Doris Chen

Directions: This test is closed books/notes. Complete the following problems by hand.

1. (5 pts) We are interested in constructing a decision tree to determine if a borrower will default on their loan.
   (a) (4 pts) Use Hunt's algorithm, the training data given below, and classification error rate to construct a decision tree for classifying borrowers.
   (b) (1 pt) What is the final classification error rate of your decision tree?

	name	status	home owner	default
1.	Jones	married	Y	N
2.	Jackson	married	N	N
3.	Johnson	single	N	Y
4.	James	single	Y	N
5.	Jennings	single	Y	Y

a) ① default = N
(3, 2)

② status
married / \ single
default = N (2, 0)   (1, 2)

③ status
married / \ single
default = N (2, 0)   home owner
Y / \ N
default = N (1, 1)   default = Y (0, 1)

∴ b) error rate $= \frac{1}{5} = 20\%$

2. (5 pts) Consider the training data given below where columns A_1 and A_2 are attributes and column T is the target. Use the Naive Bayes classifier to determine the probability the target is a + given that the attribute values are $A_1 = 0$ and $A_2 = 1$. You must show your calculations.

	A_1	A_2	T
1.	0	1	+
2.	1	1	+
3.	1	0	−
4.	0	1	−
5.	1	0	−

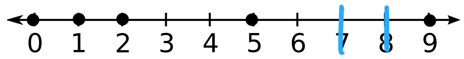
A_1	A_2	T
0	1	?

$$P(+ \mid A_1=0,\ A_2=1) = \frac{P(0,1\mid +)\,P(+)}{P(0,1)} = \frac{\frac{1}{2}\cdot 1\cdot \frac{2}{5}}{\frac{4}{15}} = \frac{1}{\cancel{5}}\cdot\frac{\cancel{15}\,3}{4} \boxed{= \frac{3}{4} = 0.75}$$

$$P(0,1) = P(0,1\mid +) + P(0,1\mid -)$$
$$= P(A_1=0\mid +)\,P(A_2=1\mid +)\,P(+) + P(A_1=0\mid -)\,P(A_2=1\mid -)\,P(-)$$
$$= \frac{1}{\cancel{2}}\cdot 1\cdot\frac{\cancel{2}}{5} + \frac{1}{3}\cdot\frac{1}{\cancel{3}}\cdot\frac{\cancel{3}}{5} = \frac{4}{15}$$

3. K-Means clustering is applied to the five points shown below with the number of clusters set equal to $K = 2$.

0 1 2 5 9



(a) (3 pts) What two clusters does K-means converge to if the K-means algorithm starts with centroids $c_1 = 7$ and $c_2 = 8$. Compute the SSE (sum of squared errors) for the two clusters that K-means algorithm converged to.

Centroid	C_1	C_2
ite 0	7	8
1	2	9

$C_1 = \{0, 1, 2, 5\}$ $\quad C_2 = \{9\}$

$C_1 = \{0, 1, 2, 5\}$ $\quad C_2 = \{9\}$

$$\boxed{SSE = (0-2)^2 + (1-2)^2 + (2-2)^2 + (5-2)^2 + (9-9)^2 = 14}$$

(b) (2 pt) Determine two clusters with a smaller SSE than the two clusters that K-Means converges to in part (a). What is the SSE equal to for these clusters?

$C_1 = \{0, 1, 2\}$ $\quad c_1 = 1$ $\quad SSE = (0-1)^2 + (1-1)^2 + (2-1)^2 = 2$

$C_2 = \{5, 9\}$ $\quad c_2 = 7$ $\quad SSE = (5-7)^2 + (9-7)^2 = 8$

$$\boxed{\text{Total } SSE = 2 + 8 = 10}$$

4. (5 pts) Consider the cluster of points $\{1, 2, 6\}$. Show that setting the centroid, c, equal to the mean of the cluster minimizes SSE. <u>Hint:</u> Set a derivative equal to zero and solve. $c = 3$

$i = 1$

N_i – Number of points in cluster $= 3$

c_i – centroid of i^{th} cluster $= 3$

$\vec{x}_{ij}$ – $\mathbb{R}^d$ (d-demensional space) could be 100

$\vec{x}_{ij}$ is the jth point

$j = 1\ 2 \cdots N_i$

in ith cluster

$$SSE = \sum_{k=1}^{K} \underbrace{\sum_{j=1}^{N_i} dist(x_{ij}, c_i)^2}_{SSE\text{ of } c_i \quad SSE_i} = \sum_{i=1}^{K} SSE_i$$

Proof for 4: (In purple pen)

$$SSE_i = \sum_{j=1}^{3} dist(x_{1j}, c_i)^2$$

$$\frac{dSSE_i}{dc_i} = \frac{d}{dc_i} \sum_{j=1}^{3} (x_{1j} - c_i)^T (x_{1j} - c_i)$$

$$= \sum_{j=1}^{3} 2(x_{1j} - c_i)(-1) = 0$$

$$\cancel{-2} \sum_{j=1}^{3} (x_{1j} - c_i) = 0$$

Thm

The centroid c_i minimizes SSE_i

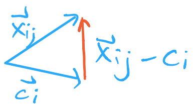
if $c_i = \frac{1}{N_i}\sum_{j=1}^{N_i} x_{ij}$ $\quad i=1,2,\cdots k$

Proof

$\left(\frac{dSSE}{dc_i} = 0 \text{ solve for } c_i\right)$

$$SSE_i = \sum_{j=1}^{N_i} dist(x_{ij}, c_i)^2$$

$$dist(x_{ij}, c_i)^2 = (x_{ij}-c_i)^T(x_{ij}-c_i)$$



$$SSE_i = \sum_{j=1}^{N_i} (x_{ij}-c_i)^T(x_{ij}-c_i)$$

$$\frac{dSSE_i}{dc_i} = \frac{d}{dc_i}\sum_{j=1}^{N_i}(x_{ij}-c_i)^T(x_{ij}-c_i)$$

$$= \sum_{j=1}^{N_i}\frac{d}{dc_i}(x_{ij}-c_i)^T(x_{ij}-c_i) \quad * \frac{d}{dx}(x^T x) = 2x$$

vector derivative

$$= \sum_{j=1}^{N_i} 2(x_{ij}-c_i)(-1) = 0$$

try to find c_i with minimize SSE $\therefore$ set $=0$

$$-2\sum_{j=1}^{N_i}(x_{ij}-c_i) = 0$$

$$\sum_{j=1}^{N_i} x_{ij} - \sum_{j=1}^{N_i} c_i = 0$$

$$\sum_{j=1}^{N_i} x_{ij} = \sum_{j=1}^{N_i} c_i$$

$$= \underbrace{c_i + c_i + \cdots + c_i}_{N_i}$$

$$\sum_{j=1}^{N_i} x_{ij} = N_i c_i$$

$$\frac{1}{N_i}\sum_{j=1}^{N_i} x_{ij} = c_i$$

$$\sum_{j=1}^{3} x_{ij} - \sum_{j=1}^{3} c_i = 0$$

$$\sum_{j=1}^{3} x_{ij} = \sum_{j=1}^{3} c_i$$

$$1+2+6 = 3c_i$$

$$9 = 3c_i$$

$$\boxed{c_i = 3}$$

$$\sum_{j=1}^{3} x_{ij} = N_i c_i = 3c_i$$

$$\frac{1}{N_i}\sum_{j=1}^{N_i} x_{ij} = c_i$$

$$\frac{1}{3}\sum_{j=1}^{3} x_{ij} = 3$$