CS 189 Introduction to Machine Learning
Spring 2018

EXAMREVIEW1

After the exam starts, please write your student ID (or name) on every odd page. On question 2, and select *all* correct answers. On questions 3-6, justify and show all your work. You may consult only one sheet of notes. Calculators, phones, computers, and other electronic devices are not permitted. There are **7** single sided pages on the exam. **Notify a proctor immediately if a page is missing.** You may, without proof, use theorems and lemmas that were proven in the notes and/or in lecture, unless we explicitly ask for a derivation. **You have 120 minutes**: there are 6 questions on this exam. Question 5 is out of scope for the Fall 2018 midterm.

Exam Location: Earth

PRINT and SIGN Your Name: ____________________, ____________________, ____________________
(last) (first) (signature)

PRINT Your Student ID: ____

Person before you: ________________________________, ________________________________
(name) (SID)

Person behind you: ________________________________, ________________________________
(name) (SID)

Person to your left: ________________________________, ________________________________
(name) (SID)

Person to your right: ________________________________, ________________________________
(name) (SID)

Row (front is 1): _____________ Seat (leftmost is 1): _____________ (*Include empty seats/rows.*)

1 Pre-exam Questions (4 points)

(a) **(2 points)** What is your favorite fruit? Your favorite vegetable?

(b) **(2 points)** What was your favorite machine learning topic?

Do not turn this page until your instructor tells you to do so.

Extra page. If you want the work on this page to be graded, mention it on the problem's main page.

2 Multiple Choice and Short Answer Questions

(a) What is the primary purpose of PCA?

○ Dimension reduction

○ Outlier removal

○ Linear regression

○ Optimization

(b) Which of the following is **not** always true about a Multivariate Gaussian distribution?

○ The isocontours are ellipses.

○ Covariance matrix has positive entries.

○ The PDF through the mean is Gaussian.

○ The PDF parallel to an axis is Gaussian.

(c) Which of the following is true about training and testing error?

○ Training measures bias; Testing measures variance.

○ Training measures variance; Testing measures bias.

○ Training measures bias; Testing measures both bias and variance.

○ Training measures variance; Testing measures both bias and variance.

(d) When preforming polynomial regression, how does training error, validation error, and testing error change with the degree of the polynomial?

○ Training Error decreases; Validation and Testing Error initial decrease, then increase.

○ Training Error initially decreases, then increases; Validation and Testing Error initial decrease, then increase;

○ Training Error decreases; Validation and Testing Error decrease.

○ Training Error initially decreases, then increases; Validation and Testing Error decrease.

(e) The left singular vectors of a rectangular matrix A are also:

○ Eigenvectors of AA^T

○ Eigenvectors of A^2

○ Eigenvectors of A^TA

○ Eigenvalues of AA^T

(f) The left singular vectors of a square matrix A are also:

○ Eigenvectors of AA^T

○ Eigenvectors of A^2

○ Eigenvectors of A^TA

○ Eigenvalues of AA^T

(g) Which of the following functions is **not** convex?

○ $f(x) = e^{-x}$ ○ $f(x) = \sin x$

○ $f(\mathbf{x}) = ||\mathbf{x}||_2^2$ ○ $f(x) = \max\{x, 0\}$

3 Multiple choice, multiple may be correct

For these questions, multiple options could be correct. You will only get credit if you provide all the correct choices, and no partial credit.

(a) Assume you have two zero mean random variables $X \in \mathbb{R}^{d_1}$ and $Y \in \mathbb{R}^{d_2}$. Let their covariance matrices by given by $\Sigma_{XX}, \Sigma_{YY}, \Sigma_{XY}$. Also define the random variables $X' = A^\top X$, and $Y' = B^\top Y$, where $A \in \mathbb{R}^{d_1 \times d_1}$ and $B \in \mathbb{R}^{d_2 \times d_2}$ are both matrices. Let $W_1 \in \mathbb{R}^{d_1 \times d_1}$ and $W_2 \in \mathbb{R}^{d_2 \times d_2}$ represent two unitary matrices. Also, for any matrix M, we let $U(M)$ denote the matrix of its left singular vectors, and $V(M)$ denote the matrix of right singular vectors.

Which of the following is/are correct?

(a) Choosing $A = \Sigma_{XX}^{-1/2}$ whitens X'.

(b) Choosing $B = \Sigma_{XX}^{-1/2}$ whitens Y'.

(c) Choosing $B = W_2 \Sigma_{YY}^{-1/2}$ whitens Y'.

(d) Choosing $A = U(\Sigma_{XX})$ decorrelates the entries of X'.

(e) Choosing $A = U(\Sigma_{XY})$ and $B = V(\Sigma_{XY})$ leads to diagonal $\Sigma_{X'Y'}$.

(f) If we want diagonal $\Sigma_{X'Y'}$ and whitened X', Y', using unitary matrices A, B is insufficient.

(b) Increasing λ in ridge regression has the following effect(s):

(a) The bias blows up without bound irrespective of the true model.

(b) The variance increases.

(c) The regularization strength increases.

(d) The model complexity decreases.

(e) The bias when $\lambda = 0$ is 0 if data is generated by a linear model.

(f) The bias when $\lambda = 0$ is always 0 irrespective of how the data is generated.

4 Estimation in linear regression

In linear regression, we estimate a vector $y \in \mathbb{R}^n$ by using the columns of a feature matrix $A \in \mathbb{R}^{n \times d}$. Assume that the number of training samples $n \geq d$ and that A has full column rank. You saw in homework how well we could predict y; let us now see how well we can estimate the regression coefficients.

Assume that the true underlying model for our noisy training observations is given by $Y = Aw^* + Z$, with $Z \in \mathbb{R}^n$ having iid $Z_j \sim \mathcal{N}(0,1)$ representing the random noise in the observation Y. Here, the $w^* \in \mathbb{R}^d$ is something arbitrary and not random. After obtaining $\hat{w} = \arg\min_w ||Y - Aw||_2^2$, we would like to bound the error $||\hat{w} - w^*||_2^2$, which is our error in estimating the underlying parameters w^*. Note that this is a random variable.

Having a good estimate of the parameters is the ultimate goal, since we then know exactly how the underlying model is generated.

(a) Using the standard closed form solution to the ordinary least squares problem, **show that**

$$\|\hat{w} - w^*\|_2^2 = \|(A^\top A)^{-1} A^\top Z\|_2^2.$$

(b) Use the (full) SVD of the matrix $A = U\Sigma V^\top$ to conclude that

$$\|\hat{w} - w^*\|_2^2 = \|V\Sigma' U^\top Z\|_2^2,$$

where we have denoted

$$\Sigma' = \begin{bmatrix} \Sigma_{\mathsf{inv}} & \mathbf{0}_{d\times(n-d)} \end{bmatrix}.$$

Here, we have used $\Sigma_{\mathsf{inv}} \in \mathbb{R}^{d\times d}$ to denote a diagonal matrix consisting of the reciprocals of the singular values of A, and $\mathbf{0}_{d\times(n-d)}$ denotes the $d \times (n-d)$ matrix of zeroes.

(c) What is the distribution of $U^\top Z$? Use unitary invariance of the ℓ_2-norm and the distribution you calculated to conclude that

$$\|\hat{w} - w^*\|_2^2 = \|\Sigma' Z'\|_2^2,$$

where $Z' \in \mathbb{R}^n$ is also i.i.d. standard Gaussian.

(d) Now conclude that

$$\mathbb{E}\left[\|\hat{w} - w^*\|_2^2\right] = \mathsf{trace}[(A^\top A)^{-1}].$$

Which of the following matrices A is better for estimation of the parameters?

i) $A_1 = \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix}$

ii) $A_2 = \begin{bmatrix} 10 & 0 \\ 0 & 2 \end{bmatrix}$

5 (NOT IN SCOPE FOR SP18 MIDTERM EXAM) Convergence Rate of Gradient Descent for Quadratic Functions

Show that the following problems have a geometric convergence rate when applying gradient descent with a fixed step size. Recall from Homework 6, Problem 3 that for a constant step size $\gamma = \frac{1}{\lambda_{\min}(A^T A) + \lambda_{\max}(A^T A)}$ and $A^T A$ positive definite,

$$\min_x \frac{1}{2}\|Ax - b\|_2^2$$

has geometric convergence. That is, for $Q = \frac{\lambda_{\max}(A^T A)}{\lambda_{\min}(A^T A)}$, we have

$$f(x_k) - f(x^*) = \frac{\lambda_{\max}(A^T A)}{2} \left(\frac{Q-1}{Q+1}\right)^{2k} \|x_0 - x^*\|_2^2$$

You may use the above result (without rederivation) if required for the following parts.

(a) Consider a matrix $A \in \mathbb{R}^{n\times d}, b \in \mathbb{R}^n, k \in \mathbb{R}$ such that $A^T A \succeq mI$ for $m > 0$:

$$\min_x \frac{1}{2}\|Ax - b\|_2^2 + k$$

Show geometric convergence for this problem.

(b) Consider $A \in \mathbb{R}^{n\times n}$ (a square matrix) and $b, c \in \mathbb{R}^n$, such that $A \succeq mI$ for $m > 0$:

$$\min_x \frac{1}{2}\|Ax - b\|_2^2 + c^T x$$

Show geometric convergence for this problem.

(c) Consider $A, C \in \mathbb{R}^{n\times d}, b, d \in \mathbb{R}^n$ such that $A^T A \succeq m_1 I, C^T C \succeq m_2 I$ for $m_1, m_2 > 0$:

$$\min_x \frac{1}{2}\|Ax - b\|_2^2 + \frac{1}{2}\|Cx - d\|_2^2$$

Show geometric convergence for this problem.

6 Parameter Estimation

Assume that $X_1, X_2, ..., X_n$ are i.i.d. samples from an exponential distribution $p(X = x) = \lambda e^{-\lambda x}$.

(a) Compute the maximum likelihood estimation of λ given $X_1, ..., X_n$.

(b) Now assume the the prior on λ has a gamma distribution with parameters α, β:

$$P(X = x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x},$$

where $\Gamma(\alpha)$ is the gamma function, which you should think of as a constant normalization that ensures that the PDF integrates to 1.

Show that the posterior random variable $\lambda | (X_1, X_2, \ldots, X_n)$ also has a gamma distribution.

(c) Compute the maximum a posteriori estimation of λ. Compare the result with the maximum likelihood estimation of λ when the sample size is large.

(d) Now assume that $X_1, X_2, ..., X_n$ are i.i.d. samples from a gamma distribution $P(X = x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$. In the next parts we will compute the gradient descent step for the maximum likelihood estimation of α and β. Assume that

$$\frac{d\Gamma(x)}{dx} = g(x)$$

Compute the partial derivative of log likelihood with respect to β and find β as a function of α.

(e) Using your result for the previous part, compute the gradient descent step for α.

The PDF of a multivariate n-dimensional Gaussian with mean μ and covariance matrix Σ is given by

$$f(x) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right).$$

Doodle page! Draw us something if you want or give us suggestions or complaints. You can also use this page to report anything suspicious that you might have noticed.