MATH 571/CSP 571 – Data Preparation & Analysis

Fall 2021 – All Sections

Final Exam

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

1. We fit a piecewise polynomial for a regression, selecting 3 *cut-points* in our data - how many constraints need to be introduced at each *cut-point* to obtain a *cubic spline*? Explain each constraint. What would be the total *degrees-of-freedom* for the curve?

2. With a data sample containing $d = 250$ features and $N = 5{,}000{,}000$ observations, what are the number of *eigenvectors* and *eigenvalues* that we would obtain via decomposition of the sample covariance matrix? If we select only the top 10% of the eigenvectors for dimensionality reduction, what would be the dimensions of the *projected* data sample matrix?

3. Given the first layer of a *convolutional* neural network with 100 units, and which inputs in grayscale images that have $30\ X\ 30$ pixel dimensions, what is the size of the weight matrix W_1 for this CNN? Describe what the rows/columns represent.

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 $p = 225$ predictors, of which one is considered a strong predictor. We use random forests to reduce model variance over regular bagging, setting $m = \sqrt{p} = 15$ for the estimation. On average, what number of splits for a given 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 $n = 100$ 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. How many new exoplanets was NASA able to recently discover via the existing data from the Kepler spacecraft?

2. What area of mathematics did professors from Oxford leverage DeepMind for in order to develop novel new conjectures?

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 recently introduced the use of AI for analysis of extreme weather data which impact business supply chains?

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?