ORIE 4741 Final Exam Review Section

December 4, 2019

Problem 1

Generalized low rank models. Which of the following statements are true?

A. Singular value decomposition (SVD) can be used to solve the PCA problem.

B. Alternating minimization can be used to solve the PCA problem.

C. PCA is an unsupervised learning technique.

D. Alternating minimization can be used to fit any generalized low rank model

E. Alternating minimization always finds the global minimum of any generalized low rank model.

Problem 2

Computational methods.

2.1 Least Squares

List three computational methods that can be used to solve a least squares problem. Give an example of when you would choose one method over the other two, and say why.

2.2 PCA

List two computational methods to fit PCA to the matrix $Y \in \mathbb{R}^{n \times d}$, minimize $\|Y - XW\|_F^2$, for $X \in \mathbb{R}^{n \times r}$, $W \in \mathbb{R}^{r \times d}$. Give an example of when you would choose one method over the other.

Problem 3

Designing the train/test split. Recall Hoeffding's inequality: Let $z_i \in \{0, 1\}$, $i = 1, \ldots, n$, be independent Boolean random variables with mean $z_i = \mu$. Define the sample mean $\nu = \frac{1}{n} \sum_{i=1}^{n} z_i$. Then for any $\epsilon > 0$,

$$\mathbb{P}[|\nu - \mu| > \epsilon] \leq 2 \exp(-2\epsilon^2 n).$$

Suppose we want to estimate the error rate μ of our binary classification model $g : \mathbb{R}^d \to \{0, 1\}$. Our procedure will be to leave out n data points in the test set when we fit our model, and to evaluate the model on each example (x_i, y_i) in the test set. We will compute the fraction of the time the model correctly predicts the output, $g(x_i) = y_i$, and use that as our error estimate:

$$\nu = \frac{1}{n} 1_{\{g(x_i) \neq y_i\}}.$$

If we want to be sure that, with 95% probability, the sample error rate ν is within .1 of true error rate μ, then how many test data points n do we need?
(Giving a formula of the kind you might type into a calculator, rather than a numerical answer, is fine; but be sure no variables remain in your answer.)