MATH 571/CSP 571 – Data Preparation & Analysis

Fall 2020 – All Sections

Final Exam

Part I – Short Answer (Show Points/Results) – 5 points each, 40 points total

1. 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?

2. 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?

3. 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?

4. The lasso regression method also performs shrinkage as the ridge regression method does - what does lasso perform in addition to shrinkage? In terms of the β_j coefficients, what values do these take on under the lasso which results in a sparse model?

5. When applying bagging to a decision (regression) tree estimation, we perform a bootstrap of size B, obtaining trained trees $\hat{f}^{*b}(x)$ for $b = \{1..B\}$. Given an out-of-sample observation x_{os}, what would the prediction $\hat{f}_{bag}(x_{os})$ be? If we were performing classification instead of regression, how would this solution change?

6. Assuming we have a dataset with p predictors, of which one is considered a strong predictor. We use random forests to reduce model variance over regular bagging, setting $m = \sqrt{p}$ for the estimation. On average, what number of splits in a give tree in the random forest will not consider the strong predictor? How does this process reduce variance?

7. Given a data matrix D with $d = 5$ columns with a total variance of 100, an analyst performs a PCA via eigenvalue decomposition, with the resulting eigenvalues as $[35,25,20,15,5]$. If the analyst wishes to reduce dimensionality with $80\,\%$ of variance explained, how many dimensions would the analyst be able to reduce down to? What would be the standard deviations of the data for these selected dimensions?

8. During a k-means clustering process, we obtain $K = 2$ clusters with the following single-dimensional observations assigned to each: $C_1 = \{1,2,3,4\}$ and $C_2 = \{-9, -8, -7, -6\}$. What would the centroids of these two clusters be? As the k-means algorithm searches through a large parameter space to obtain a local optimum, what is the total number of clusterings we would have to check to find a global optimum given the above data?

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

1. When using smoothing splines for regression, the tuning parameter λ determines the smoothness of the fitting function via application of a penalty term to a loss function. Determine whether the effective degrees of freedom df_λ increase or decrease as λ varies between $[0,\infty)$. What minimum and maximum values does df_λ take on?

2. When using *soft-margin* estimation for a SVM/SVC, we allow for a certain number of training observations out of n samples to violate the margin. Given a budget $C = m$ as our selected tuning parameter value, can more than m training observations be located on the other side of the margin? On the other side of the hyperplane? Show your answer in terms of constraints on *slack* variables ϵ_i.

3. A node within a decision (classification) tree contains 100 observations split evenly between two class types: C_1 and C_2. What is the Gini coefficient value at this node? What is the Entropy value at this node? Assuming an optimal split is performed into two leaf nodes, what are the Gini and Entropy values at each? (**Hint:** Assume $0 \log_2 0 = 0$).

4. Within a hierarchical clustering process, a cluster contains two observations: point p_1 with coordinates $(4,5)$ and point p_2 with coordinates $(6,13)$. The algorithm needs to merge in point p_3 with coordinates $(1,1)$. What would be the dissimilarity (distance) values be for single linkage? Complete linkage? If the final merged cluster with all three points was used for k-means initialization, what would be the coordinates of the centroid that is used?

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

1. Discuss the difference between a *maximal margin classifier* and a *support vector classifier*. What role does the *separating hyperplane* play for each, and what change is introduced to the optimization problem used in estimation? Which classifier has lower bias/higher variance, and how is the bias-variance trade-off controlled in each case?

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

1. What deep learning pioneer recently commented that AI has become an arms race between governments and a few large firms?

2. What animal colonies are an Israeli startup using AI to monitor in order to address parasites and population declines?

3. What new type of neural networks which extend CNN models were recently released by Hinton?

4. When running IQ tests for all major AI assistants (Google, Apple, Amazon, Microsoft, etc.), what human age were all models shown to be under?

5. What firm was shown to be calculating lower credit limits/scores for women as part of their new credit/payment offering, highlight algorithmic bias against women?

6. What model from DeepMind at Google was able to recover missing characters from ancient texts in order to aid historical reconstruction and translation?

7. What recent furniture item from Ikea was shown to be assembled by a robot using deep learning models?