Homework Assignment 2

Caregivers & Machine Learning Program 2023 - Vector Institute

Deadline: Monday, April 3rd, 2023 at 11:59 pm
Submission: Submit answers to theory questions (`hw2_yourname.pdf/tex/doc/etc)` and practical questions `hw2_yourname.ipynb/py)` through `https://learn.vectorinstitue.ai` course page.

1. **Theory** [25 pts]

   (a) Explain what a decision tree is. How does it make classification predictions? [5 pts]

   (b) In a classification problem, using your own words explain how attributes are split. What mathematical principle can be used to define a good split? [5 pts]

   (c) What is information gain and why it is useful in the training of a decision tree? [5 pts]

   (d) Given training, validation, and test sets, explain the process of selecting the best hyperparameters of a classification algorithm. [5 pts]

   (e) If a Decision Tree is overfitting the training set, is it a good idea to try decreasing or increasing maximum depth? Why? [5 pts]

   (f) Information theory. Recall the definition of the entropy of a discrete random variable X with probability mass function p: [10 bonus pts/optional]

   $$H(X) = \sum_{x} p(x) log_2(1/p(x)) \tag{1}$$

   Here the summation is over all possible values of $x \in X$, which (for simplicity) we assume is finite (e.g., $1, 2, ..., N$). Prove that the entropy H(X) is non-negative [1 pt].

2. **Practical** [25 pts]

   (a) Utilize ChatGPT (`https://chat.openai.com/auth/login`) to provide you with example python code of a decision tree classifier for one of the classification sklearn.datasets (e.g., iris/breast cancer, wine) with training, validation, and test splits. This implementation will produce an `X_train, X_val, X_test, y_train, y_val and y_test`. [1 pt]

   (b) Why do we need training/test splits? Why training/validation/test spltis?[5 pts]

   (c) What are the dimensions d of the `X_test` matrix and what do they mean? How many data points are in the validation, training, and test sets? (HINT: use 'shape') [5 pts]

(d) Prompt ChatGPT to write you a function (e.g., selectmodel) which trains a decision tree classifier using 2 different values of `max_depth`, as well as two different split criteria (information gain and Gini coefficient), evaluates the performance of each one on the validation set, and prints the resuls of each model. Expand the number of `max_depth` values to 10 and run your function [2]

(e) What are the best choices of hyperparameters (split criteria, depth of the tree)? Explain your answer. What is the test accuracy of the best model? [3 pt]

(f) Use the best hyperparameters you found in (e). Visualize the first two layers of the tree. Decision trees are easily interpretable. Why is the clarification of the machine learning decision process important? [5 pts]

(g) With the help of ChatGPT, replace the decision tree classifier with Random forest or XGBoost. XGBoost is currently the state of the art for heterogeneous tabular data and builds decision trees sequentially such that each subsequent tree corrects earlier mistakes. XGBoost can reveal important features in the decision process. Identify the most important features for your classification task. [5 pts]

`https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html`