Assignment II
Due: 23:55 November 27th, 2018

November 13, 2018

1 Decision Tree

Table 1 summarizes a dataset with three attributes A, B, C and two class labels +,-. Each attribute is a binary variable with values T and F. Build a two-level decision tree.

Table 1: Dataset

A	**B**	**C**	**Number of +**	**Number of -**
T	T	T	5	0
F	T	T	0	20
T	F	T	20	0
F	F	T	0	5
T	T	F	0	0
F	T	F	25	0
T	F	F	0	0
F	F	F	0	25

a) According to the misclassification error rate, which attribute would be chosen as the first splitting attribute? For each attribute, show the contingency table and the gains in misclassification error. How about the gains in entropy and Gini index?

b) Repeat for the two children of the root node.

c) How many instances are misclassified by the resulting decision tree?

d) Compute the description length of the tree, which is given by the following equation:

$$Cost(tree, data) = Cost(tree) + Cost(data|tree) \tag{1}$$

Note: You can define the $Cost(tree)$ and $Cost(data|tree)$ by yourself and explain why.

2 SVM

Derive the dual Lagrangian for the liner SVM with non-separable data where the objective function is,

$$f(\mathbf{w}) = \frac{\|\mathbf{w}\|^2}{2} + C(\sum_{i=1}^{N} \xi_i)^2 \tag{2}$$

also write down the KKT conditions for this objective function.

3 Naive Bayes

Use Naive Bayes to classify the dataset in Table 1.

a) Estimate the conditional probabilities for $P(A = T|+)$, $P(B = T|+)$, $P(C = T|+)$, $P(A = T|-)$, $P(B = T|-)$, $P(C = T|-)$

b) Given a test sample $(A = T, B = T, C = T)$, predict the class sample using the Naive Bayes approach.

4 Case Study: Document Classification

In this case, we have two categories of emails, in which one category is about hockey and the other is about baseball.

a) Firstly preprocess the documents into numerical data (Record data). The preprocessing guidelines can be found in the **introduction slides (Assignment II-case)**, consider using $tf - idf$ (referring to Question 1 in Assignment I).

b) Use SVMs to classify the documents and test the classification results with 5-fold cross validation. You should report the precision, recall, and F1-measure of each fold and the average values. (*Recommend LIBSVM to implement SVMs. You can refer to the introductive slides in evaluating the results.*)

c) **Bonus (5 extra points)**. Implement Sequential Minimal Optimization (SMO) by following the introductive slides.

Note: If you have problems in preprocessing the documents, contact the teacher or the TA as soon as possible.

5 XGBoost

Read the paper XGBoost: *A Scalable Tree Boosting System* and summarize the main contributions by answering the following questions.

i) What is the general idea of XGBoost?

ii) How the Gradient Boosting is used in XGBoost?

iii) What techniques are used to prevent overfitting and why?

iv) How does XGBoost handles the sparsity issues of features?