COMP6200 - FINAL EXAM

Learning Tasks

- Unsupervised (instances are not labeled): Clustering
- Supervised (instances are labeled): Regression and Classification

Evaluation Metrics for Supervised Learning

- MSE, R-square, Accuracy, Precision, Recall, F1-score.

Tricky Issues

- Overfitting/Underfitting, Curse of Dimensionality, Tuning Hyper-parameters
- Weaknesses and Strengths of Each Model

1. Unsupervised
   Clustering K-means algorithm (compute centroid, clustering points to their closest centroids, update centroids)
   Preprocess: Data normalization
   Overfitting probably occurs if K is a large number

   Hierarchical Clustering algorithm (bottom up learning process by merging two closest points or clusters)
   How to compute the distance between two clusters
   How to choose the number of clusters based on Dendrogram

   Advantages and Disadvantages of KMeans and Hierarchical Clustering

2. Supervised: K Nearest Neighbour algorithm
   Distinguish regression and classification
   A learning task has: 1) Dataset; 2) Learning Algorithm; 3) Performance Measure

   Learning Process
   - Learning with the training dataset and testing with the testing dataset. For KNN, how to define similarity? Euclidean or Hamming distance
   - K is the number of nearest neighbors to look at before classifying a new instance

   Overfitting probably occurs if K is small

   How to split the dataset for training and testing?

   Cross Validation Curse of Dimensionality in KNN Advantages and Disadvantages of KNN

3. **Supervised: Naïve Bayes Model**
   Bayes Rules

$$P(C|\mathbf{X}) = \frac{P(\mathbf{X}|C)P(C)}{P(\mathbf{X})}$$

Perceive features as random variables P(X|C) where C is the class and X is the vector of features.
Given features X, find class C such that

$$\text{Label}(x) \leftarrow \arg\max_{\mathcal{C}_k}\{p(\mathcal{C}_k|\boldsymbol{x})\}$$

Important Assumption: All features are independent random variables
Naïve Bayes model estimates P(Xi|C) separately from the training dataset, where Xi is the feature and C is the class

Objective is to maximize a posterior:

$$[P(x_1|c^*)\cdots P(x_n|c^*)]P(c^*) > [P(x_1|c)\cdots P(x_n|c)]P(c), \qquad c \neq c^*, c = c_1,\cdots,c_L$$

How to deal with continuous random variable
Advantages and Disadvantages of Naïve Bayes model

4. **Supervised: Neural Network**

Logistic Regression for classification
Objective of Logistic Regression: Cross Entropy
Three activation functions for Neural Network
Understand the role of regularization to prevent overfitting

How to train a neural network?
1/ Feedforward Neural Network: How to compute output from input and edge weights
2/ Backpropagation Algorithm: How to backpropagate error from output to hidden layers
3/ Understand the updating of weights

Conduct 1,2,3 iteratively for multiple rounds.
Advantages and disadvantages of Neural Networks

Pros:

- They form the basis of state-of-the-art models and can be formed into advanced architectures that effectively capture complex features given enough data and computation.

Cons:

- Larger, more complex models require significant training time, data, and customization.
- Careful preprocessing of the data is needed.
- A good choice when the features are of similar types, but less so when features of very different types.

5. **Supervised: Decision Tree (classification)**

Constructing a Decision Tree is a process to split a dataset into subsets based on features.
The objective is to purify the labels of instances in each subset
The criteria to split the dataset is Information Gain.
Information Gain is the change of Entropy before and after dataset splitting
Know how to compute Entropy and Information Gain
How to prevent overfitting of a Decision Tree
Advantages and Disadvantages of Decision Trees:
Advantages

- Easy to Understand
- Useful in Data exploration
- Less data cleaning required
    - It is not influenced by outliers and missing values to a fair degree.
- Data type is not a constraint
    - It can handle both numerical and categorical variables.

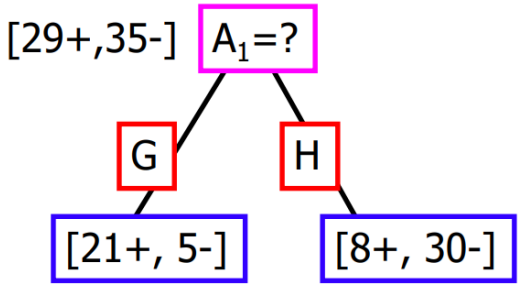
Disadvantages

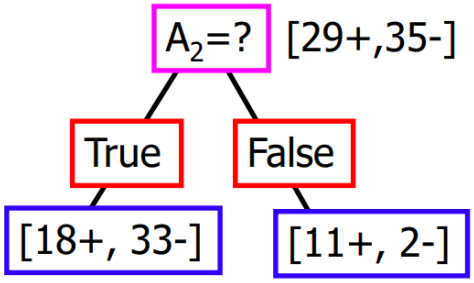
- Over fitting
- Not fit for continuous variables
    - While working with continuous numerical variables, decision tree loses information when it categorizes variables in different categories.

- Gain(S,A): expected reduction in entropy due to sorting S on attribute A

$$Gain(S, A) \equiv Entropy(S) - \sum_{v \in D_A} \frac{|S_v|}{|S|} Entropy(S_v)$$

Entropy([29+,35-]) = -29/64 log2 29/64 – 35/64 log2 35/64
= 0.99





Entropy([21+,5-]) = 0.71
Entropy([8+,30-]) = 0.74
Gain(S,A1)=Entropy(S)
-26/64*Entropy([21+,5-])
-38/64*Entropy([8+,30-])
=0.27

Entropy([18+,33-]) = 0.94
Entropy([11+,2-]) = 0.62
Gain(S,A2)=Entropy(S)
-51/64*Entropy([18+,33-])
-13/64*Entropy([11+,2-])
=0.12