

CS725: Foundations of Machine Learning, Fall 2009, Homework 5

Roll: _____

Due date: 13 November 2009

Name: _____

Mode: Credit/Audit _____

Write all your answers in the space provided. Do not spend time/space giving irrelevant details or details not asked for. You are expected to solve most of the questions on your own. If you discuss any solution with someone mention alongside that question the person with whom you discussed it.

1. Show that $1 - x \leq e^{-x}$.

..1

2. Show that the VC-dimension of hyperplanes in d -dimensions is at least $d + 1$.

..2

3. Let $\theta_1, \dots, \theta_K$ follow a Dirichlet distribution with parameters $\alpha_1, \dots, \alpha_K$ where $\alpha = \sum_{k=1}^K \alpha_k$. Prove the following.

(a) The variance of θ_i is $\frac{\alpha_i(\alpha - \alpha_i)}{\alpha^2(\alpha + 1)}$

..2

(b) The covariance between θ_i and θ_j is $\frac{-\alpha_i\alpha_j}{\alpha^2(\alpha + 1)}$

..2

(c) For $\alpha_i > 1$, show that the mode of the distribution is obtained at each θ_i set to $\frac{\alpha_i-1}{\alpha-K}$.

..2

4. Consider running the Structure Preserving Embedding algorithm with the goal of preserving the clusters returned by the single link clustering algorithm (defined in Question 2 of HW 4). That is, given

- (a) a set of C clusters over N points
- (b) algorithm A for single link clustering on K

Your goal is to find a similarity matrix \mathcal{K} such that running A on it will recover the same set of clusters.

What linear constraints will you add for this goal?

..5

5. Consider a version of Spatial scan statistics where we have a set of background points B in two dimensions and each point (x_i, y_i) is associated with a real value v_i instead of a color. Note values can be positive or negative. How will you find the rectangle R of highest discrepancy for the following discrepancy measures?

- (a) Sum of values associated with points in B_R .

..3

(b) Sum of squares of values associated with points in B_R .

..2

(c) Average of values associated with points in B_R .

..2

(d) Variance of values associated with points in B_R . (Extra credit question)

Total: 21