# CMPT 459 Spring 2020
# Martin Ester
# TAs: Ruijia Mao and Ruchita Rozario

## Midterm Exam with Solution

### Problem 1 (Multiple Choice) 24 marks

Mark your answers for the following questions directly in the table with a BIG X. Note that, for each question, multiple answers or none may be correct. When marking this problem, each of the four possible answers will be considered as a TRUE/FALSE question, and you will receive one mark if you correctly marked or correctly not marked that answer.

| | | | | |
|---|---|---|---|---|
| a) Compared to k-means, k-medoid algorithms such as CLARANS . . . | Are more efficient | **Are more generally applicable** | **Are more robust to outliers** | Produce more natural clusters |
| b) The EM clustering algorithm . . . | Assigns every point to exactly one cluster | **Assigns every point to multiple clusters** | Represents a cluster through one point | **Represents a cluster through a probability density distribution** |
| c) Overfitting is an issue for . . . | **Decision tree** | **Linear SVM** | **Logistic regression** | Nearest neighbour classifier |
| d) What are advantages of lazy classifiers compared to eager classifiers? | **shorter training time** | shorter testing time | **local decision surface** | global decision surface |
| e) The following are high-bias classifiers: | Decision tree | **Logistic regression** | Nearest neighbour classifier | **SVM** |
| f) The following ensemble classifiers reduce the bias . . . | Bagging | **Boosting** | Random Forests | **Stacking** |

### Problem 2 (Preprocessing) 14 marks

a) Given two text documents D1 and D2 containing the terms a, b, c, and d as follows:

D1: a, b, c, a, b, c
D2: a, c, c, d

Show the TFIDF (Term Frequency Inverse Document Frequency) vectors for D1 and for D2.

D1 = [2/2, 2/1, 2/2, 0] = [1, 2, 1, 0]
D2 = [0.5, 0, 1, 1]

where the dimensions correspond to terms a, b, c, d in that order.

b) Assume the following dataset of six objects with a single categorial attribute:

(r)
(s)
(r)
(t)
(s)
(u)

Explain how you can transform the categorical attribute into numerical attribute(s). Show the transformed dataset.

We create one numerical (Boolean) attribute per value of the categorical attribute. Exactly one of the attributes, i.e. the one corresponding to the categorical value, will have a value of 1.

(1, 0, 0, 0)
(0, 1, 0, 0)
(1, 0, 0, 0)
(0, 0, 1, 0)
(0, 1, 0, 0)
(0, 0, 0, 1)

**Problem 3 (Distance function) 20 marks**

Assume a dataset of objects $o = (o1, \ldots, o10, o11, \ldots o20)$ where the first 10 attributes are numerical and the second 10 attributes are categorical.

a) Define a distance function for pairs of objects, i.e. dist(p, q).

$$\boldsymbol{dist}(\boldsymbol{p}, \boldsymbol{q}) =$$
$$\boldsymbol{dist}((p1, \ldots, p10, p11, \ldots p20), (q1, \ldots, q10, q11, \ldots q20)) =$$
$$\boldsymbol{\lambda} * \boldsymbol{dist1}\big((p1, \ldots, p10), (q1, \ldots q10)\big) + (1 - \lambda) * \boldsymbol{dist2}\big((p11, \ldots, p20), (q11, \ldots q20)\big)$$

where dist1 is the Euclidean distance and dist2 is the Hamming distance.

b) Show that your proposed function satisfies the three requirements for distance functions.

We use the fact that dist1 and dist2 are both distance functions, i.e. they satisfy the three requirements.
1) Non-negativity
The sum of two non-negative numbers is non-negative.

2) dist(p,q)=0 if and only if p=q
p=q implies $\boldsymbol{dist1}\big((p1,\ldots,p10),(q1,\ldots,q10)\big)=0$ and
$\boldsymbol{dist2}\big((p11,\ldots p20),(q11,\ldots q20)\big)=0$ and therefore
dist(p,q)=0
dist(p,q)=0 implies, because of the non-negativity of dist1 and dist2, that
$\boldsymbol{dist1}\big((p1,\ldots,p10),(q1,\ldots,q10)\big)=0$ and
$\boldsymbol{dist2}\big((p11,\ldots p20),(q11,\ldots q20)\big)=0$
which implies $(p1,\ldots,p10)=(q1,\ldots,q10)$ and
$(p11,\ldots,p20)=(q11,\ldots,q20)$
and therefore p=q
3) Symmetry
Follows from the symmetry of dist1 and dist2.

**Problem 4 (Decision Tree classifier) 27 marks**

Given the following dataset of employees with categorical attributes Department, Status, Age and Salary:

| Department | Status | Age | Salary |
|---|---|---|---|
| Sales | senior | 31..35 | 46K..50K |
| Sales | junior | 26..30 | 26K..30K |
| Sales | junior | 31..35 | 31K..35K |
| Systems | junior | 21..25 | 46K..50K |
| Systems | senior | 31..35 | 66K..70K |
| Systems | junior | 26..30 | 46K..50K |
| Systems | senior | 41..45 | 66K..70K |
| Marketing | senior | 36..40 | 46K..50K |
| Marketing | junior | 31..35 | 41K..45K |
| Secretary | senior | 46..50 | 36K..40K |

We want to use this dataset to train a classifier to predict the Status of an employee, i.e. we are using Status as the class label. Suppose that the gini index is used to determine the split attributes.

a) Show all the relevant gini indexes that you have to compute in order to choose the split attribute in the root.

gini(D) = $1 - \Sigma\, p_i^2$ for i from 1 to 2, i.e. for the classes "senior" and "junior"

gini(D|department) = 3/10*gini(D|department=Sales)
+ 4/10* gini(D|department=Systems)
+ 2/10* gini(D|department=Marketing)
+ 1/10* gini(D|department=Secretary)
= 0.3*0.44 + 0.4*0.5 + 0.2*0.5 + 0.1*0 = 0.132 + 0.2 + 0.1 + 0
= 0.432

gini(D|department=Sales) = $1-(1/3)^2-(2/3)^2 = 1-5/9=4/9=0.44$
gini(D|department=Systems) = $1-(1/2)^2-(1/2)^2 = 1-1/2 = 0.5$
gini(D|department=Marketing) = $1-(1/2)^2-(1/2)^2 = 1-1/2=0.5$
gini(D|department=Secretary) = $1-(1/1)^2= 1-1=0$

gini(D|age) = 2/10*gini(D|age=21) + 1/10*gini(D|age=26) + 4/10*gini(D|age=31)
+ 1/10*gini(D|age=36) + 1/10*gini(D|age=41) + 1/10*gini(D|age=46)
= 0.2*0.0 + 0.1*0.0 + 0.4*0.5 + 0.1*0.0 + 0.1*0.0 + 0.1*0.0
= 0.2
gini(D|age=21) = $1-(1/1)^2 = 0.0$
gini(D|age=26) = $1-(1/1)^2 = 1-1.0=0.0$
gini(D|age=31) = $1-(1/2)^2 -(1/2)^2= 1-1/4 -1/4=0.5$
gini(D|age=36) = $1-(1/1)^2= 1-1.0=0.0$
gini(D|age=41) = $1-(1/1)^2 = 1-1.0=0.0$
gini(D|age=46) = $1-(1/1)^2 = 1-1.0=0.0$

gini(D|salary) = 1/10*gini(D|salary=26) + 1/10*gini(D|salary=31) +
1/10*gini(D|salary=36)
+ 1/10*gini(D|salary=41) + 4/10*gini(D|salary=46) +
2/10*gini(D|salary=66)
= 0.1*0.0 + 0.1*0.0 + 0.1*0.0 + + 0.1*0.0 + 0.4*0.5 + 0.2*0.0
= 0.2
gini(D|salary=26) = $1-(1/1)^2=0.0$
gini(D|salary=31) = $1-(1/1)^2=0.0$
gini(D|salary=36) = $1-(1/1)^2=0.0$
gini(D|salary=41) = $1-(1/1)^2=0.0$
gini(D|salary=46) = $1-(1/2)^2-(1/2)^2=1-1/4 -1/4=0.5$
gini(D|salary=66) = $1-(1/1)^2=0.0$

b) Which split attribute does the decision tree algorithm select for the root of the decision tree?

Age and Salary both have the smallest gini index of 0.2, and either of them can be chosen.

c) Given a labelled training dataset T. When computing the gini index for a numerical attribute A, a decision tree classifier
needs to determine the best possible split point t. If A is chosen as the next split attribute, the decision tree will be expanded by two nodes, one for $A \leq t$ and another for $A > v$.

Describe an algorithm that returns the best split point t for attribute A.

Let V be the set of all values of A that appear in T. Compute the gini index for the split $A \leq v$ for all $v \in V$. Return the v with the smallest gini index.

**Problem 5 (Active Learning) 15 marks**

Consider a situation where you want to learn a classifier and have a small labeled dataset L (with features and class labels) and a large unlabeled dataset U (only features). You have already trained a classifier on training dataset L, which for a test object returns the probability distribution over the set of all classes. You also have a domain expert who can label unlabeled data, which can then be added to the training dataset and to retrain (and improve the accuracy of) your classifier. However, labeling is expensive, and you want to minimize the number of unlabeled objects labeled.

a) How do you choose the next unlabelled object to be labelled?

   Apply the classifier to U. For every unlabelled object, compute the entropy of the predicted class distribution. Choose the object with the largest entropy to be labelled next. In the case of binary classification (two classes), you can alternatively choose an object whose difference between the probabilities of the two classes is minimal.

b) Explain the reason for your choice.

   The reason for that choice is that this is the object for which the classifier is the most uncertain in its prediction.