Assignment 1

CS 480/680 Winter 2019

1 Decision Trees [20 pts]

1.1 Decision Trees and Boolean Expressions [10 pts]

A boolean function can be written represented by a decision tree, i.e., the decision tree equivalent to that boolean function will output 1 if the expression is satisfied, 0 otherwise. Please write down a decision tree that is equivalent to the following:

$$(C \wedge \neg A \wedge \neg B \wedge \neg E) \vee (C \wedge A \wedge D) \vee (\neg C \wedge B)$$

1.2 Train a decision Tree [10 pts]

Below is a dataset that determines the survival rate of passengers on the Titanic based on three attributes—Class (C), Gender (G) and Age (A). Train a depth 1 decision tree (i.e., a decision tree with only 1 decision node) using information gain as the criterion for splitting, using $IG(S, x) = H(S) - H(S|x)$, where $x \in \{C, G, A\}$. Which node is at the root of the tree? Show all your work (i.e., include all intermediate steps) in calculating information gain for each attribute. Round to 4 decimal points for all calculations, and use log base 2.

Class	Gender	Age	No	Yes	Total
1st	Male	Child	0	5	5
1st	Male	Adult	118	57	175
1st	Female	Child	0	1	1
1st	Female	Adult	4	140	144
Lower	Male	Child	35	24	59
Lower	Male	Adult	1211	281	1492
Lower	Female	Child	17	27	44
Lower	Female	Adult	105	176	281
		Total:	1490	711	2201

Class	No	Yes	Total
1st	122	203	325
Lower	1368	508	1876

Gender	No	Yes	Total
Male	1364	367	1731
Female	126	344	470

Age	No	Yes	Total
Child	52	57	109
Adult	1438	654	2092

2 K-nearest neighbor [30 pts]

Implement k-nearest neighbours from scratch using any programming language of your choice. Download the dataset (knn-dataset.zip) posted on the course web page, which contains the training data and labels for each fold. Classify each input x according to the most frequent class amongst its k nearest neighbours as measured by the Euclidean distance (L2-norm). Break ties at random. Test the algorithms by 10-fold cross validation.

What to hand in:

- Your code for K-nearest neighbor and cross validation
- Find the best k by 10-fold cross validation. Draw a graph that show the accuracy as k increases from 1 to 30.

3 Perceptron [20 pts]

Consider a threshold perceptron that predicts $y = 1$ when $\mathbf{w}^T\mathbf{x} + w_0 \geq 0$ and $y = 0$ when $\mathbf{w}^T\mathbf{x} + w_0 < 0$. It is interesting to study the class of Boolean functions that can be represented by a threshold perceptron.

Assume that the input space is $\mathbf{X} = \{0, 1\}^2$ and the output space is $Y = \{0, 1\}$. For each of the following Boolean functions, indicate whether it is possible to encode the function as a threshold perceptron. If it is possible, indicate some values for $\mathbf{w}$ and w_0. If it is not possible, indicate a feature mapping $\phi : X \to \hat{X}$ with values for w and w_0 such that $\mathbf{w}^T\phi(\mathbf{x}) + w_0$ is a linear separator that encodes the function.

- and
- or
- exclusive-or
- iff

4 Empirical Evaluation [30 pts]

In this exercise, we will be working with the Breast Cancer Wisconsin (Diagnostic) Dataset (WDBC) from the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets.html). The dataset consists of 569 samples of biopsied tissue. The tissue for each sample is imaged and 10 characteristics of the nuclei of cells present in each image are characterized, including

- radius (mean of distances from center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter
- area
- smoothness (local variation in radius lengths)
- compactness (perimeter^2 / area - 1.0)
- concavity (severity of concave portions of the contour)
- concave points (number of concave portions of the contour)
- symmetry
- fractal dimension (“coastline approximation” - 1)

Each of the 569 samples used in the dataset consists of a feature vector of length 30. The first 10 entries in this feature vector are the mean of the characteristics listed above for each image. The second 10 are the standard deviation and last 10 are the largest value of each of these characteristics present in each image. Each sample is also associated with a classification label, which is M for malignant and B for benign.

The dataset has already been divided into a training set (wdbc-train.csv) and test set (wdbc-test.csv), which you can download from the Assignments page on the course website.

(a) Use your favourite machine learning package, and train (a) a decision tree, (b) kNN algorithm on the WDBC dataset, using all 30 features. Choose the best k by 10-fold cross validation. Make sure you shuffle the training data randomly before splitting it into groups to be used for cross validation.

(b) Report the training performance using 10-fold cross validation, in terms of precision, recall, accuracy, sensitivity and specificity. Assume that M is the positive class.

(c) Report the testing performance, in terms of precision, recall, accuracy, sensitivity and specificity. Assume that M is the positive class.

(d) Suppose you were to recommend an algorithm to a conservative doctor who is afraid of making a mistake in diagnosis, which classifier would you recommend? Explain why?