

Due Monday, 20 Jan 2020, by 11:59pm to Gradescope.

Covers material up to Introduction to machine learning.

100 points total.

1. (25 points) **Linear algebra refresher.**

(a) (12 points) Let \mathbf{A} be a square matrix, and further let $\mathbf{A}\mathbf{A}^T = \mathbf{I}$.

- i. (3 points) Construct a 2×2 example of \mathbf{A} and derive the eigenvalues and eigenvectors of this example. Show all work (i.e., do not use a computer's eigenvalue decomposition capabilities). You may not use a diagonal matrix as your 2×2 example. What do you notice about the eigenvalues and eigenvectors?
- ii. (3 points) Show that \mathbf{A} has eigenvalues with norm 1.
- iii. (3 points) Show that the eigenvectors of \mathbf{A} corresponding to distinct eigenvalues are orthogonal.
- iv. (3 points) In words, describe what may happen to a vector \mathbf{x} under the transformation $\mathbf{A}\mathbf{x}$.

(b) (8 points) Let \mathbf{A} be a matrix.

- i. (4 points) What is the relationship between the singular vectors of \mathbf{A} and the eigenvectors of $\mathbf{A}\mathbf{A}^T$? What about $\mathbf{A}^T\mathbf{A}$?
- ii. (4 points) What is the relationship between the singular values of \mathbf{A} and the eigenvalues of $\mathbf{A}\mathbf{A}^T$? What about $\mathbf{A}^T\mathbf{A}$?

(c) (5 points) True or False. Partial credit on an incorrect solution may be awarded if you justify your answer.

- i. Every linear operator in an n -dimensional vector space has n distinct eigenvalues.
- ii. A non-zero sum of two eigenvectors of a matrix \mathbf{A} is an eigenvector.
- iii. If a matrix \mathbf{A} has the positive semidefinite property, i.e., $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$ for all \mathbf{x} , then its eigenvalues must be non-negative.
- iv. The rank of a matrix can exceed the number of distinct non-zero eigenvalues.
- v. A non-zero sum of two eigenvectors of a matrix \mathbf{A} corresponding to the same eigenvalue λ is always an eigenvector.

2. (22 points) **Probability refresher.**

- (a) (9 points) A jar of coins is equally populated with two types of coins. One is type “H50” and comes up heads with probability 0.5. Another is type “H60” and comes up heads with probability 0.6.

- i. (3 points) You take one coin from the jar and flip it. It lands tails. What is the posterior probability that this is an H50 coin?
 - ii. (3 points) You put the coin back, take another, and flip it 4 times. It lands T, H, H, H. How likely is the coin to be type H50?
 - iii. (3 points) A new jar is now equally populated with coins of type H50, H55, and H60 (with probabilities of coming up heads 0.5, 0.55, and 0.6 respectively. You take one coin and flip it 10 times. It lands heads 9 times. How likely is the coin to be of each possible type?
- (b) (3 points) Consider a pregnancy test with the following statistics.
- If the woman is pregnant, the test returns “positive” (or 1, indicating the woman is pregnant) 99% of the time.
 - If the woman is not pregnant, the test returns “positive” 10% of the time.
 - At any given point in time, 99% of the female population is not pregnant.

What is the probability that a woman is pregnant given she received a positive test? The answer should make intuitive sense; given an explanation of the result that you find.

- (c) (5 points) Let x_1, x_2, \dots, x_n be identically distributed random variables. A random vector, \mathbf{x} , is defined as

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

What is $\mathbb{E}(\mathbf{Ax} + \mathbf{b})$ in terms of $\mathbb{E}(\mathbf{x})$, given that \mathbf{A} and \mathbf{b} are deterministic?

- (d) (5 points) Let

$$\mathbf{cov}(\mathbf{x}) = \mathbb{E}((\mathbf{x} - \mathbb{E}\mathbf{x})(\mathbf{x} - \mathbb{E}\mathbf{x})^T)$$

What is $\mathbf{cov}(\mathbf{Ax} + \mathbf{b})$ in terms of $\mathbf{cov}(\mathbf{x})$, given that \mathbf{A} and \mathbf{b} are deterministic?

3. (13 points) **Multivariate derivatives.**

- (a) (2 points) Let $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{y} \in \mathbb{R}^m$, and $\mathbf{A} \in \mathbb{R}^{n \times m}$. What is $\nabla_{\mathbf{x}} \mathbf{x}^T \mathbf{A} \mathbf{y}$?
- (b) (2 points) What is $\nabla_{\mathbf{y}} \mathbf{x}^T \mathbf{A} \mathbf{y}$?
- (c) (3 points) What is $\nabla_{\mathbf{A}} \mathbf{x}^T \mathbf{A} \mathbf{y}$?
- (d) (3 points) Let $\mathbf{A} \in \mathbb{R}^{n \times n}$, and $f = \mathbf{x}^T \mathbf{A} \mathbf{x} + \mathbf{b}^T \mathbf{x}$. What is $\nabla_{\mathbf{x}} f$?
- (e) (3 points) Let $f = \text{tr}(\mathbf{AB})$. What is $\nabla_{\mathbf{A}} f$?

4. (10 points) **Deriving least-squares with matrix derivatives.**

In least-squares, we seek to estimate some multivariate output \mathbf{y} via the model

$$\hat{\mathbf{y}} = \mathbf{W}\mathbf{x}$$

In the training set we're given paired data examples $(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$ from $i = 1, \dots, n$. Least-squares is the following quadratic optimization problem:

$$\min_{\mathbf{W}} \quad \frac{1}{2} \sum_{i=1}^n \left\| \mathbf{y}^{(i)} - \mathbf{W} \mathbf{x}^{(i)} \right\|^2$$

Derive the optimal \mathbf{W} .

Hint: you may find the following derivatives useful:

$$\begin{aligned} \frac{\partial \text{tr}(\mathbf{W} \mathbf{A})}{\partial \mathbf{W}} &= \mathbf{A}^T \\ \frac{\partial \text{tr}(\mathbf{W} \mathbf{A} \mathbf{W}^T)}{\partial \mathbf{W}} &= \mathbf{W} \mathbf{A}^T + \mathbf{W} \mathbf{A} \end{aligned}$$

5. (30 points) **Hello World in Jupyter.**

Complete the Jupyter notebook `linear_regression.ipynb`. Print out the Jupyter notebook and submit it to Gradescope.