MATH60629A. Homework.

Due date: October 18, 2024

Instructions:

- Please include your name and HEC ID with submission.
- The homework is due by 11:59pm on the due date. Please upload a PDF version of your assignment on ZoneCours.
- The homework is worth 20% of the course's final grade.
- Assignments are to be done individually.

1 Machine Learning principles (3pt)

Your answers to these questions should be short (max. 5 lines).

1. (1pt) Explain the difference between the training error and the generalization error. Make sure to describe how to estimate the generalization error of a model in practice including pitfalls of this approach (in other words, describe conditions under which this approach is reliable).
2. (1pt) Explain the effect of varying the regularization strength (for example when using Ridge regularization) on the bias-variance trade off.
3. (1pt) A colleague trained a model to predict the median price of houses using Montreal-housing data. The resulting model works well according to their validation data. However, this model performs poorly when used to predict the prices of houses in the Toronto housing market. Explain why that is and suggest a way of obtaining a better model for the Toronto data.

2 Classification (16pt)

- We will use a synthetic dataset that I created. It is available here.

  Once the data are accessible from your current working directory, you can load them using the following code:

```
data = np.load("a24_devoir_q2-classification.npz")
X = data["X"]
y = data["y"]
```

The data are encoded in a *numpy array*.

The task at hand is to predict the class of each datum from its two features.

1. (4pt) Following an initial data exploration, what do you notice? What can you say about the approximate test performance (in terms of accuracy) if using a a model that uses only linear decision boundaries?
2. (2pt) Divide your dataset into a training and a test set. The test set must make up 20% of the total original dataset. Make sure to use this parameter upon calling the appropriate sklearn function: `random_state=123`. We ask that you provide the few lines of code you used to divide the data.
3. (1pt) What does the fact that we do not keep a validation set tells you?

You will now explore the Naive Bayes model.

5. (1pt) Which Naive Bayes classifier should you use? In other words, what is a reasonable distribution for the class conditional densities? Justify your answer.
6. (2pt) Train the model on the training data and obtain its accuracy for both train and test. Please provide the few lines of code to do these and provide the resulting train and test accuracies.
7. (4pt) Re-train the model using uniform class priors (provide the code for doing so). What do you notice in terms of the training and test accuracies? Explain this result.
8. (2pt) Do you think this observation is likely to hold in general?

3 Regression (38pt)

You will now train a k-NN and a neural network model for the task of predicting the rating of a text review.

The data to download are here. Each datum is a review in text format of an Amazon product.[1]

In the data file, each line corresponds to a datum. Each datum contains a target (y) followed by a short text (x). The target variable is the rating given

[1] For your information, the complete datasets are available here. We use a subset of the *Toys and Games* category.

by a user to a product. It is an integer value between 1 and 5. The text is the review. To pre-process the data you will first have to separate the targets from the features.

Hint: you can use the `split('\t')` function. There are also functions in pandas that will allow you to easily load this dataset.

1. (2pt) We will model this task as a regression problem (and so you can use mean squared error to measure performance). List one advantage and one disadvantage of instead modelling the problem as a classification problem with 5 classes.

Data pre-processing (5pt)

2. (2pt) First divide the datasets into training (80% of the data), validation (10%), and test sets (10%). For this set the random seed to 1234 (`random_state=1234`)

   Provide the few lines of code you used to divide the data in two.

3. (3pt) Now you must obtain a bag-of-words representation of the features. sklearn provides several functions for doing so.

   To limit the required training time, please use a maximum of 2,000 words in your vocabulary (`max_features=2000`) and the list of English stop words from sklearn (`stop_words="english"`). Words on this list will be automatically removed from the data since they are, a priori, less predictive for the task at hand. Use the default value for all other function parameters.

   Provide the few lines of sklearn code you used to encode (and only those) the training, validation, and test data.

K-nearest neighbours (13pt)

4. (3pt) Which of the following three distance functions 'cosine', 'euclidean', and 'manhattan' do you deem more appropriate for this problem? Please justify.

5. (5pt) Given your previous answer, train an adequate k-NN model for this task. We ask that you train models with 1, 10, 50, 100, and 1000 neighbours.

   What is the performance of each model on the training and validation sets?

   Provide the few lines of code that you used to train this model and evaluate its performance

6. (5pt) What value of the hyperparameter provides the best results? Explain.

Neural networks (10pt)

Upon instantiating your neural networks, fix the random seed to 1234 (that is `random_state=1234`).

7. (5pt) You will now train a series of neural networks using different hyperparameters. Use the option `early_stopping=True` and find the hyperparameters that obtain the best results on the validation dataset (to give you an idea, I imagine that you will train around 50 different models). I suggest that you explore the following three hyperarameters: learning rate, size of the network, and the strength of the L2 regularization term.

   In your answer, please include the code used to train these networks and to obtain the performance on the validation set (and only that part of the code).

   What is the best combination of hyperparameters and what is there performance on the validation set?

8. (5pt) What did you find out about the importance of the various hyperparameters?

Comparison (8pt)

9. (4pt) If you were to keep a single feature (that is a single word), which one would it be and why?

10. (2pt) What is the final performance of each model (k-NN and neural network)? *Please provide the few lines of code you used for this.*

11. (2pt) Find an example for which the prediction of the two models differ by more than 2.0. Explain the reason behind this difference in prediction. (You can come up with new examples for this.)

Online system (11pt)

In this question you are tasked with finding a system online that could be making use of machine learning for making predictions (of course, you might not know for sure whether it does or does not). One example of such a system is Google Translate, but you cannot use this example. We ask you to explore the system along several axes.

1. (2pt) Describe the ML task the system is trying to solve.

2. (3pt) Describe data (or its type) that would be needed to train this model and a suitable performance measure for it (you don't necessarily need to use a measure that exists).

3. (2pt) What are the limitations of the system? List at least two, ideally you would demonstrate their effect on the system.

4. (4pt) For each limitation above, given you machine learning knowledge, how would you try to correct it?