ELEC 378 — Machine Learning: Concepts & Techniques
Midterm Exam
Take-Home Due: 5pm Friday 10 March 2023

INSTRUCTIONS

1. This exam is CLOSED BOOK, CLOSED NOTES, and CLOSED ANY OTHER RESOURCE (including calculators and computers), except that you are allowed ONE $8\frac{1}{2} \times 11$ inch sheet of notes (both sides).

2. You have 3 hours to complete the exam; take it at one sitting. You should write in an exam booklet or on sheets of blank paper.

3. This test is to be completed on your own — NO COLLABORATION WITH OTHERS ALLOWED.

4. Include your note sheet at the end of your test. Minus 10 points if you don't include it.

5. Write clearly; if we can't read it, you won't get credit. Show your work.

6. Sign the pledge when you are finished.

7. **Submit the test on Gradescope.** Late tests will *not* be accepted.

NOTATION

- A **training data set** for supervised learning consists of n labeled data points $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$ with each $\mathbf{x}_i \in \mathbb{R}^p$.
- A **training data set** for unsupervised learning consists of n unlabeled data points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ with each $\mathbf{x}_i \in \mathbb{R}^p$.

1. **QUIKKIES**$^{\text{TM}}$ (32 points; 4 points each)

   (a) Provide an example of a sparse vector in $\mathbb{R}^5$.

   (b) Why is data centering crucial for the operation of PCA? A sketch will be useful in your explanation.

   (c) Why is either data centering or appending a column of ones to the data matrix crucial for the operation of linear regression? Provide an example, using formulae and/or plots, illustrating its importance.

   (d) Why do we introduce the ridge and/or Lasso penalty terms into linear regression? (That is, what issue do they solve?) When might one prefer the ridge penalty over Lasso and vice versa?

   (e) Where is the "linear" relationship that gives linear regression its name? Which relationship can be nonlinear (in linear regression)? When can this nonlinearity be useful? Give an example.

   (f) What is "overfitting?" Explain in terms of the (size) of the data matrix.

   (g) What is the property of the gradient of a convex function that makes it useful for optimization?

   (h) Explain the difference between stochastic gradient descent and gradient descent.

2. **Regression vs. PCA** (10 points)

Consider the problem of finding the "best" line (through the origin) that fits a collection of n centered data points $\{x_i \in \mathbb{R}, y_i \in \mathbb{R}\}_{i=1}^{n}$.

(a) How does linear regression solve this problem? Include a detailed sketch in your explanation.

(b) If we define the data matrix $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\intercal \\ \vdots \\ \mathbf{x}_n^\intercal \end{bmatrix} = \begin{bmatrix} x_1 & y_1 \\ \vdots & \vdots \\ x_n & y_n \end{bmatrix}$, how does PCA solve this problem? Include a detailed sketch in your explanation.

(c) Are the two lines found by linear regression and PCA the same? Why or why not?

(d) How do the linear regression and PCA objective functions differ in how they each define the "best" line?

3. **Orthogonality Principle** (15 points)

Recall the **orthogonality principle** that states that the optimal error vector $\mathbf{e}^* = \mathbf{y} - \mathbf{X}\mathbf{w}^* \in \mathbb{R}^n$ in a least-squares linear regression problem is orthogonal to any linear combination of the data vectors $\mathbf{X}\mathbf{w}$, where $\mathbf{X}$ is the $n \times p$ data matrix, $\mathbf{w}$ is an arbitrary vector in $\mathbb{R}^p$, and $\mathbf{w}^*$ contains the optimal least-squares linear regression coefficients.

(a) Draw a labeled depiction of this situation including the labels $\mathbf{y}$, the set of all linear combinations of the data vectors $\mathbf{X}\mathbf{w}$, the optimal predictor $\mathbf{X}\mathbf{w}^*$, and the optimal error vector $\mathbf{e}^*$.

(b) Starting from the orthogonality principle, derive an expression for the optimal least-squares linear regression coefficients $\mathbf{w}^*$ in terms of the data matrix $\mathbf{X}$ and labels $\mathbf{y}$. (i.e., **without** using the least-squares objective function or its gradient.)

4. **Linear Regression and Data Set Centroid** (15 points)

   Without loss of generality (but to make the math cleaner), consider least squares linear regression that fits a straight line to the data with $x_i, y_i \in \mathbb{R}$, i.e., $p = 1$. Assume that the data is *not* centered.

   Prove that the optimal least squares line always passes through the **centroid** of the data, i.e., the point $(\overline{x}, \overline{y})$ where

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \qquad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i.$$

   Hint: If you need to show that $\overline{e} = \frac{1}{n}\sum_{i=1}^{n} e_i = 0$, then you can appeal to the normal equations.