IST407/707 Data Mining

HW6: naïve Bayes and decision tree Comparative Analysis

Now that we have learned two classification algorithms, decision tree and naïve Bayes, let’s think further on the question of choosing algorithms for a specific task. Note that there is no silver bullet in terms of algorithm comparison – no algorithm would outperform all other algorithms on all data sets. Therefore, choosing appropriate algorithms is an important decision, and it requires knowledge of both the data set and the candidate algorithms.

Task description:

A telecommunications company is concerned about the number of customers leaving their landline business for cable competitors. They need to understand who is leaving. Imagine that you are an analyst at this company and you have to find out who is leaving and why.

The data set includes information about:

* Customers who left within the last month – the column is called Churn
* Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies
* Customer account information – how long they’ve been a customer, contract, payment method, paperless billing, monthly charges, and total charges
* Demographic info about customers – gender, age range, and if they have partners and dependents

**Report structure:**

Section 1: Introduction

Briefly describe the classification problem and general data preprocessing. Note that some data preprocessing steps maybe specific to a particular algorithm. Report those steps under each algorithm section.

Section 2: Decision tree

Build a decision tree model. Tune the parameters, such as the pruning options, and report the 3-fold CV accuracy.

Section 3: Naïve Bayes

Build a naïve Bayes model. Tune the parameters, such as the discretization options, to compare results.

Section 4: Algorithm performance comparison

Compare the results from the two algorithms. Which one reached higher accuracy? Which one runs faster? Can you explain why?

Section 5: Model Evaluation

Referencing our learnings on Model Evaluation (ie:Precision,Recall,F-Measure), which accuracy measure do you believe would be best for the Telco company to use and why? Does the use of an alternative evaluation measure effect your decision on which model you would choose? Why or why not?