ELG 20225: Applied Machine Learning

Assignment 2

Due date posted in Bright Space

Submission

You must submit two documents. First, a report of the solutions including important code snippets as a PDF file. Second, the whole code should be in a separate python file (Notebooks are accepted). The file name must include your group number and assignment number, for example **Group1_HW2.pdf and Group1_HW2.py**.

Assignment must be submitted on-line with Bright Space. This is the only method by which we accept assignment submissions. We do not accept assignments sent via email, and we are not able to enter a mark if the assignment is not submitted on Bright Space! The deadline date is firm since you cannot submit an assignment passed the deadline. It is your responsibility to ensure that the assignment has been submitted properly.

Part 1: Calculations

1. Suppose we have some data collected from a cloth shop, and the dataset contains three features. The first feature is the cloth color (x1), the second feature is the consumer's gender(x2), and the third feature is the price (x3) (we simplify the problem and use high, medium and low to present different prices). The label TARGET (y) is whether the consumer buy the cloth at last. Suppose we have the following training data including 15 training samples. Using Bayesian Rule Based Classifier to make prediction when Color = G, Gender = F, Price=H . Please include the detailed calculation process. **(20 Marks)**

   Notes: In the color row, R, G, and Y are short for Red, Green and Yellow; in the Gender row, M and F mean Male and Female, respectively; in the Price row, H, M and L stand for High Prices, Medium Price and Low Prices, respectively and in the Target row, N and Y are No and Yes.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Color(x1)	R	R	R	R	R	G	G	G	G	G	Y	Y	Y	Y	Y
Gender(x2)	M	M	F	F	M	M	M	M	F	F	F	F	M	M	F
Price(x3)	H	L	L	H	M	M	H	L	L	M	L	H	M	L	M
TARGET(y)	N	N	Y	Y	N	N	N	Y	Y	Y	Y	Y	Y	Y	N

2. Consider the following loss table, which contains three actions and two classes. Calculate the expected risk of three actions, and determine the rejection area of P(Class1| x). **(20 Marks)**

Target	Class1	Class2
a1(Choose Class1)	5	2
a2(Choose Class2)	0	5
a3(Rejection)	4	4

Part 2: Programming

1. Use scikitlearn or other python packages to implement a naïve Bayesian classifier (GaussianNB), and show the precision, recall, F1 of the testing set. Use wine dataset in the question, which you can load with the following python codes: **(20 Marks)**

```
from sklearn.datasets import load_wine
data = load_wine()
```

   There are 3 classes in this dataset, and each sample in this dataset has 13 features.

   (a) You need to use train_test_split function in scikitlearn to split the dataset into a training set, a testing set. **(5 Marks)**

   (b) Consider to use classification_report function to help you calculate precision, recall and f1. **(5 Marks)**

   (c) Plot the decision boundary on the test set (for simplicity, you can only consider 2 features when plotting the decision boundary in this question). **(10 Marks)**

2. Use scikit-learn or other python packages to implement a KNN classifier (KNeighborsClassifier). In this question, we use car-evaluation-dataset, which can be downloaded from their official website or Kaggle: **(40 Marks)**

   (a) In this dataset, there are 1728 samples in total. Firstly, you need to shuffle the dataset and split the dataset into a training set with 1000 samples and a validation set with 300 samples and a testing set with 428 samples. Use python to implement this data preparation step. **(5 Marks)**

   (b) Since some attributes are represented by string values. If we choose a distance metric like Euclidean distance, we need to transform the string values into numbers. Use python to implement this preprocessing step. **(5 Marks)**

   (c) Try to use different number of training samples to show the impact of number of training samples. Use 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% and 100% of the training set for 10 separate KNN classifiers and show their performance (accuracy score) on the validation set and testing set. You can specify a fixed K=2 value (nearest neighbor) in this question. Notably, X axis is the portion of the training set, Y axis should be the accuracy score. There should be two lines in total, one is for the validation set and another is for the testing set. **(10 Marks)**

   (d) Use 100% of training samples, try to find the best K value, and show the accuracy curve on the validation set when K varies from 1 to 10. **(5 Marks)**

(e) Analysis the training time when use different number of training samples. Consider the following 4 cases: **(10 Marks)**

- 10% of the whole training set and K = 2
- 100% of the whole training set and K = 2
- 10% of the whole training set and K = 10
- 100% of the whole training set and K = 10.

Plot a bar chart figure to show the prediction time on the testing set.

(f) Provide your conclusions from the experiments of question (c), (d) and (e) in this question. **(5 Marks)**

Important Note

Report should include answers for all question briefly. All plots must have titles and proper axis labels. **Otherwise, you will lose one point for each missing item.** The code file is requested in case of need to verify.