COMPSCI 762 2022 S1 Week 4 Questions – Fundamentals of Learning

Luke Chang

March 17, 2022

Question 1

There are many ways to evaluate the performance of a machine learning model. Your team decides to try out different train-test splitting strategies as the following: 2-fold *Cross Validation* (CV), 10-fold CV, leave-one-out, and using a single 70/30 percent train/validation split.

1. What effect will this have on your results? (Explain it in term of sample size, bias and variance.)
2. Which strategy will give you the best representative value to the "unseen test set"?
3. Does it matter how large your original dataset is? Will you use a different approach if the dataset is very big (e.g., 142.8 million user reviews from Amazon) or very small (e.g., 54 compounds in a molecular dataset).
4. Rank the methods above from fastest to slowest in terms of computation, and explain why.

Question 2

2.1 Part 1

Will it rain tomorrow? Your team built a machine learning model to predict the weather based on the humidity. The actual weather and predictions were recorded in Table 1. Answer the following questions based on Table 1 the results above:

Day	True Label	Prediction
1	Shower	Shower
2	Clear	Shower
3	Shower	Clear
4	Shower	Shower
5	Clear	Shower
6	Shower	Shower
7	Clear	Shower
8	Clear	Clear
9	Clear	Clear
10	Shower	Shower

Table 1: A model that predicts the weather of tomorrow.

Day	True Label	Probability Estimate
1	Shower	0.95
2	Clear	0.85
3	Shower	0.78
4	Shower	0.66
5	Clear	0.6
6	Shower	0.55
7	Clear	0.53
8	Clear	0.52
9	Clear	0.51
10	Shower	0.4

Table 2: The probability estimates indicate how much the predictor believes it will rain tomorrow.

1. Draw a confusion matrix.
2. Compute Accuracy, Precision, Recall, and F1 score.
3. Can you evaluate the performance of this predictor based on these metrics?

2.2 Part 2

You noticed that the many classifiers implemented in `scikit-learn` (Link to the documentation) can also output probability estimates. Table 2 shows the probability estimates instead of predicting the class labels. Can you draw a *Receiver Operating Characteristic* (ROC) curve and compute the *Area Under the Curve* (AUC) based on Table 2? Assuming the thresholds are from 0 to 1 with a 0.2 increment.

Question 3

Your team decides to participate a machine learning competition on Kaggle to win the 100k prize. After data preprocessing and initial analysis, you decide to use a decision tree classifier, but not sure about the optimal hyper-parameters. One of your teammate suggests to use a 5-fold CV. The dataset which provided by Kaggle has 10 million samples. Explain how to split the data, train the model, and which model you will submit to Kaggle.