Reading assignment 2

Minzhe Zhang mxz163730

Chapter 1:

1. What are the phases in the ML lifecycle? What type of evaluation is suitable for each phase?

Mainly two. First is involves prototyping, we try out different models to find the best one. Then we deploy it into production, and validate it in live data.

The evaluation for first phase is called offline evaluation, the evaluation for second phase is called online evaluation.

2. Why is the process of evaluation complicated?

First, online and offline evaluation use different measurement matrics. Second, historical data and live data may change over time, which is called distribution drift.

3. What is the difference between model parameter and hyperparameter? Give examples of each for different models that you have learned. For example, in case of decision tree, what would be a model parameter and what would be a hyperparameter?

   A model parameter is a variable that is internal to the model which can be estimated from data, and is part of the model when.

   A model hyperparameter is a variable external to the model which cannot be estimated from data, and should be tunned manually.

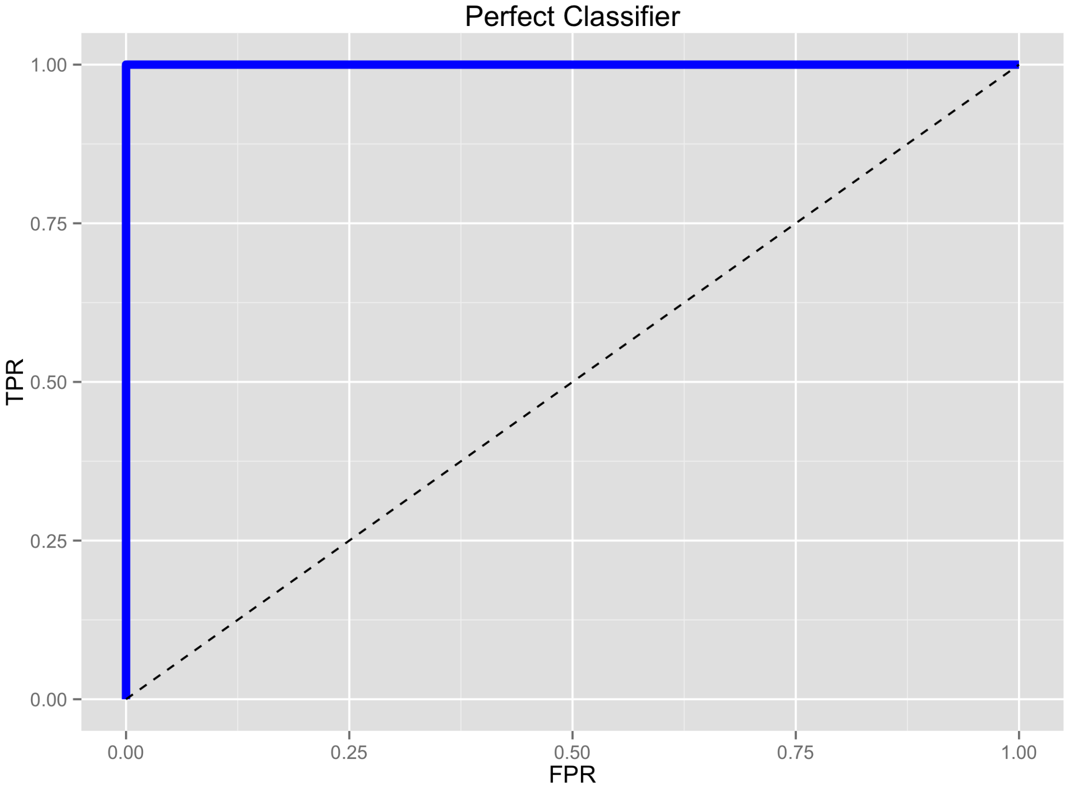
   - Decision Tree: model parameters would be which parameter used to split, what value cutoff used to split the node; model hyperparameter would be the depth of the tree, pruning factor etc.
   - Neural Network: model parameters would be weights for each node; model hyperparameter would be the depth of the hidden layer, number of nodes in a hidden layer etc.
   - SVM: model parameters would be which data points to select as support vector; model hyperparameter would be which kernel used to transform data etc.
   - Logistic Regression: model parameters would be weight for each parameter; model hyperparameter would be regularization term etc.

Chapter 2:

1. What would be the equation for log-loss metric for a binary class dataset when using Logistic Regression as the model? Write the equation in the most simplified format.

$$\text{log-loss} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log p_i + (1 - y_i)\log(1 - p_i)$$

2. Read the section on ROC and draw the curve for the perfect classifier i.e. one that makes no mistakes.



3. Understand the concepts of precision, recall, and F1 score. Is it possible for precision and recall both to go up at the same time i.e. do they have a positive or negative correlation? Explain.

They can both go up at the same time, they have positive relationship. Because precision is true positive divided by the sum of true positive and false positive, recall is true positive divided by the sum of true positive and false negative. If the true positive goes up, precision and recall should also both go up.

Chapter 3:

1. Why is model selection and hyperparameter tuning done using results of validation dataset and not training dataset?

   If we tune the hyperparameter in training data, the model can easily be overfitted, because we are using the training data twice.

2. If you are given just one dataset, what are the 3 ways in which you can obtain validation dataset(s) from the given dataset? Explain each and give advantages and disadvantages of each.

   - Hold-out validation: simply randomly hold out part of the data for validation. It is simple to program and fast to run, but it does not use data effectively.
   - Cross-validation: data is used more effectively than hold-out validation, especially when dataset is small.
   - Bootstrap: it is a resampling technique, one can obtain new datasets from given dataset using bootstrap, but it may contain same instance for multiple times.

4. We live in an era of Big Data. Why don' t we just get more data rather than obtaining validation data from given data?

   Data collection is expensive, and not easy to manage, sometimes it is even impossible to collect enough data from real world.

5. Why should training, and evaluation data never be mixed?

   Otherwise, we don' t have a fair way to evaluate the model since the information in validation data is leaked. The model will learn from it and be overfitted.

Chapter 4:

1. What is the role of a hyperparameter in a model? When are model parameters learned? Do you think a model is more strongly influenced by model parameters or hyperparameters? Explain you reasoning?

   Hyperparameter usually control some trade-off in model parameter estimation, like accuracy-overfitting trade off, accuracy-speed trade-off, which model itself can not decide.

   Model parameters are learned when model is given the training data.

   I think model parameters influence the model more. As I mentioned above, hyperparameters usually control some trade-off, it helps to make your model better. While parameters are what your model learned directly from data, it control if your model will work or not.

2. What are some techniques for hyperparameter tuning? Explain each briefly and in your own words.
   - Grid search: list possible hyperparameters, try them out and choose the one that gives the best result.
   - Random search: randomly search some of the hyperparameters in the whole spce in grid search, actually in statistics, it will have high probability that we can find good hyperparameters in the top. It is computationaly more efficient.
   - Smart tuning: compare to grid o random search, each time, the algorithm will evaluate the hyperparameters in this run and then decide the direction to search in the next run.