Machine Learning

Assignment 9.2

Submitted By: Ranji Raj

January 6, 2021

a) Difference between regression and classification

Algorithm	labels/classes	attributes
Linear Regression	Continuous	Numeric
Regression Tree	Continuous	Categorical/Numeric
Locally Weighted Regression	Continuous	Numeric

Table 1: Regression algorithms

Reasoning

- For a Regression Tree algorithm the predictor/attributes can be either categorical or continuous.
- For Linear Regression and Locally Weighted Regression where there involves a mathematical computation ($\sum_{i=1}^{n} w_i x_i$) for predicting the target value all the predictors/attributes must strictly be transformed to continuous value.

Algorithm	labels/classes	attributes
Naive Bayes	Categorical/Continuous	Categorical/Numeric
Decision Tree	Categorical	Categorical/Numeric
Logistic Regression	Categorical/Continuous	Numeric
Perceptron	Categorical/Continuous	Numeric
Neural Networks	Categorical/Continuous	Numeric
kNN	Categorical/Continuous	Numeric

Table 2: Classification algorithms

Reasoning

- For Naive Bayes algorithm the predictor/attributes can be either categorical or continuous. For handling continuous attributes there is a provision of multivariate Gaussian estimate.
- For Perceptrons, Logistic Regression and Neural Nets where there involves a mathematical computation ($\sum_{i=1}^{n} w_i x_i$) for classifying the target value all the predictors/attributes must be strictly be transformed to continuous value.
- For Decision Tree the predictor/attributes can be either categorical or continuous. For handling continuous attributes there is a provision of discretization into bins.
- For kNN predictor/attributes must be strictly continuous because distance metrics are used for determining the nearest neighbours.

b) Estimating performance of regression and classification algorithms

- Performance of regression algorithms like linear regression can be estimated by **R-squared** or **Adjusted R-squared** values, whereas, for Regression trees can be done by **Coefficient of Variation**.
- Performance of classification algorithms like Naive Bayes can be determined by **confusion matrix (Precision, Recall, Accuracy)** etc., For perceptrons and neural nets by means of **boxplots**, and for kNN by means of **ROC/AUC**.

c) Overfitting in IB learning

- For **smaller** values of k instance-based algorithms like kNN suffers from overfitting and would produce a non-smooth decision surface.
- Overfitting in Decision Trees is due to lack of data with sufficient distribution.
- Overfitting in Neural nets is when training is halted when the training error is minimum which cannot generalize to unseen data.

d) Optimal k in kNN

- Cross validation run on a validation set can be used to pick the best k.
- If k is low, non-linear functions can be approximated, but also can capture noise. Low bias and high variance (underfitting).
- If k is high, output is much smoother, less sensitive to data variation. high bias and low variance (overfitting).