MTH/CSE 4224 Midterm Exam

Deadline: Saturday, March 16 by 11:59 PM
Max score: 110/100

To receive full credit on computational problems, **show all mathematical work**.

You may use notes, books, or other sources, but you **may not communicate with other people (except me) about the exam**. If you use content from any source other than your brain, it **must be cited**.

You must submit exactly 1 document with **handwritten** solutions in Canvas.

Problems

1. What is the difference between supervised and unsupervised learning? Give an example of each in the context of economics. [4]

2. What is the difference between a train/test split and K-fold cross-validation? When is each ideal? [4]

3. What is bootstrapping? Why is it used? [4]

4. What information does a covariance matrix contain? How do you compute a covariance matrix for a dataset D containing n points in $\mathbb{R}^d$? [4]

5. What supervised learning methods we have covered require covariance matrices? [4]

6. Why is a radial basis function expansion model so much more challenging to fit to a dataset than linear regression or an LBF expansion? [6]

7. What are precision and recall? If a model has high precision and low recall, what does this imply about your model's errors practically? [4]

8. Derive a formula for the optimal parameters for a linear basis function expansion model using functions $h_1, ..., h_m$ with sum of squared errors loss. [20]

9. In Fisher's LDA, what is the difference between μ_i and m_i? Derive the formulas for each in terms of the original points $x_1, ..., x_n \in \mathbb{R}^d$. Then, derive a formula for the optimal projection vector w for binary classification. [20]

10. Give end-to-end descriptions of how random forests and XGBoost are trained and how they perform inference. What are the strengths and weaknesses of each? [20]

11. Derive formulas for the discriminants for each class in a QDA classifier. Is it more or less complex than LDA? [10]

12. Compare and contrast nearest neighbor and decision tree classifiers. Include computational costs, interpretability, and other important factors. [10]