

# CS573: Homework 1 Solution

## 1 Elements of Data Mining (5 pts)

Read the following paper at a high-level (don't worry about the low-level details):

M. Deodhar and J. Ghosh (2006). Consensus Clustering for Detection of Overlapping Clusters in Microarray Data. *Proceedings of the ICDM 2006 Workshop on Data Mining in Bioinformatics (DMB 2006)*. (<http://www.ideal.ece.utexas.edu/papers/deodhar06overlap.pdf>)

Identify the following components of the work:

1. The task

*The task in this paper is to cluster genes into groups, such that each gene can be a member of more than one group.*

2. The data representation

*The data representation is tabular, iid data with sets of individual measurements for each gene, specifically microarray data with gene expression levels under a range of environmental stress conditions.*

3. The knowledge representation

*MCLA: Cluster assignments for each data point, each instance can be associated with more than one cluster.*

*SKK: Probability distribution for each data point, representing the likelihood that the instance was generated from each cluster.*

4. The learning technique (search method + scoring function)

*MCLA: Search is inside the METIS hypergraph partitioning algorithm, scoring function is not described.*

*SKK: Search is expectation maximization (EM) to maximize likelihood with unknown cluster assignments, scoring function is  $\sum_{i=1}^n \sum_{j=1}^k h_{x_i}^{c_j} \log(P(c_j)P(\psi(x_i)|\theta_j))$ .*

5. The inference technique (if applicable) and evaluation method

*The results of the clustering are evaluated using precision, recall, and F-value. The clustering is not applied to new data, the cluster assignments on the sample data are compared to ground truth groupings (i.e., true class labels).*

## 2 Probability (3 pts)

Suppose that we have three colored boxes  $r$  (red),  $b$  (blue), and  $g$  (green). Box  $r$  contains 3 apples, 4 oranges, and 3 limes; box  $b$  contains 1 apple, 1 orange, and 0 limes; box  $g$  contains 3 apples, 3 oranges, and 4 limes. If a box is chosen at random with probabilities  $p(r) = 0.2, p(b) = 0.2, p(g) = 0.6$ , and a piece of fruit is removed from the box (with

equal probability of selecting any of the items in the box), then what is the probability of selecting an apple? If we observe that the selected fruit is in fact an orange, what is the probability that it came from the green box?

**Solution:**

$$\begin{aligned} P(a) &= P(r)P(a|r) + P(g)P(a|g) + P(b)P(a|b) \\ &= 0.2 * 0.3 + 0.2 * 0.5 + 0.6 * 0.3 \\ &= 0.34 \end{aligned}$$

$$\begin{aligned} P(g|o) &= \frac{P(g,o)}{P(o)} = \frac{P(g)P(o|g)}{P(o)} \\ &= \frac{0.6 * 0.3}{0.2 * 0.4 + 0.2 * 0.5 + 0.6 * 0.3} \\ &= 0.5 \end{aligned}$$

### 3 Probability distributions (3 pts)

The form of the Bernoulli( $p$ ) distribution is not symmetric between the two values of  $X$ . In some situations, it will be more convenient to use an equivalent formulation for which  $x \in \{-1, 1\}$ , in which case the distribution can be written as:

$$P(x|p) = \left(\frac{1-p}{2}\right)^{(1-x)/2} \left(\frac{1+p}{2}\right)^{(1+x)/2}$$

Show that this distribution is normalized (i.e., sums to 1) and evaluate its mean, variance, and entropy.

**Solution:**

To show the distribution is normalized, its probability mass function should sum to 1:

$$P(x|p) = P(1|p) + P(-1|p) = \frac{1+p}{2} + \frac{1-p}{2} = 1$$

Mean:

$$\begin{aligned} E(X) &= 1 \cdot P(1|p) + (-1) \cdot P(-1|p) \\ &= \frac{1+p}{2} - \frac{1-p}{2} \\ &= p \end{aligned}$$

Variance:

$$\begin{aligned} Var(X) &= E(X^2) - [E(X)]^2 \\ &= \left[ \sum_x x^2 P(x|p) \right] - p^2 \end{aligned}$$

$$\begin{aligned}
&= 1^2 \cdot \left(\frac{1+p}{2}\right) + (-1)^2 \cdot \left(\frac{1-p}{2}\right) - p^2 \\
&= \frac{1+p}{2} + \frac{1-p}{2} - p^2 \\
&= 1 - p^2
\end{aligned}$$

Entropy:

$$\begin{aligned}
Entropy(X) &= -P(-1|p) \log P(-1|p) - P(1|p) \log P(1|p) \\
&= -\frac{1+p}{2} \log \frac{1+p}{2} - \frac{1-p}{2} \log \frac{1-p}{2}
\end{aligned}$$

#### 4 Independence (3 pts)

Prove that two events  $A$  and  $B$  are independent if and only if the following pairs of events are also independent:

1.  $A$  and  $B'$ .
2.  $A'$  and  $B$ .
3.  $A'$  and  $B'$

where  $A'$  refers to the complement of  $A$ .

**Solution:**

1. If  $A$  and  $B$  are independent:

$$\begin{aligned}
P(A, B) &= P(A)P(B) \\
P(A, B) + P(A, B') &= P(A)P(B) + P(A, B') \\
P(A) &= P(A)P(B) + P(A, B') \\
P(A) - P(A)P(B) &= P(A, B') \\
P(A)[1 - P(B)] &= P(A, B') \\
P(A)P(B') &= P(A, B')
\end{aligned}$$

$\Rightarrow A$  and  $B'$  are independent.

If  $A$  and  $B'$  are independent:

$$\begin{aligned}
P(A, B') &= P(A)P(B') \\
P(A, B') + P(A, B) &= P(A)P(B') + P(A, B) \\
P(A) &= P(A)P(B') + P(A, B) \\
P(A) - P(A)P(B') &= P(A, B) \\
P(A)[1 - P(B')] &= P(A, B) \\
P(A)P(B) &= P(A, B)
\end{aligned}$$

$\Rightarrow A$  and  $B$  are independent.

2. If A and B are independent:

$$\begin{aligned}
 P(A, B) &= P(A)P(B) \\
 P(A, B) + P(A', B) &= P(A)P(B) + P(A', B) \\
 P(B) &= P(A)P(B) + P(A', B) \\
 P(B) - P(A)P(B) &= P(A', B) \\
 P(B)[1 - P(A)] &= P(A', B) \\
 P(B)P(A') &= P(A', B)
 \end{aligned}$$

$\Rightarrow A'$  and  $B$  are independent.

If  $A'$  and  $B$  are independent:

$$\begin{aligned}
 P(A', B) &= P(A')P(B) \\
 P(A', B) + P(A, B) &= P(A')P(B) + P(A, B) \\
 P(B) &= P(A')P(B) + P(A, B) \\
 P(B) - P(A')P(B) &= P(A, B) \\
 P(B)[1 - P(A')] &= P(A, B) \\
 P(B)P(A) &= P(A, B)
 \end{aligned}$$

$\Rightarrow A$  and  $B$  are independent.

3. If A and B are independent:

$$\begin{aligned}
 P(A, B) &= P(A)P(B) \\
 P(A, B) - P(A) - P(B) &= P(A)P(B) - P(A) - P(B) \\
 -P(A \cup B) &= P(A)P(B) - P(A) - P(B) \\
 1 - P(A \cup B) &= 1 + P(A)P(B) - P(A) - P(B) \\
 P(A', B') &= 1(1 - P(A)) - P(B)(1 - P(A)) \\
 P(A', B') &= (1 - P(B))(1 - P(A)) \\
 P(A', B') &= P(B')P(A')
 \end{aligned}$$

$\Rightarrow A'$  and  $B'$  are independent.

If  $A'$  and  $B'$  are independent:

$$\begin{aligned}
 P(A', B') &= P(A')P(B') \\
 P(A', B') - P(A') - P(B') &= P(A')P(B') - P(A') - P(B') \\
 -P(A' \cup B') &= P(A')P(B') - P(A') - P(B') \\
 1 - P(A' \cup B') &= 1 + P(A')P(B') - P(A') - P(B') \\
 P(A, B) &= 1(1 - P(A')) - P(B')(1 - P(A')) \\
 P(A, B) &= (1 - P(B'))(1 - P(A')) \\
 P(A, B) &= P(B)P(A)
 \end{aligned}$$

$\Rightarrow A$  and  $B$  are independent.

## 5 Expectation (4 pts)

Consider two variables  $X$  and  $Y$  with joint distribution  $P(X, Y)$ . Prove the following two results:

1.  $E[X] = E_Y[E_X[X|Y]]$
2.  $Var[X] = E_Y[Var_X[X|Y]] + Var_Y[E_X[X|Y]]$

Here  $E_X[X|Y]$  denotes the expectation of  $X$  under the conditional distribution  $P(X|Y)$ , with a similar notation for the conditional variance.

**Solution:**

1.

$$\begin{aligned} E_Y E_X(X|Y) &= E_Y \sum_x x P(x|Y) \\ &= \sum_y \sum_x x P(x|y) P(y) \\ &= \sum_{x,y} x P(x, y) \\ &= E(X) \end{aligned}$$

2.

$$\begin{aligned} &E_Y(Var_X(X|Y)) + Var_Y(E_X(X|Y)) \\ &= E_Y(E_X(X^2|Y) - E_X^2(X|Y)) + E_Y E_X^2(X|Y) - E_Y^2 E_X(X|Y) \\ &= E_Y E_X(X^2|Y) - E_Y E_X^2(X|Y) + E_Y E_X^2(X|Y) - E_Y^2 E_X(X|Y) \\ &= E(X^2) - E^2(X) \\ &= Var(X) \end{aligned}$$

## 6 Covariance (3 pts)

Prove the following. If  $X$  and  $Y$  are random variables and  $a$  and  $b$  are constants, then:

1.  $Cov(aX, bY) = abCov(X, Y)$ .
2.  $Cov(X + a, Y + B) = Cov(X, Y)$ .
3.  $Cov(X, aX + b) = aVar(X)$ .

**Solution:**

1.

$$\begin{aligned} Cov(aX, bY) &= E((aX - E(aX))(bY - E(bY))) \\ &= abE((X - EX)(Y - EY)) \\ &= abCov(X, Y) \end{aligned}$$

2.

$$\begin{aligned} \text{Cov}(X + a, Y + b) &= E((X + a - E(X + a))(Y + b - E(Y + b))) \\ &= E((X - EX)(Y - EY)) \\ &= \text{Cov}(X, Y) \end{aligned}$$

3.

$$\begin{aligned} \text{Cov}(X, aX + b) &= E((X - EX)(aX + b - E(aX + b))) \\ &= E((X - EX)(aX - aEX)) \\ &= aE((X - EX)^2) \\ &= a\text{Var}(X) \end{aligned}$$

## 7 Maximum likelihood estimation (3 pts)

Let  $X$  have an exponential density:

$$P(x|\theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

1. Plot  $P(x|\theta)$  versus  $x$  for  $\theta = 1$ . Plot  $P(x|\theta)$  versus  $\theta$ , ( $0 \leq \theta \leq 5$ ), for  $x = 2$ .
2. Suppose that  $n$  samples  $x_1, \dots, x_n$  are drawn independently according to  $P(x|\theta)$ . Show that the maximum likelihood estimate for  $\theta$  is given by:

$$\hat{\theta} = \frac{1}{\frac{1}{n} \sum_{k=1}^n x_k}$$

3. On your graph generate with  $\theta = 1$  in part (1), mark the maximum likelihood estimate  $\hat{\theta}$  for large  $n$ .

**Solution:**

1. Please refer to *Figure 1* and 2.
- 2.

$$\begin{aligned} L &= \prod_k \theta e^{-\theta x_k} = \theta^n e^{-\theta \sum x_k} \\ \frac{dL}{d\theta} &= n\theta^{n-1} e^{-\theta \sum x_k} + \theta^n e^{-\theta \sum x_k} (-\sum x_k) = 0 \\ 0 &= n - \theta \sum x_k \\ \hat{\theta} &= \frac{n}{\sum_k x_k} \end{aligned}$$

3. For large  $n$ ,  $\hat{\theta}$  will be close to 1.

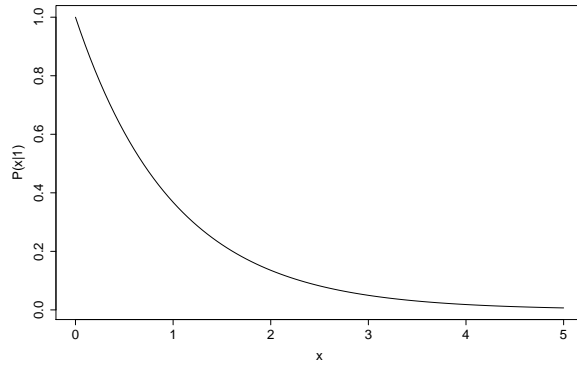


Figure 1: 7.1  $P(x|1)$

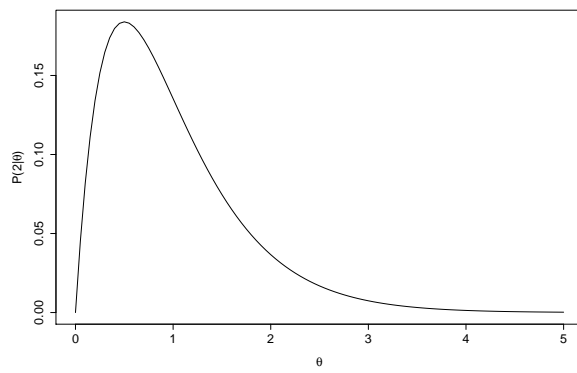


Figure 2: 7.1  $P(2|\theta)$

## 8 Sampling (2 pts)

You are given a set of  $m$  objects that is divided into  $K$  groups, where the  $i^{\text{th}}$  group is of size  $m_i$ . If the goal is to obtain a sample of size  $n < m$ , what is the difference between the following two sampling schemes (assume sampling with replacement):

1. We randomly select  $n \cdot \frac{m_i}{m}$  elements (rounded to the nearest integer) from each group.
2. We randomly select  $n$  elements from the data set, without regard for the group to which an object belongs.

**Solution:** The first method is called stratified sampling which selects items proportional to the size of the group, while the second method cannot guarantee that. Stratified sampling has lower variance when group size is skewed or sample size is small. But since it will always maintain the group proportions, it will not capture natural variance in group sizes (if this is expected). Also, it can't be used if the amount of data in the subgroups is not equal but each subgroup is of equal importance—in this case stratified sampling would give more importance to larger subgroups.

## 9 Noise and Outliers (2 pts)

Distinguish between noise and outliers. Be sure to consider the following questions:

1. Is noise ever interesting or desirable?  
*Generally no. Most of the time noise is not interesting, nor desirable, as it adds bias (if the effect is systematic) and/or variance (if the effect is random) to the measurements.*
2. Can noise objects be outliers?  
*Yes. Noise objects can be outliers if their deviation from "normal" is large enough. However, with the exception of systematic bias that only affects a few examples, most noise effects will either appear to be normal behavior (since they affect all points) or their impact will be relatively small.*
3. Are noise objects always outliers?  
*No. See explanation above.*
4. Are outliers always noise objects?  
*No. See explanation above.*
5. Can noise make a typical value into an unusual one, or vice versa?  
*Yes. Sometimes noise can even make a typical value into an unusual one or vice versa depending on the magnitude of the effect and its relation to the "true" observed value.*

## 10 Distance and Correlation Measures (4 pts)

1. Derive the mathematical relationship between cosine similarity and Euclidean distance when each data object has an  $L_2$  length of 1.
2. Derive the mathematical relationship between correlation and Euclidean distance when each data point has been standardized by subtracting its mean and dividing by its standard deviation.

**Solution:**

1. Let  $\mathbf{x} = (x_1, x_2)$  and  $\mathbf{y} = (y_1, y_2)$  be two arbitrary data points, with unit length, then their cosine similarity is:

$$\cos \theta = \frac{x_1 y_1 + x_2 y_2}{\sqrt{x_1^2 + x_2^2} \sqrt{y_1^2 + y_2^2}} = x_1 y_1 + x_2 y_2$$

Their Euclidean distance is:

$$\begin{aligned} d(\mathbf{x}, \mathbf{y}) &= \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \\ &= \sqrt{(x_1^2 + x_2^2) + (y_1^2 + y_2^2) - 2(x_1 y_1 + x_2 y_2)} \\ &= \sqrt{2 - 2 \cos \theta} \end{aligned}$$

Thus we obtain the relationship the following relationship between Euclidean distance and cosine similarity:  $d(\mathbf{x}, \mathbf{y}) = \sqrt{2 - 2 \cos \theta}$ .

2. For normalized data points,  $\mu_x = \mu_y = 0$  and  $\sigma_x = \sigma_y = 1$ . The correlation of the points is:

$$\begin{aligned} \rho &= \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} \\ &= \frac{E[(X - \mu_x)(Y - \mu_y)]}{\sigma_x \sigma_y} \\ &= E[XY] \\ &= \frac{1}{n} \sum_{i=1}^n x_i y_i \end{aligned}$$

The Euclidean distance can be expressed as:

$$\begin{aligned} d(\mathbf{x}, \mathbf{y}) &= \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \\ &= \sqrt{\sum_{i=1}^n x_i^2 + \sum_{i=1}^n y_i^2 - 2 \sum_{i=1}^n x_i y_i} \\ &= \sqrt{2n - 2n\rho} \end{aligned}$$

Thus we obtain the relationship the following relationship between Euclidean distance and correlation:  $d(\mathbf{x}, \mathbf{y}) = \sqrt{2n - 2n\rho}$

## 11 Proximity (3 pts)

Proximity is typically defined between a pair of instances.

1. Outline two ways in which you might define the proximity among a group of instances.  
*The proximity can be defined as:*

$$P_1 = \frac{1}{\sum \|x_i - \bar{x}\|_p}, (p > 0)$$

or

$$P_2 = \frac{1}{\sum \|x_i - x_j\|_p}, (p > 0)$$

2. How might you define the distance between two sets of points in Euclidean space?  
*Given two sets of points  $X = x_1, x_2, \dots, x_m$ , and  $Y = y_1, y_2, \dots, y_n$ , where  $\bar{X} = \frac{1}{m} \sum x_i$  and  $\bar{Y} = \frac{1}{n} \sum y_i$ . The distance can be defined as:*

$$D(X, Y) = \|\bar{X} - \bar{Y}\|_p, (p > 0)$$

3. How might you define the proximity between two sets of data instances? (Make no assumption about the instances, except that a proximity measure is defined between any pair of instances.)  
*Let  $prox(x_i, y_j)$  be a proximity measurement defined on any pair  $x_i, y_j$ , then between two sets of points  $X_m$  and  $Y_n$ , we can define the set proximity as:*

$$prox_{set}(X, Y) = \frac{1}{mn} \sum prox(x_i, y_j)$$