# ORIE 4741 Final Exam

December 15, 2016

**Rules for the exam.**

- Write your name and NetID at the top of the exam.
- The exam is 2.5 hours long.
- Every multiple choice or true false question is worth 1 point. Every free answer question is worth 3 points.
- You may use one ($8.5 \times 11$) page of notes (back and front).
- No computers, calculators, phones, or other electronic devices are allowed.
- Only answers written in pen will be considered for regrading.
- For multiple choice questions, no explanations of your answers are needed. Note that more than one choice may be true.
- The questions in the exam are roughly ordered by how long we think they'll take you. Questions later in the exam tend to be less factual and more open-ended (and interesting).

**Honor code.** Write the following statement, and sign your name below:

> I certify that all work on this exam is my own.

1. *Loss functions for classification.* Which of the following statements are true?

   (a) 0-1 Loss is sensitive to outliers that are correctly classified.

   (b) 0-1 Loss is continuous.

   (c) 0-1 Loss is differentiable.

   (d) Quadratic loss is sensitive to outliers that are correctly classified.

   (e) Quadratic loss is continuous.

   (f) Quadratic loss is differentiable.

   (g) Hinge loss is sensitive to outliers that are correctly classified.

   (h) Hinge loss is continuous.

   (i) Hinge loss is differentiable.

2. *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.

3. *k-means.* Recall the $k$-means problem

   - given data points $y_i \in \mathbf{R}^d$, $i = 1, \ldots, n$
   - find $k$ centers $w_l \in \mathbf{R}^d$, $l = 1, \ldots, k$
   - and assignments $c_i \in \{1, \ldots, k\}$, $i = 1, \ldots, n$
   - to minimize

   $$\sum_{i=1}^{n} \|y_i - w_{c_i}\|^2$$

   The $k$-means algorithm alternately minimizes this objective over the assignments $c_i \in \{1, \ldots, k\}$, $i = 1, \ldots, n$, and over the centers $w_l \in \mathbf{R}^d$, $l = 1, \ldots, k$. Which of the following statements is true?

   (a) The $k$-means algorithm will always find a global optimum of the problem.

   (b) If the $k$-means algorithm is initialized with a specific set of centers, it will always return the same assignments.

   (c) It's a good idea to run the $k$-means algorithm multiple times with different initial starting centers to see how stable the solution is, and to find the best solution.

4. Suppose you split your data set into a training set, a validation set, and a test set. You fit a regularized least squares problem with 100 different values of $\lambda$ to your training data set. Then you evaluate each model on the validation set to compute its validation error, and pick the model with the smallest error: call that model $g$. You evaluate the error of $g$ on your test set. Which of the following statements are true?

   (a) If the validation error of $g$ is much higher than the training error, then adding more features into our data set may resolve it.

   (b) If the validation error of $g$ is much higher than the training error, then increasing the size of the training set may resolve it.

   (c) If the validation error of $g$ is much higher than the training error, then increasing the size of the validation set may resolve it.

   (d) If the test error of $g$ is much higher than the validation error, then considering more values of $\lambda$ may resolve it.

   (e) If the test error of $g$ is much higher than the validation error, then increasing the size of the validation set may resolve it.

5. *Iterative methods.* The first two statements below refer to the stochastic proximal (sub)gradient method. Which of the following four statements is true?

   (a) When using a fixed step size $\alpha_t = \alpha$, iterates quickly converge and then wander within a small ball.

   (b) When using a decreasing step size $\alpha_t = 1/t$, iterates converge slowly to a solution.

   (c) Gradient descent cannot be used to solve all problems because some regularizers and loss functions aren't differentiable.

   (d) The Lasso problem cannot be solved using the proximal gradient method.

6. *Supervised and unsupervised.* Explain the difference between supervised learning and unsupervised learning.

7. *Standardizing.* Explain what it means to standardize data. When is it a good idea? When is it a bad idea?

8. *Overfitting and underfitting.* Explain the difference between overfit and underfit models. Why are they each problematic? How can we reduce overfitting? How can we reduce underfitting? Is it possible to underfit and overfit at the same time?

9. *Proximal gradient.*

   (a) In what situation would you use the proximal gradient method rather than gradient descent?

(b) Describe in words the output of the proximal operator

$$\mathbf{prox}_r(z) = \operatorname*{argmin}_w (r(w) + \frac{1}{2}\|w - z\|_2^2),$$

where $r : \mathbf{R}^d \to \mathbf{R}$. Is the output a number or a vector?

(c) Solve for the proximal operator when $r(w) = \mathbf{1}(w \geq 0)$ and $w \in \mathbf{R}$.

(d) Write out pseudocode for proximal gradient method for solving the problem

$$\text{minimize } \ell(w) + r(w).$$

10. *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 $\mathbb{E}z_i = \mu$. Define the sample mean $\nu = \frac{1}{n}\sum_{i=1}^n z_i$. Then for any $\epsilon > 0$,

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

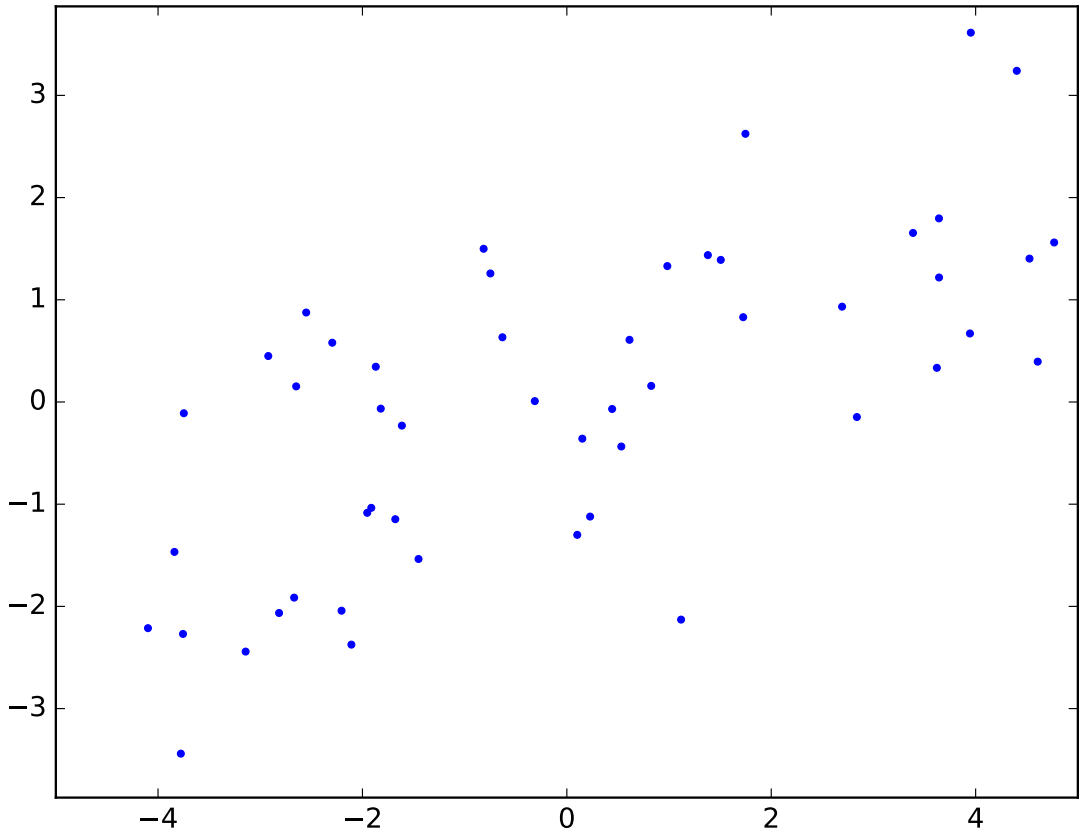
Suppose we want to estimate the error rate $\mu$ of our binary classification model $g : \mathbf{R}^d \to \{0, 1\}$. Our procedure will be to leave out $n$ data point in the test set when we fit our model, and to evaluate the model on each example $(x_i, y_i)$ in the test set. We will compute the fraction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

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

If we want to be sure that, with 95% probability, the the sample error rate $\nu$ is within .1 of true error rate $\mu$, 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.)

11. *Interpreting principal components.* We have a data set $x_1, \ldots, x_n \in \mathbf{R}^2$. The data are plotted below.




We stack the data points, and compute the singular value decomposition of the data,

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_n^T \end{bmatrix} = \sum_{i=1}^{2} \sigma_i u_i v_i^T \in \mathbf{R}^{n \times 2}.$$

(a) What is $v_1^T v_2$?

(b) Draw a vector on the plot corresponding to $v_1$ and another corresponding to $v_2$. (Label them clearly.)

(c) Looking at the plot, estimate $\frac{\sigma_1}{\sigma_2}$.

(d) Suppose you know $v_1$, $v_2$, $\sigma_1$, and $\sigma_2$, but not $u_1$ or $u_2$. Show how to find $u_1$. What is the geometric interpretation of your formula for $u_1$?

12. *More rats!* Recall Bob and Alice, who work in a lab studying rat DNA, from Homework

3. Bob and Alice are now collecting data on the telomere[1] length of rat chromosomes in order to further their research on the effects of aging. The organize their data into a matrix $X$, where each row corresponds to a specific rat, and each column corresponds to a specific chromosome of interest. The entry $X_{ij}$ records the length of the telomere of chromosome $j$ of rat $i$. The response is a vector $y$ which records the apparent age of each rat (assessed by an expert guesser of rat ages). Bob and Alice hypothesize that rats with shorter telomeres will look older than rats with longer telomeres.

In their analysis, both Bob and Alice try out least squares, ridge regression, and nonnegative least squares to fit the data. The only difference is the entries in Alice's matrix are in units of nucleotide *bases* (A,T,G,C), while the entries in Bob's matrix are in units of *tens of bases*.

(a) Why might Alice and Bob use ridge regression instead of least squares regression?

(b) Alice and Bob wonder whether the effect of telomere length on aging is mediated by only a small number of chromosomes; however, they see that the optimal regression coefficients found by both least squares and ridge regression are dense. Suggest an alternative method that might identify a small number of chromosomes which predict aging.

(c) Bob asks Alice to guess whether his least squares or ridge regression analysis produced a lower objective value. How would you suggest she figure out the answer?

(d) Alice tells Bob that the objective value she obtained from nonnegative least squares is the same as the objective value she obtained from standard least squares. She asks if he'd like to bet that her least squares solution has only nonnegative coefficients: he wins the best if her least squares solution is nonnegative; and otherwise loses the bet[2]. What should he check before accepting the wager?

(e) Alice and Bob finally compare their results. It turns out they get the same answer for linear regression, but a wildly different result from ridge regression (even when they use the same regularization parameter). Why? How can they arrive at a solution they both can agree on?

(f) Alice runs some more tests of telomere length on a new batch of rats, but discovers that her telomerometer[3] has malfunctioned and failed to record the lengths some of the rat telomeres. State at least three ways that Alice could salvage this data so that she could use it in her regression analysis. Which should she choose, and why? Is there anything else you'd want to know about the malfunctioning telomerometer to answer this question?

[1] A telomere is a sequence of repetitive nucleotide base pairs occuring at the end of a chromosome. They do not code for any protein. It is believed that they decrease in length with each cell division, limiting the number of possible cell divisions. Hence they are sometimes used as a proxy for the biological age of a cell.

[2] The winner of the bet gets to decide author order on the publication they coauthor.

[3] A telomerometer is an (imaginary) device for measuring the length of telomeres.

13. *Ordinal regression.* The problem of ordinal regression is to fit a model to an output space $\mathcal{Y} = \{1, 2, \ldots, k\}$ consisting of ordinal levels, which we represent as the integers $1, \ldots, k$. One method for ordinal regression is to define a loss function

$$\ell(y, z) = \sum_{i=1}^{k-1} \ell^{\text{bin}}(\psi(y)_i, z_i),$$

based on a binary loss function $\ell^{\text{bin}} : \{-1, 1\} \times \mathbf{R} \to \mathbf{R}$, where

$$\psi(y) = (1, \ldots, 1, \overbrace{-1}^{y\text{th entry}}, \ldots, -1) \in \{-1, 1\}^{k-1}$$

encodes the ordinal level $y \in \mathcal{Y}$ as a boolean vector. Here we will use hinge loss as our binary loss function

$$\ell^{\text{bin}}(\psi_i, z_i) = \ell_{\text{hinge}}(\psi_i, z_i) = (1 - (\psi_i z_i))_+.$$

Prove that the following inequality holds $\forall i \in \mathcal{Y}$, and explain its meaning:

$$\ell(j_1, \psi(i)) \leq \ell(j_2, \psi(i)), \text{if } |i - j_1| \leq |i - j_2|,$$

# Extra workspace