COMP9318 (16S2) ASSIGNMENT 2

DUE ON 23:59 1 NOV, 2016 (TUE)

Q1. (*35 marks*)

(1) Consider the following training dataset.

A	B	C	$Class$
0	0	0	+
0	0	1	+
0	1	0	+
0	1	1	−
1	0	0	+
1	0	0	+
1	1	0	−
1	0	1	+
1	1	0	−
1	1	0	−

Illustrate the decision tree constructed by the ID3 algorithm. You need to show your steps.

(2) What is the precision of the constructed decision tree on the training dataset?

(3) What is the precision of the constructed decision tree on the following testing dataset?

A	B	C	$Class$
0	0	0	+
0	1	1	+
1	1	0	+
1	0	1	−
1	0	0	+

(4) Considre the ID3 decision tree induction algorithm. Show that the entropy of the input data never increases after splitting it using any of its attribute.

(5) Consider a Logistic Regression classifier with the parameter

$$\boldsymbol{w}^\top = \begin{bmatrix} 0.2 & 0.3 & -0.1 & 0.4 \end{bmatrix}.$$

Compute the *log likelihood* of the training data under this classifier.

Hint 1. *A function $f(x)$ is convex iff $\forall x_1, x_2, \forall \lambda_1, \lambda_2 \in [0, 1]$ such that $\lambda_1 + \lambda_2 = 1$, $f(\lambda_1 x_1 + \lambda_2 x_2) \leq \lambda_1 f(x_1) + \lambda_2 f(x_2)$. This can be generalized to more than two variables (and is generally known as the Jensen's inequality).*

Q2. (*35 marks*)

Algorithm 1: k-means(D, k)

Data: D is a dataset of n d-dimensional points; k is the number of clusters.

```
1  Initialize k centers C = [c_1, c_2, ..., c_k];
2  canStop ← false;
3  while canStop = false do
4      Initialize k empty clusters G = [g_1, g_2, ..., g_k];
5      for each data point p ∈ D do
6          c_x ← NearestCenter(p, C);
7          g_{c_x}.append(p);
8      for each group g ∈ G do
9          c_i ← ComputeCenter(g);
10 return G;
```

Consider the (slightly incomplete) k-means clustering algorithm as depicted in Algorithm 1.

(1) Assume that the stopping criterion is till the algorithm converges to the final k clusters. Can you insert several lines of pseudo-code to the algorithm to implement this logic? You are **not** allowed to change the first 7 lines though.

(2) The cost of k clusters is just the total cost of each group g_i, or formally

$$cost(g_1, g_2, \ldots, g_k) = \sum_{i=1}^{k} cost(g_i)$$

$cost(g_i)$ is the sum of squared distances of all its constituent points to the center c_i, or

$$cost(g_i) = \sum_{p \in g_i} dist^2(p, c_i)$$

$dist()$ is the Euclidean distance. Now show that the cost of k clusters as evaluated at the end of each iteration (i.e., after Line 9 in the current algorithm) never increases. (You may assume $d = 2$)

(3) Prove that the cost of clusters obtained by k-means algorithm always converges to a local minima. You can make use of the previous conclusion even if you have not proved it.

Hint 2. *In fact, the two loops (Lines 5–7 and Lines 8–9) never increases the cost.*

Q3. (*30 marks*)

Consider binary classification where the class attribute y takes two values: 0 or 1. Let the feature vector for a test instance be a d-dimension column vector $\boldsymbol{x}$. A linear classifier with the model parameter $\boldsymbol{w}$ (which is a d-dimension column vector) is the following function:

$$y = \begin{cases} 1 & \text{, if } \boldsymbol{w}^\top \boldsymbol{x} > 0 \\ 0 & \text{, otherwise.} \end{cases}$$

We make additional simplifying assumptions: $\boldsymbol{x}$ is a binary vector (i.e., each dimension of $\boldsymbol{x}$ take only two values: 0 or 1).

- Prove that if the feature vectors are d-dimension, then a Naïve Bayes classifier is a linear classifier in a $d+1$-dimension space. You need to explicitly write out the vector $\boldsymbol{w}$ that the Naïve Bayes classifier learns.
- It is obvious that the Logistic Regression classifier learned on the same training dataset as the Naïve Bayes is also a linear classifier in the same $d+1$-dimension space. Let the parameter $\boldsymbol{w}$ learned by the two classifiers be $\boldsymbol{w}_{\text{LR}}$ and $\boldsymbol{w}_{\text{NB}}$, respectively. Briefly explain why learning $\boldsymbol{w}_{\text{NB}}$ is much easier than learning $\boldsymbol{w}_{\text{LR}}$.

Hint 3. *It is a common trick in ML (e.g., c.f., Linear Regression) to enhance the feature vector by a dummy dimension x_0 and set $x_0 = 1$.*

Submission

Please write down your answers in a file named `ass2.pdf`. You **must write down your name and student ID on the first page**.

You can submit your file by

```
give cs9318 ass2 ass2.pdf
```

Late Penalty. -10% per day for the first two days, and -30% for each of the following days.