CS 189 Introduction to Machine Learning
Fall 2017

EXAMMIDTERM

After the exam starts, please write your student ID (or name) on every odd page. On questions 2, 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 **22** 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. Parts 5d and 6f are optional bonus questions. The multivariate Gaussian PDF is included on the last page; it is useful but may not be needed.

Exam Location: 310 Soda

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 Questions (14 points)

For these questions, select *all* the answers which are correct. You will get full credit for selecting all the right answers. On some questions, partial credit will be assigned.

(a) Increasing λ in ridge regression can be interpreted as performing a MAP estimate with a

◯ Gaussian prior with smaller variance

◯ Uniform prior with smaller range

◯ Gaussian prior with larger variance

◯ Uniform prior with larger range

(b) How does total least squares (TLS) compare to ordinary least squares (OLS)?

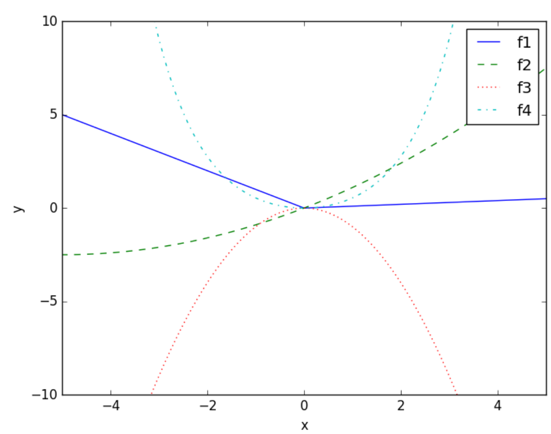
◯ TLS allows errors in X and y. OLS only allows errors in X.

◯ TLS only allows errors in X. OLS allows errors in X and y.

◯ TLS allows errors in X and y. OLS only allows errors in y.

◯ TLS only allows errors in y. OLS allows errors in X and y.

(c) (NOT IN SCOPE FOR SP18 EXAM) Which of the following functions are convex? They are drawn in the following plot and defined explicitly below:



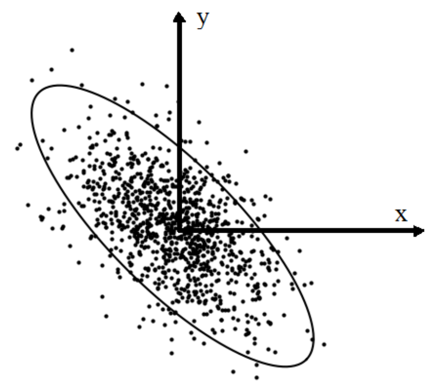
◯ $f_1(x) = \max\{-x, 0.1x\}$

◯ $f_2(x) = x + \frac{x^2}{10}$

◯ $f_3(x) = -x^2$

◯ $f_4(x) = \frac{e^x + e^{-x}}{2} - 1$

(d) Select which covariance matrix was most likely used to generate the following multivariate Gaussian distribution

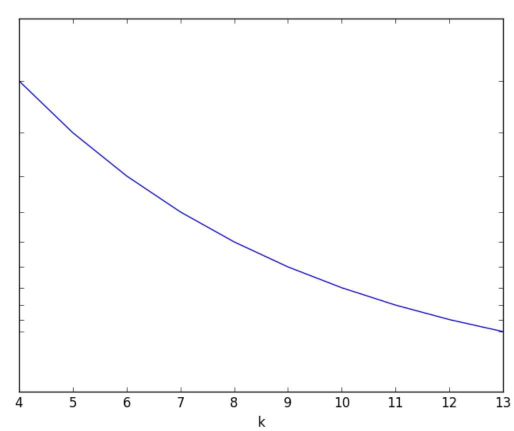


where the positive x direction is to the right and the positive y direction is up.

○ $\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$ ○ $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$

○ $\begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$ ○ $\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$

(e) Your friend at Google is training a machine learning model to predict a user's next search query based on their past k searches. She generates the following plot, where the value of k is on the x axis.

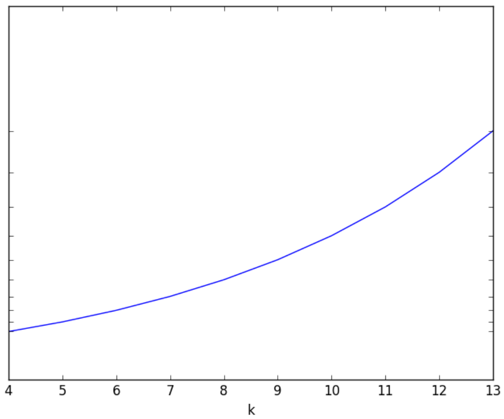


What might the y axis represent?

○ Training Error ○ Bias

○ Validation Error ○ Variance

(f) Your friend at Google is training a machine learning model to predict a user's next search query based on their past k searches. She generates the following plot, where the value of k is on the x axis.



What might the y axis represent?

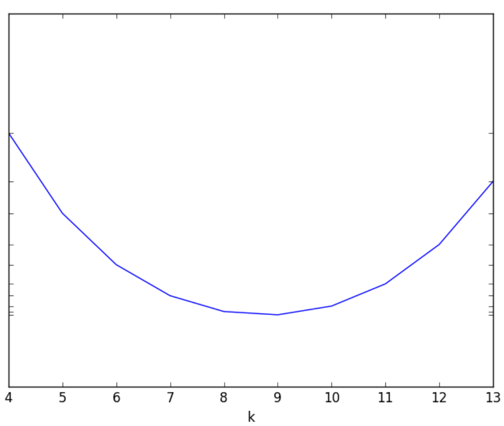
◯ Training Error

◯ Bias

◯ Validation Error

◯ Variance

(g) Your friend at Google is training a machine learning model to predict a user's next search query based on their past k searches. She generates the following plot, where the value of k is on the x axis.



What might the y axis represent?

◯ Training Error

◯ Bias

◯ Validation Error

◯ Variance

3 Short Answer Questions (7 points)

For these questions, you only need to give an answer. You do not need to show your work. You will only be judged on the correctness of your answer.

(a) (3 points) Suppose we have a covariance matrix for zero-mean X:

$$E[XX^T] = \Sigma = \begin{bmatrix} 4 & a \\ a & 1 \end{bmatrix}$$

What is the set of values that a can take on such that Σ is a valid covariance matrix?

(b) (4 points) Suppose that we observe the position and velocity of an object moving **along a line** in 3D space. At any point on the line, the object can have any speed. Our position observations measure the x, y, and z coordinates of the object, and the velocity observations measure the x, y, and z components of the velocity. We collect a large set of observations and run PCA on the set. **How many principal components would we expect to use to represent this data set?**

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

4 Parameter Estimation (19 points)

Assume that $X_1, X_2, ..., X_n$ are i.i.d. samples from a Poisson distribution:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

where $\lambda > 0$ and x is a non negative-integer. $P(X < 0) = 0$ and $P(X = x) = 0$ for all non-integer x.

(a) (3 points) **Compute the log likelihood of drawing samples** $X_1 = x_1, \ldots, X_n = x_n$ **for a given** λ**, i.e.,** $\ln p(x_1, x_2, \ldots, x_n|\lambda)$**.**

(b) (8 points) **Show that the negative of the log likelihood above is a convex function with respect to λ.** Use this fact to **compute the maximum likelihood estimate of λ given samples** $X_1 = x_1, \ldots, X_n = x_n$**.**

(c) (8 points) Now assume that we have a prior on λ with an exponential density $f(\lambda) = \alpha e^{-\alpha\lambda}$. **Compute the maximum a posteriori estimate of λ given samples $X_1 = x_1, \ldots, X_n = x_n$. Compare the MAP and MLE. What happens when $n \to \infty$?**

5 Linear regression (16 points)

In this problem, we study the loss function for ridge regression:

$$\frac{1}{2}\|Xw - y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2, \tag{1}$$

where X a data matrix in $\mathbb{R}^{n\times d}$ and y is the response in $\mathbb{R}^n$. Let the regularization weight $\lambda > 0$. Recall that the closed-form solution to ridge regression is

$$\hat{w} = (X^T X + \lambda I_d)^{-1} X^T y. \tag{2}$$

where $I_d \in \mathbb{R}^{d\times d}$ is an identity matrix of dimension d.

(a) (8 points) Augment the matrix X with d additional rows $ce_1^T, ce_2^T, \ldots, ce_d^T$ to get the matrix $X' \in \mathbb{R}^{n+d\times d}$, where c is a given constant and e_i^T is a unit vector whose ith element is 1 and the rest of the elements are zero, and augment y with d zeros to get $y' \in \mathbb{R}^{n+d}$:

$$X' = \begin{bmatrix} X \\ \\ ce_1^T \\ ce_2^T \\ \vdots \\ ce_d^T \end{bmatrix} \qquad y' = \begin{bmatrix} y \\ \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Write out the closed form solution for w' for the ordinary least squares problem on X', y'.

Derive the concrete value of c that corresponds to the ridge regression problem (1) in terms of λ. Conclude that the ridge regression estimate for w can be obtained by doing ordinary least squares on an augmented dataset.

(b) (8 points) **(NOT IN SCOPE FOR SP18 MIDTERM) Prove that applying gradient descent on the ridge regression loss function with a suitable fixed step size results in geometric convergence.** If $X^\top X$ has maximum and minimum eigenvalues M and m, **what fixed step size should we choose as a function of M, m, and ridge weight λ?**

Hint: Reduce the problem to ordinary least squares. Recall that applying gradient descent on the least squares problem

$$\min_x f(x) = \min_x \frac{1}{2}\|Ax - b\|_2^2$$

results in geometric convergence when $A^T A$ is positive definite and using a constant step size $\gamma = \frac{2}{\lambda_{\min}(A^T A) + \lambda_{\max}(A^T A)}$. Geometric convergence means that $f(x_k) - f(x^*) \leq c' Q^k$ for some $0 \leq Q < 1$ and for some $c' > 0$. You may use this result without proof.

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

6 Finding noisy low-rank matrices (21 points)

Assume you have an *unknown* matrix $M^* \in \mathscr{L}$, where $\mathscr{L}$ represents the set of all $d \times d$ square matrices of rank up to k. You observe M^* through noise as $Y = M^* + N$, where $N \in \mathbb{R}^{d \times d}$ is a noise matrix with $N_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$.

(a) (8 points) **Show that the maximum likelihood estimate of the matrix M^* is given by the optimization problem**

$$\widehat{M} = \arg\min_{M \in \mathscr{L}} \|Y - M\|_F^2, \tag{3}$$

where $\|X\|_F^2 = \sum_{i,j} X_{ij}^2$ is the squared Frobenius norm.

Hint: Begin by writing the likelihood $P(Y|M)$ for some matrix M. The exact definition of $\mathscr{L}$ is not relevant to this part; we will use this in the next part.

(b) (8 points) Recall that $\mathscr{L}$ was the class of $d \times d$ matrices of rank at most k. Assume that Y is full rank (i.e. invertible). Let us now consider the equivalent optimization problem

$$\widehat{M} = \arg \min_{M:\mathsf{rank}(M)=k} \|Y - M\|_F^2. \tag{4}$$

Write down a closed form expression to solve this problem when $k = d - 1$ in terms of the singular values and singular vectors of the matrix Y.

Hint: Your knowledge of solving total least squares problems may come in handy. (No need to derive the solution, if you can justify it with TLS.)

(c) (5 points) **BONUS: Write down a closed form expression to the problem above for arbitrary k in terms of the singular values and singular vectors of the matrix Y.**

7 (NOT IN SCOPE FOR SP18 MIDTERM) Multi-view Regression with CCA (30 points)

In robotics and other problem domains, simulations are a cheap source of training data. However, simulated data isn't perfect and might not present things in the same way that real data does. In this problem, we will show how CCA can help us use simulated data to learn by allowing us to focus on those dimensions of the simulation that actually correspond to the real world.

Let $X \in \mathbb{R}^d$ be a zero-mean random variable representing simulated data. Let $Q \in \mathbb{R}^d$ be a zero-mean random variable representing real-world data. Our goal is to learn the output $Y \in \mathbb{R}$ which is some bounded control signal (i.e. $Y \in [-1,1]$).

For example, one view Q could be images of an object taken from the real world and the other view X could be a corresponding image from a simulator. While the views are different we expect the robot to grasp (using control Y) the object in the same way.

Suppose that we know the simulation/real-world correspondences via their covariance matrix $\mathbb{E}[XQ^T] = \Sigma_{XQ}$. We also know the individual positive-definite covariance matrices Σ_{XX} and Σ_{QQ}.

(a) (8 points) In order to leverage CCA, we compute the canonical variates of our data and transform the data it into a space where the views are most correlated. Denote the canonical correlation matrix

$$C = \Sigma_{XX}^{-\frac{1}{2}} \Sigma_{XQ} \Sigma_{QQ}^{-\frac{1}{2}}.$$

Let $C = U \Lambda V^T$ be its singular value decomposition. We can transform our data to the canonical variates via the following:

$$\hat{X} = U^T \Sigma_{XX}^{-\frac{1}{2}} X \qquad \hat{Q} = V^T \Sigma_{QQ}^{-\frac{1}{2}} Q$$

Show that $\mathbb{E}[\hat{X}\hat{Q}^T] = \Sigma_{\hat{X}\hat{Q}} = \Lambda$ and $\Sigma_{\hat{X}\hat{X}} = \Sigma_{\hat{Q}\hat{Q}} = I$.

Here, we use the symmetric square root: For a positive definite matrix Σ having eigenvalue decomposition $\Sigma = \bar{U}\bar{\Lambda}\bar{U}^\top$, the symmetric square-root is given by $\Sigma^{\frac{1}{2}} = \bar{U}\bar{\Lambda}^{\frac{1}{2}}\bar{U}^T$.

(b) (4 points) To focus attention on the dimensions of the simulation that most correspond to the real world, we can use regularization while we attempt to learn the best linear map from simulated X to control Y. We define

$$\hat{w} = \underset{w \in \mathbb{R}^d}{\operatorname{argmin}} \; \mathbb{E}[(Y - w^T \hat{X})^2] + ||w||^2_{\text{CCA}} \tag{5}$$

where

$$||w||^2_{CCA} = \sum_{i=1}^{d} \frac{1-\lambda_i}{\lambda_i}(w_i)^2.$$

Here, the λ_i correspond to the diagonal elements of Λ, and satisfy $0 < \lambda_i \leq 1$ by the properties of CCA. The weighted norm here penalizes dependencies on those aspects of simulated X that are less correlated to the real Q. The expectation is taken over both random variables $\hat{X} \in \mathbb{R}^d$ and $Y \in \mathbb{R}$.

Show the following is true

$$\mathbb{E}[(Y - w^T \hat{X})^2] = \mathbb{E}[Y^2] + \|w\|_2^2 - 2w^T \mathbb{E}[Y\hat{X}].$$

(c) (8 points) **Show the global minimizer $\hat{w}$ in Eq.** (5) **has coordinate i given by**

$$\hat{w}_i = \lambda_i \mathbb{E}[Y(\hat{X})_i],$$

where $(\hat{X})_i$ is the i-th coordinate of $\hat{X}$.

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

(d) (10 points) Let us take n samples $(x^1, y^1), ..., (x^n, y^n)$ of the view X and the corresponding outputs Y, where each sample $x^j \in \mathbb{R}^d$ and $y^j \in \mathbb{R}$. We now transform the data to the canonical variates to obtain $\hat{x}^j = U^T \Sigma_{XX}^{-\frac{1}{2}} x^j$. We now use the data to create empirical estimates for $\hat{w}_i = \lambda_i \mathbb{E}[Y(\hat{X})_i]$ by defining

$$\tilde{w}_i = \lambda_i \frac{1}{n} \sum_{j=1}^{n} y^j (\hat{x}^j)_i$$

for each $i \in \{1, 2, \ldots, d\}$. As before, $(\hat{x}^j)_i$ denotes the ith coordinate of $\hat{x}^j$.

Let us examine how fast $\tilde{w}$ converges to $\hat{w}$. Notice that $\mathbb{E}[\tilde{w}] = \hat{w}$, therefore

$$\mathbb{E}[||\tilde{w} - \hat{w}||_2^2] = \underbrace{\mathbb{E}||\tilde{w} - \mathbb{E}(\tilde{w})||_2^2}_{\text{Variance}}$$

Show that

$$\mathbb{E}[||\tilde{w} - \hat{w}||_2^2] \leq \sum_{i=1}^{d} \frac{\lambda_i^2}{n}.$$

Hint: Remember the random variable $Y \in [-1, 1]$ and so $Y^2 \leq 1$.

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.