School of Computing and Information Systems
The University of Melbourne
COMP90049, Introduction to Machine Learning, Semester 2 2022

Assignment 2: Exploring the Naive Bayes Classifiers

Due:	5pm, Friday, September 2
Submission:	Jupyter Notebook with source code (in Python) and (inline) responses
Marks:	The Project will be marked out of 20, and will contribute 20% of your total mark. This will be appropriately weighted between implementation and responses to the questions.

Overview

In this Project, you will implement a Naive Bayes classifier, apply it to various classification datasets, and explore evaluation paradigms as well as the impact of individual features. You will then answer some conceptual questions about the Naive Bayes classifier, based on your observations.

Naive Bayes classifiers

The "Naive Bayes" lecture included some suggestions for implementing your learner. You should implement your Naive Bayes classifier from scratch with epsilon smoothing strategy (i.e., do *not* use existing implementations/learning algorithms from libraries like sklearn). Otherwise, you may decide on the specifics of your implementation and may use libraries to help you with data processing, visualization, evaluation, or mathematical operations. For marking purposes, a minimal submission will include the following functions:

- `preprocess()`, which opens the data file, and converts it into a usable format.[0.5 mark]
- `train()`, where you calculate statistics from the training data, to build a Naive Bayes(NB) model.[3 marks]
- `predict()`, where you use the model from train() to predict a class (or class distribution) for the test data.[1.5 marks]
- `evaluate()`, where you will output your evaluation metric(s).[1 mark]
- `main()`, where you call the above functions in order to train and test the Naive Bayes classifier on the full data sets provided (i.e., no train-test splitting). [1 mark]

The assignment materials include an iPython notebook 2022S2-a2.ipynb that summarises these, which you should use as a template. You may define the function inputs and outputs to suit your needs, and you may write other helper functions as you require. Please place the jupyter notebook into the same folder as the input data.

Data Sets

This assignment includes adapted versions of three data sets from the UCI machine learning repository (https://archive.ics.uci.edu/ml/index.html):

Bank Marketing, Obesity, and Student

These datasets vary in terms of number of instances, number of attributes, number of different class labels. For the purpose of this assignment, you should assume that all features are nominal. Each data set is provided in .csv format with one instance per line. The first column (ID) contains a unique instance identifier. The last column (Label) specifies the class label. All other columns specify the data-set specific features. We briefly describe each data set below. The README provided as part of this assignment lists and explains all features and labels (note that not necessarily all original instances or features are included in our data sets).

Bank Marketing You predict if a client will subscribe a term deposit depending on a number of personal and financial features such as job, education level, housing loan, etc.

Obesity You predict whether a patient is obese or not based on various personal and habitual attributes such as alcohol consumption, exercise level, gender, etc.

Student You predict a student's final grade {A+, A, B, C, D, F} based on a number of personal and performance related attributes, such as school, parent's education level, number of absences, etc.

Your submission must automatically process every one of these datasets. As for the questions 1–4, it is technically possible to answer each question by examining only two of the datasets. However, it is strongly recommended that you examine all of the data available, so that you reduce the likelihood that you arrive at faulty conclusions due to a small sample space.

1 Implementation Tips

The "Naive Bayes" lecture included several tips on how to implement the classifier. At training time, you will need to fill in data structures that hold the prior class probabilities $P(c_j)$ as well as data structures that hold the parameters of the likelihood for each feature under each class, i.e., $P(x_i|c_j)$.

At prediction time, you will combine the prior and likelihood terms (one per feature) into a final prediction score:

$$P(c_j)\prod_i P(x_i|c_j)$$

Multiplying a large number of probabilities can lead to underflow. You can equivalently add the log probabilities to avoid numerical instability of your solution:

$$logP(c_j) + \sum_i logP(x_i|c_j)$$

Questions

After implementing and running your classifier, the following questions will give you the chance to think more deeply about its performance. You should answer all questions. Question weights are indicated below. When responding, you should refer to the data wherever possible:

- Question 1[3 marks]

  a In order to understand whether our machine learning model is performing well, we typically compare its performance against alternative models. Implement a One-R baseline model from scratch as introduced in the evaluation lecture. Print the feature name and its corresponding error rate that your One-R classifier selects. (do *not* use existing implementations/learning algorithms from libraries like sklearn)

  b How does the performance of the Naive Bayes classifier compare against your baseline model? Explain your observations on Student and one additional dataset of your choice and justify your response.

- Question 2[3 marks]

  Evaluating the model on the same data that was used to train the model is considered to be a major mistake in Machine Learning. Implement a `cross-validation` evaluation strategy, and inspect how your estimate of effectiveness changes, compared to testing on the training data? Explain why. You must consider Bank Marketing and one additional dataset of your choice in your answer.

- Question 3[3 marks]

  a The "Feature selection" lecture discussed properties of valuable features. Using Mutual information, plotting and data visualization, explore the utility of different features in the Obesity and one additional dataset of your choice. Are different features of different utility to your classification task? Explain why. [Your notebook should display all the plots and numbers you refer to in your answer.]

  b Explain the 'naivety' assumption underlying Naive Bayes. (1) Why is it necessary? (2) Why can it be problematic? Ground your discussion in the features of two (or all) of the data sets provided for this assignment.

- Question 4 [4 marks]

  As machine learning practitioners, we should be aware of possible ethical considerations around the applications we develop. The classifier you developed in this assignment could for example be used to classify college applicants into `admitted` vs `not-admitted` – depending on their predicted grade in the Student dataset.

  a Discuss ethical problems which might arise in this application and lead to unfair treatment of the applicants. Ground your discussion in the set of features provided in the student data set.

  b Remove all ethically problematic features from the data set (use your own judgment), and train your Naive Bayes classifier on the resulting data set. How does the performance change in comparison to the full classifier?

c The approach to fairness we have adopted is called "fairness through unawareness" – we simply deleted any questionable features from our data. Is removing all problematic features as done in part (b) guarantee a fair classifier? Explain Why or Why not?

Submission

Submission will be made via the LMS, as a single jupyter notebook file. You should answer all questions directly in the notebook. You may want to use markdown in your notebook to format your answer.

The submission procedure, late submission policy and academic misconduct information is specified in the Jupyter Notebook template provided with this assignment.

Data References

P. Cortez and A. Silva.
Using Data Mining to Predict Secondary School Student Performance.
In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7
https://archive.ics.uci.edu/ml/datasets/student+performance#

S. Moro, P. Cortez and P. Rita.
A Data-Driven Approach to Predict the Success of Bank Telemarketing.
In Decision Support Systems, Elsevier, 62:22-31, June 2014
https://archive.ics.uci.edu/ml/datasets/bank+marketing

F. M. Palechor and A. de la Hoz Manotas.
Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico.
In Data in Brief, 104344.
https://archive.ics.uci.edu/ml/datasets/Estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition+