CSP 571 – Data Preparation & Analysis

Spring 2023 – All Sections

Midterm Exam - Sample Questions

Part I – Short Answer (Show Points/Results) – 5 points each, 30 points total

1. Given the following observations: $x_1 = (3,4); x_2 = (5,12); x_3 = (8,15)$ in $\mathbb{R}^2$, what would the Euclidean (ℓ_2) norm of each X_j feature vector be in $\mathbb{R}^3$? What would the Manhattan (ℓ_1) norm be? What are the distances between the points under each norm?

2. A regression result contains a coefficient β_3 with an estimated confidence interval having the range [-1.23, 1.21]. Would β_3 have a t-statistic that is significant for rejecting the null hypothesis that $\beta_3 = 0$? Why or why not?

3. Given a dataset with $n = 10000$ observations, what is the size of a training set and validation set be for k-fold cross-validation with $k = 12$? If we wished to decrease the correlation of models fitted under each fold, should we increase or decrease k?

Part II – Long Answer (Show Reasoning/Calculations) – 10 points each, 20 points total

1. Given a regression result with residual sum of squares (RSS) of 256, with n=18 observations with a single (p=1) predictors, provide the residual standard error (RSE) for the model. What property of the error term ϵ does this value provide an estimate of?

2. A classification analysis on a dataset D, where each record is a member of one of K=5 classes, involves the use of LDA for modeling. Please outline the assumptions that LDA makes regarding the probability distribution of the features of D. Specifically, what is the form of the likelihood function $f_k(x)$ for a given class k?