predictive modeling, in which you attempt to predict a real numeric value, is known as **regression**; and it will be discussed in more detail later in the course. For now, the focus of this lab is on visually exploring the data to determine which features may be useful in predicting automobile prices.

By the completion of this lab, you will:

1. Use summary statistics to understand the basics of a data set.
2. Use several types of plots to display distributions.
3. Create scatter plots with different transparency.
4. Use density plots and hex bin plots to overcome overplotting.
5. Apply aesthetics to project additional dimensions of categorical and numeric variables onto a 2d plot surface.
6. Create pair-wise scatter plots and conditioned plots to create displays with multiple axes.

4. What can you conclude from the plotted histograms?

○ A. There are more cars with high miles per gallon than low miles per gallon
○ B. Most of the cars have larger than 3000 curb weight
○ C. There are only a few cars with engine size lower than 150
[✓ ○ D. Most of the cars cost less than 30,000]

5. From the kde plots, which feature shows the closest resemblance, in terms of distribution, to price?

○ A. curb_weight
[✓ ○ B. engine_size]
○ C. city_mpg

?

6. Select three relationships that are now apparent in the scatter plots:

□ A. Both gas and diesel turbo cars are generally more expensive than standard cars.

□ B. Turbo cars appear to have worse city_mpg at a given price point than standard cars.

□ C. Standard cars generally has diesel engine.

□ D. Turbo cars have greater horsepower at a given price point.

7. Select three relationships that are now apparent in the conditioned plots:

□ A. The distribution of the values generally increases for length and curb_weight for rear wheel drive (rwd) cars, with the values for 4 wheel drive (4wd) and rear wheel drive (rwd) overlapping.

□ B. Generally, 4wd cars have the highest engine size.

□ C. Cars with fwd have the highest city_mpg, whereas, 4wd and rwd in a similar range.

□ D. Generally, 4wd cars have the lowest price, with rwd cars having the widest range.

8. Which combination produces blank plots?

○ A. fwd and convertible

○ B. 4wd and hatchback

○ C. fwd and hardtop

○ D. 4wd and convertible

Lab 3 Visualizing Data for Classification

In this part, your goal is to explore a dataset that includes information about German bank credit to understand the relationships for a **classification** problem. In classification problems, the label is a categorical variable.

Visualization for classification problems shares much in common with visualization for regression problems. Colinear features should be identified so they can be eliminated or otherwise dealt with. However, for classification problems, you are looking for features that help **separate the label categories**. Separation is achieved when there are distinctive feature values for each label category. Good separation results in a low classification error rate.

By the completion of this lab, you will:

1. Examine the imbalance in the label cases using a frequency table.
2. Find numeric or categorical features that separate the cases using visualization..

Lab Steps

1. Make sure that you have completed the setup requirements as described in the Lab Overview section.
2. Now, run jupyter notebook and open the “VisualizingDataForClassification.ipynb” notebook under Module 2 folder.
3. Examine the notebook and answer the questions along the way.

9. From the created plots, which two features seem to separate the good and bad credits?

□ A. payment_pcnt_income

□ B. number_loans

□ C. age

□ D. loan_amount

10. From the created plots, which feature seems to separate the good and bad credits?

2

○ A. foreign_worker

○ B. telephone

○ C. job_category

[✓ ○ D. checking_account_status]

Data Preparation

Data preparation is a vital step in the machine learning pipeline. Just as visualization is necessary to understand the relationships in data, proper preparation or **data munging** is required to ensure machine learning models work optimally.

The process of data preparation is highly interactive and iterative. A typical process includes at least the following steps:

1. **Visualization** of the dataset to understand the relationships and identify possible problems with the data.
2. **Data cleaning and transformation** to address the problems identified. It many cases, step 1 is then repeated to verify that the cleaning and transformation had the desired effect.
3. **Construction and evaluation of a machine learning models**. Visualization of the results will often lead to understanding of further data preparation that is required; going back to step 1.

By the completion of this lab, you will:

1. Recode character strings to eliminate characters that will not be processed correctly.
2. Find and treat missing values.
3. Set correct data type of each column.
4. Transform categorical features to create categories with more cases and likely to be useful in predicting the label.
5. Apply transformations to numeric features and the label to improve the distribution properties.
6. Locate and treat duplicate cases.

Lab Steps

1. Make sure that you have completed the setup requirements as described in the Lab Overview section.
2. Now, run jupyter notebook and open the “DataPreparation.ipynb” notebook under Module 3 folder.
3. Examine the notebook and answer the questions along the way.

11. What can you conclude about aggregating the hardtop and convertible categories to hardtop_convert?

1

[✓ ○ A. It seems like a good idea because hardtop_convert category does appear to have values distinct from the other body style.]

○ B. It seems like a bad idea because hardtop_convert category does appear to have values distinct from the other body style.

○ C. It seems like a good idea because hardtop_convert category does NOT appear to have values distinct from the other body style.

○ D. It seems like a bad idea because hardtop_convert category does NOT appear to have values distinct from the other body style.

2

12. From the scatter plots, it appears that the relationships between curb_weight and log_price and city_mpg and log_price are more linear, compared to the relationships between curb_weight and price and city_mpg and price respectively. What can you conclude from that?

?

○ A. It is likely that curb_weight is better in predicting log_price than city_mpg.

○ B. It is likely that curb_weight is better in predicting price than city_mpg.

○ C. It is likely that curb_weight is better in predicting log_price than price.

[✓ ○ D. It is likely that city_mpg is better in predicting log_price than curb_weight.]

3

13. How many cases have duplicates?

[✓ ○ A. 12]

○ B. 22

○ C. 1000

○ D. 1012

Lab Steps

1. Make sure that you have completed the setup requirements as described in the Lab Overview section.
2. Now, run jupyter notebook and open the “IntroductionToRegression.ipynb” notebook under Module 4 folder.
3. Now, run jupyter notebook and open the “ApplyingLinearRegression.ipynb” notebook under Module 4 folder.
4. Now, run jupyter notebook and open the “Classification.ipynb” notebook under Module 4 folder.

14. What is the RMSE for the prediction?

○ A. 1.0178480188322825

○ B. 1.008884541873986

○ C. 0.763059846639255

○ D. 0.8957743904421017

15. During the one-hot encoding process, the five categorical features were converted to 14 dummy variables. How many dummy variables came from the num_of_cylinders feature?

○ A. 1

○ B. 2

○ C. 3

○ D. 4

16. What is the RMSE for the prediction?

○ A. 0.15038686027841155

○ B. 0.11912678436796194

○ C. 0.10696449432930688

○ D. 0.921638632102606

1 submissions remaining Submit

You have reached the end of the lesson. Click and access the next lesson tab on the left side of the page. Feel free to post your queries and questions in the **Q&A Parking Lot** on the left tab.

Be the first to like this Like Subscribe Subpages

Back L1: Reading Assignment (https://www.openlearning.com/unitar/courses/2023-05-itim5123-machine-learning/l1_reading_assignment/)

Next L2: Reading Assignment (https://www.openlearning.com/unitar/courses/2023-05-itim5123-machine-learning/l2_reading_assignment/)

TOOLS & RESOURCES

Help & Support (/sso/help/?url=https%3A%2F%2Fhelp.openlearning.com%2F)

Contact us (/contact/)

Learning design toolkit (https://medium.com/learning-designers-toolkit)

Verify a certificate (/certificate-verification/)

Integrations (/sso/help/?url=https%3A%2F%2Fhelp.openlearning.com%2Fcategory%2Fintegrations)

Status (https://status.openlearning.com/)

PLATFORM

Philosophy (https://solutions.openlearning.com/learning-philosophy/)

Features (https://solutions.openlearning.com/platform-features/)

Pricing (https://solutions.openlearning.com/pricing/)

OpenCreds (https://solutions.openlearning.com/opencreds/)

Partners (/partners/)

Browse all courses (/courses/)

Create a course (/courses/create/)

COMPANY

About us (https://solutions.openlearn

Team (https://solutions.openlearn

Careers (https://solutions.openlearn

Press (https://solutions.openlearn

Investors (https://solutions.openlearn home/)

Partnerships (https://solutions.openlearn

(https://twitter.com/openlrning)
(https://www.instagram.com/openlearning_global)
(https://facebook.com/OpenLearning)
(https://www.linkedin.com/company/openlearningcom/)
(https://www.youtube.com/user/openlearningcom)

Terms of service (https://solutions.openlearning.com/terms-of-service/)

Privacy policy (https://solutions.openlearning.com/privacy-policy/)