

CS573: Homework 1

Due date: Thursday September 9, start of class

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
2. The data representation
3. The knowledge representation
4. The learning technique (search method + scoring function)
5. The inference technique (if applicable) and evaluation method

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?

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.

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 .

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.

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. $Cov(X, aX + b) = aVar(X)$.

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 θ , ($0 \leq \theta \leq 5$), 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 θ 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 .

8 Sampling (2 pts)

You are given a set of m objects that is divided into K groups, where the i^{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.

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?
2. Can noise objects be outliers?
3. Are noise objects always outliers?
4. Are outliers always noise objects?
5. Can noise make a typical value into an unusual one, or vice versa?

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.

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.
2. How might you define the distance between two sets of points in Euclidean space?
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.)