MATH60629A. Homework.

Due date: October 24, 2022

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 (12pt)

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

1. (4pt) 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. (4pt) Explain how to obtain the probability of the dependant variable (y) conditioned on the independent variables (x) in a probabilistic model, that is $P(y \mid x)$.
3. (2pt) Explain the effect of varying the regularization strength (for example when using Ridge) on the bias-variance trade off.
4. (2pt) 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 (18pt)

- 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("a22_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 linear model (a model that uses only linear decision boundaries)?
  2. (6pt) Divide your dataset into training, validation, and test sets. The validation and test sets must each make up 20% of the total original dataset (so 40% in total). Make sure to use this parameter upon calling the appropriate sklearn function: `random_state=1234`.

     Train a linear SVM on the training set for each one of these C hyperparameter values: $\{0.001, 0.01, 0.1, 1, 10\}$.

     For each value of C, what is the performance (accuracy) of the model on the training and validation sets?

     Given your answer, obtain the performance of the best model on the test set.

     We ask that you provide the few lines of code you used to divide the data, train the model, and obtain the accuracy.
  3. (4pt) Retrain the SVM model using 10-fold cross-validation for each of the C hyperparameter values from above.

     For each value of C, provide the training and validation accuracies as well as the performance on the test set of the best model.

     Careful that you must use the same test set in both cases (previous question and this question)!

     We ask that you provide the few lines of codes you used to divide the data, train the model, and obtain all accuracies.
  4. (2pt) Explain precisely how is the validation performance evaluated when doing cross validation.
  5. (2pt) Do you obtain a better model with cross validation or without it? Justify your answer and explain your result.

3 Regression (38pt)

You will now train a k-NN and neural network models 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 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 three.

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 yo 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.

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

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.

~~Naive Bayes (10pt)~~ Update: Naive Bayes is for classification and so it does not make much sense to use it in this regression setting. Work on this question will be considereed for bonus points.

7. ~~(2pt) Train the appropriate Naive Bayes model for this task. Which Naive Bayes model did you use?~~

8. ~~(3pt) What is the performance of this model on the training and validation sets?~~

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

9. ~~(5pt) The Naive Bayes model makes a naive assumption. Given your trained model, demonstrate using two examples (from your dataset or that you came up with) one limit of this assumption on the prediction of the models.~~

Neural Networks (10pt)

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

10. (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?

11. (5pt) What did you find out about the importance of the various hyperparameters?

Comparison (8pt)

12. (4pt) If you were to keep a single feature (that is a single word), which one would it be and why?

13. (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.*

14. (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.)