MATH 571/CSP 571 – Data Preparation & Analysis

Fall 2021 – All Sections

Midterm Exam

Part I – Short Answer (Show Points/Results) – 5 points each, 40 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 $\mathbb{R}^3$? Assuming the following labeling: $y_1 = 0; y_2 = 0; y_3 = 1$ for the data, what kNN ($K = 2$) nearest-neighbor label would be given to an out-of-sample point $x_0 = (0,0)$?

2. With a regression data set containing $d = 15$ features and $N = 12{,}000$ observations, what is the dimensionality of the covariance matrix of the predictors? If we were to represent the predictors with a multivariate normal distribution, how many distribution parameters would need to be estimated from the feature data?

3. Given a dataset with n observations, what is the size of a training set and validation set be for k-fold cross-validation? If we wished to perform leave-one-out cross-validation, what would be set the value of k to be?

4. Within a binary classification dataset with a single predictor, x, the two classes have means of 3.0 and 5.0 respectively. According to Bayes Rule, what property of the two classes has to be equal in order for the Bayes Decision Boundary to fall on the point $x = 4.0$ when performing classification via LDA?

5. Given a dataset of $n = 4$ of observations of a single dimensional variable: $x = [4, -13, -2,8]$, would the following be a valid bootstrap sample: $x^* = [4,4, -2, -2]$? What type of resampling does bootstrap use which allows/does not allow this?

6. What is the difference between the logistic function and the odds function? What range of values does each function take on? As an example, if an event has a $20\,\%$ chance of occurring, what are the implied odds?

7. When processing observations that have missing values for certain features, one has the option of replacing missing values with the sample mean, $\hat{\mu}$, of the feature. Will this method affect the sample variance, $\hat{\sigma}$, of the feature in any way? What value would the sample variance approach asymptotically?

8. When performing ridge regression, the tuning parameter λ can vary between $[0,\infty)$. What type of estimation does the model reduce to when $\lambda = 0$? As λ increases, what happens to the β_j coefficients of the linear model? When $\lambda = \infty$, what model do we obtain?

Part II – Long Answer (Show Reasoning/Calculations) – 10 points each, 40 points total

1. Show that with the number of observations n in a multiple regression being held constant, an increase in the number of dimensions/predictors p results a lower F-statistic for the null hypothesis that all regression coefficients are 0. What does this imply about regression results for high-dimensional data with small sample sizes?

2. Briefly explain why some non-parametric regression methods such as k-nearest-neighbor (kNN) perform poorly when feature dimensionality p increases. Specifically, provide the equation for $\hat{f}(x_0)$ and explain how the region of K neighbors in $\mathcal{N}_0$ behaves in higher dimensions.

3. The results of a regression have produced a $\hat{\beta}_1$ coefficient value of 3.92, with a standard error $SE(\hat{\beta}_1)$ of 1.25. Given these values, what is the $95\,\%$ confidence interval of $\hat{\beta}_1$ for the population value? What would the t-statistic of this coefficient be? What does the t-statistic show in terms of the number of standard deviations $\hat{\beta}_1$ is from 0? At what confidence interval would we reject the null hypothesis for this coefficient?

4. A real-valued response variable y has $n = 101$ observations with an *unbiased* sample variance of 3.75 - a regression analysis has produced a residual sum of squares (RSS) of 125. Based on these values, what would the R^2 be of this regression? If we assumed this regression was done with a single predictor x, what value would we obtain for the correlation ρ_{xy} between the predictor and response?

Part III – Essay Question (Show Argument/Proof) – 20 points each, 20 points total

1. Provide a sketch of of the three components of mean-squared-error (MSE) for a typical modeling problem: Squared Bias, Variance, and Irreducible Error - plot each of the curves as a function of model flexibility. Why are each of the curves shaped the way that they are? How does the shape of these curves explain the bias-variance trade-off?

Lucky 7 – Bonus Questions (Industry News, AI/ML Topics) – 1 point each, 7 points total

1. What top Computer Science award was recently announced for Hinton/Bengio/LeCun, some of the core founders of modern Machine/Deep Learning?

2. What did researchers at MIT use for input from the Internet in order to have an AI system generate new food combinations?

3. How much did an AI-generated portrait sell for recently at a Christie's (New York) art auction?

4. Which group of AI researchers recently claimed to their deep-learning based methods predicted weather more accurately than conventional forecasting?

5. Which university recently released a collaborative AI tool which allowed for accelerated materials science research?

6. What recent language model has been singled out for its ability to generate news stories which contribute to online disinformation?

7. What country recently started a program to pair police officers with robots designed for identifying and processing petty and nuisance crimes?