Problem 1

(20 pt)

Answer the following questions, and provide an **explanation** for each question.

(a)

(5 pt) Can linear regression work when all X values are the same? When all Y values are the same?

(b)

(5 pt) Can linear regression be used when the X values are actually categories?

(c)

(5 pt) Will the regression line be the same if you exchange X and Y?

(d)

(5 pt) Under what circumstances will there be over-fitting in linear regression?

Problem 2
(20 pt)

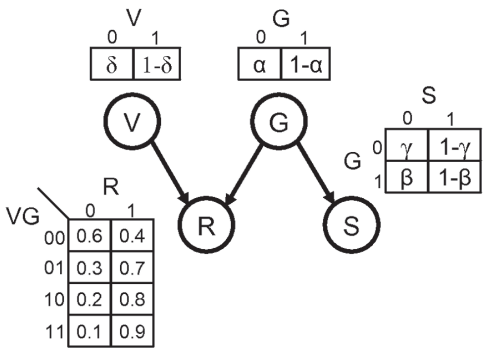


Figure 1: Weather Bayes Nets

In this question you must model a problem with 4 binary variables: $G =$"gray", $V =$"Vancouver", $R =$"rain" and $S =$"sad". Consider the directed graphical model describing the relationship between these variables shown in Figure 1.

(a)

(10 pt) Write down an expression for $P(S = 1|V = 1)$ in terms of $\alpha, \beta, \gamma, \delta$.

(b)

(5 pt) Write down an expression for $P(S = 1|V = 0)$. Is this the same or different to $P(S = 1|V = 1)$? Explain why.

(c)

(5 pt) Find maximum likelihood estimates of α, β, γ using the following data set, where each row is a training case.

V	G	R	S
1	1	1	1
1	1	0	1
1	0	0	0

Problem 3

(10 pt)

Canonical correlation analysis (CCA) handles the situation that each data point (i.e., each object) has two representations (i.e., two sets of features), e.g., a web page can be represented by the text on that page, and can also be represented by other pages linked to that page. Now suppose each data point has two representations $\mathbf{x}$ and $\mathbf{y}$, each of which is a 2-dimensional feature vector (i.e., $\mathbf{x} = [x_1, x_2]^T$ & $\mathbf{y} = [y_1, y_2]^T$). Given a set of data points, CCA finds a pair of projection directions $(\mathbf{u}, \mathbf{v})$ to maximize the sample correlation $\hat{corr}(\mathbf{u}^T\mathbf{x})(\mathbf{v}^T\mathbf{y})$ along the directions $\mathbf{u}$ and $\mathbf{v}$. In other words, after we project one representation of data points onto $\mathbf{u}$ and the other representation of data points onto $\mathbf{v}$, the two projected representations $\mathbf{u}^T\mathbf{x}$ and $\mathbf{v}^T\mathbf{y}$ should be maximally correlated (intuitively, data points with large values in one projected direction should also have large values in the other projected direction).

Now we can see data points shown in the Figure 2, where each data point has two representations $\mathbf{x} = [x_1, x_2]^T$ and $\mathbf{y} = [y_1, y_2]^T$. Note that data are paired: each point in the left figure corresponds to a specific point in the right figure and vice versa, because these two points are two representations of the same object. Different objects are shown in different gray scales in the two figures (so you should be able to approximately figure out how points are paired). In the right figure we've given one CCA projection direction $\mathbf{v}$, draw the other CCA projection direction $\mathbf{u}$ in the left figure.

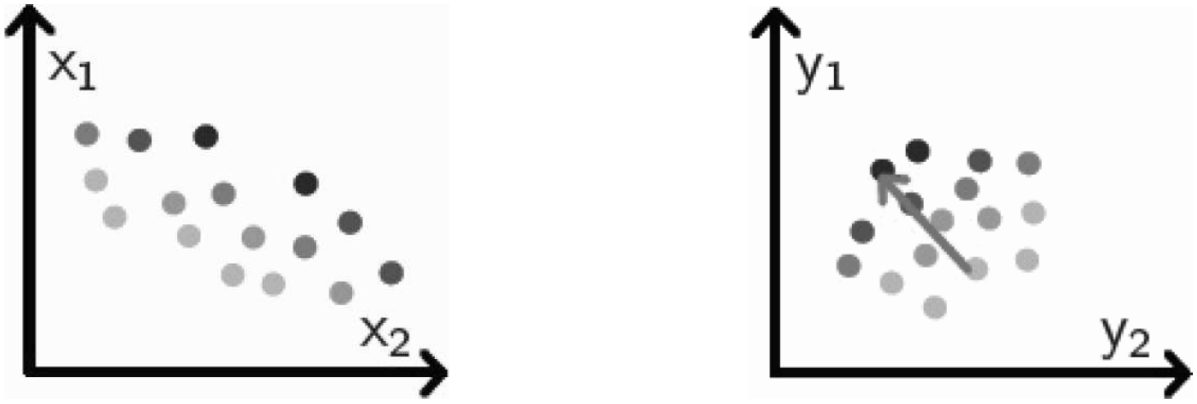
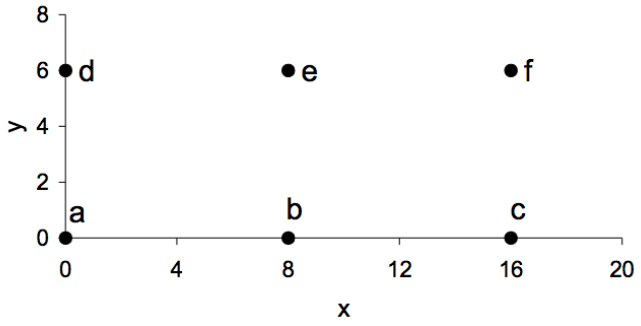


Figure 2: Draw the CCA projection direction in the left figure

Problem 4

(20 pt)

There is a set S consisting of 6 points in the plane shown as below, $a = (0,0), b = (8,0), c = (16,0), d = (0,6), e = (8,6), f = (16,6)$. Now we run the k-means algorithm on those points with $k = 3$. The algorithm uses the Euclidean distance metric (i.e. the straight line distance between two points) to assign each point to its nearest centroid. Ties are broken in favor of the centroid to the left/down. Two definitions:



- A k-starting configuration is a subset of k starting points from S that form the initial centroids, e.g. $\{a, b, c\}$.
- A k-partition is a partition of S into k non-empty subsets, e.g. $\{a, b, e\}, \{c, d\}, \{f\}$ is a 3-partition.

Clearly any k-partition induces a set of k centroids in the natural manner. A k-partition is called *stable* if a repetition of the k-means iteration with the induced centroids leaves it unchanged.

(a)

(10 pt) How many 3-starting configuration are there? (Remember, a 3-starting configuration is just a subset, of size 3, of the six data points).

(b)

(10 pt) Fill in the following table (like an example in the first line):

3-partition	Stable?	An example 3-starting configuration that can arrive at the 3-partition after running k-means (or write none if no such 3-starting configuration exists)	# of unique 3-starting configuration that arrive at the 3-partition
{a, b},{d, e},{c, f}	Y	{b, c, e}	4
{a, b, e},{c, d},{f}			
{a, d},{b, e},{c, f}			
{a},{d},{b, c, e, f}			
{a, b},{d},{c, e, f}			
{a, b, d},{c},{e, f}			

Problem 5

(30 pt)

This problem is about the EM algorithm for mixtures of Bernoullis.
The expectation of the complete-data log likelihood with respect to the posterior distribution is given by

$$Q(\theta, \theta^t) = \sum_{i=1}^{N} \sum_{k=1}^{K} r_{ik} \left[\log \pi_k + \sum_{j=1}^{D} x_{ij} \log \mu_{kj} + (1 - x_{ij}) \log(1 - \mu_{kj}) \right]$$

The pdf of the beta distribution $Beta(\alpha, \beta)$ is given by $f(x; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$.

(a)

(15 pt) Derive the M step for ML estimation of a mixture of Bernoullis.

$$\mu_{kj} = ?$$

(b)

(15 pt) Derive the M step for MAP estimation of a mixture of Bernoullis with a $Beta(\alpha, \beta)$ prior.

$$\mu_{kj} = ?$$