CS168, Spring 2016

CS168 Final Exam

(Do not turn this page until you are instructed to do so!)

Instructions: This is closed-book exam, and you are permitted to refer only to one double-sided sheet of notes, which you should have prepared in advance (though the notes should not be necessary to complete the exam). You have 3 hours and the exam is worth 138 points. Make sure you print your name legibly and sign the honor code below. All of the intended answers can be written well within the space provided. You can use the back of the preceding page for scratch work. If you want to use the back side of a page to write part of your answer, be sure to mark your answer clearly. Good luck!

The following is a statement of the Stanford University Honor Code:

A. The Honor Code is an undertaking of the students, individually and collectively:

(1) that they will not give or receive aid in examinations; that they will not give or receive unpermitted aid in class work, in the preparation of reports, or in any other work that is to be used by the instructor as the basis of grading;

(2) that they will do their share and take an active part in seeing to it that others as well as themselves uphold the spirit and letter of the Honor Code.

B. The faculty on its part manifests its codence in the honor of its students by refraining from proctoring examinations and from taking unusual and unreasonable precautions to prevent the forms of dishonesty mentioned above. The faculty will also avoid, as far as practicable, academic procedures that create temptations to violate the Honor Code.

C. While the faculty alone has the right and obligation to set academic requirements, the students and faculty will work together to establish optimal conditions for honorable academic work.

I acknowledge and accept the Honor Code.

_____

(Signature)

_____

(Print your name, legibly!)

Problem	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	Total
Score											
Maximum	25	5	7	10	14	10	16	8	15	28	138

1. **Miscellaneous short answer.** (25 points)

   (a) (3 points) List 3 strategies that can be used to make further progress after gradient descent seems stuck at a local minimum.

   (b) (3 points) You want to study the relationship between state of residence, ethnicity, and attitudes towards gun-ownership in the United States. Why might tensor methods be a good approach to try out for this problem?

   (c) (5 points) Which of the following functions are convex? You do not need to provide a proof, just circle the functions that are convex:

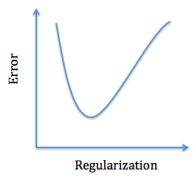
   i. e^x
   ii. e^{-x}
   iii. $\log^2(1+x)$
   iv. The ℓ_2 norm of a vector
   v. The squared ℓ_2 norm of a vector

   (d) (4 points) Which of the following sets are guaranteed to be convex? You do not need to provide a proof, just circle the convex sets:

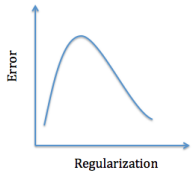
   i. The feasible region of a linear program
   ii. The intersection of convex sets
   iii. The union of convex sets
   iv. The set of rank-1 n by n matrices

(e) (3 points) What is the connection between convex functions and convex sets?

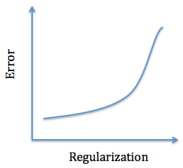
(f) (4 points) Consider a linear regression problem with ℓ_2 regularization (i.e. a penalty on the learned weight vector $(w_1, w_2, \ldots)$ of the form $\lambda \sum_i w_i^2$ for some constant λ.) Label the plot in Figure 1 that is is most likely to depict the final training error (i.e., error on training data) as a function of the amount of regularization, λ, and the plot that is most likely to depict the relationship between the test error (i.e., error on new data of the same type) of the computed prediction function, as a function of λ.



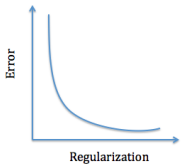
(a) Plot 1



(b) Plot 2



(c) Plot 3



(d) Plot 4

Figure 1: Possible plots of error vs. regularization

(g) (3 points) Someone claims to have invented a magical hash function that is guaranteed to spread every data set out almost equally (i.e., for each i in the function's range $\{0, 1, 2, \ldots, n-1\}$, a roughly $1/n$ fraction of the data set hashes to i). Explain why, without even looking at the description of the magical hash function, you know that this claim is false.

2. **Sampling from unit balls.** (5 points)

   (a) (3 points) How can one select a d-dimensional vector uniformly at random from the ℓ_2 unit ball (i.e. from the set of points whose euclidean distance from the origin is 1)? (Write at most one sentence.)

   (b) (1 point) How can one select a d-dimensional vector uniformly at random from the ℓ_1 unit ball (i.e. from the set of points whose ℓ_1 distance from the origin is 1)? (Write at most one sentence.)

   (c) (1 point) How can one select a d-dimensional vector uniformly at random from the ℓ_∞ unit ball (i.e. from the set of points whose ℓ_∞ distance from the origin is 1)? (Write at most one sentence.)

3. **Markov Chain Monte Carlo.** (7 points) Consider the problem MAX-3SAT, where you are given a set of disjunctions (i.e., clauses of the form $l_i \vee l_j \vee l_k$, where each l_i is either a variable x_i or its negation $\neg x_i$), and the goal is to assign $True/False$ to the variables $x_1, \ldots, x_n$ to maximize the number of clauses that are satisfied. Suppose you try to solve this problem via a Markov Chain Monte Carlo approach:

   (a) (3 points) What is the size of the state space of this Markov Chain?

   (b) (4 points) Describe a transition rule that might be effective for this Markov Chain. (Write at most two sentences.)

4. **PCA vs. Least Squares.** (10 points)

For questions (a) and (b), choose an answer from the following choices:

- PCA
- Least squares (i.e., linear regression)
- Both
- None of the above

Questions:

(a) (2 points) Finds the best fit line that minimizes the average squared Euclidean distance between the line and the data points.

(b) (2 points) Finds the best fit line that minimizes the sum of square of residuals $\sum_i (y_i - \hat{y}_i)^2$.

(c) (2 points) Given a set $(x_1, y_1), \ldots, (x_n, y_n)$ of two-dimensional points, suppose we run linear regression twice, first regressing x onto y, and then regressing y onto x. Will we get the same best-fit line in both cases?

(d) (4 points) Consider the set of points depicted in Figure 2, with the horizontal axis corresponding to the x coordinate of each point, and the vertical axis corresponding to the y-coordinate. Draw two lines, one depicting the first principal component of the point set, and the second depicting the least squares regression line corresponding to fitting y as a linear function of x. Make sure you clearly label which line is which.

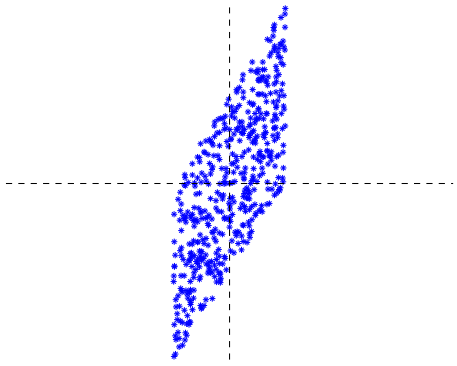


Figure 2: Draw two lines, and label one "PCA" and one "least squares".

5. **Linear regression.** (14 points)

(a) (3 points) Suppose you have one-dimensional points $x_1, ..., x_m$ with real-valued labels $y_1, ..., y_m$. Explain how to use linear regression to compute the prediction function of the form $f(x) = ax^2 + bx + c$ that minimizes (over all choices for the constants a, b, c) the mean squared error between the predictions and the labels (i.e., $\frac{1}{m}\sum_i (f(x_i) - y_i)^2$).

(b) (4 points) Consider the generalization of (a) where you allow prediction functions that are higher-degree polynomials. State one advantage and one disadvantage of allowing larger degrees.

(c) (4 points) Consider linear regression with n-dimensional points $x_1, ..., x_m \in \mathbb{R}^n$ and real-valued labels $y_1, ..., y_m$. What is the gradient of the mean-squared error function? Give a fully precise mathematical answer, and also an intuitive interpretation.

(d) (3 points) How does the gradient in (c) change in the presence of an ℓ_2 regularization term?

6. **Sampling and Estimation.** (10 points) Suppose you are conducting a poll to decide what fraction of the population support a soda tax.

   (a) (5 points) Suppose you poll 10,000 randomly chosen people, of which $5,021$ support the tax. Roughly what do you expect the percent error in the outcome of this poll to be? Explain your answer with at most two sentences. [You might find it helpful to refer to Chebyshev's inequality, which states that for a random variable X and any $c > 0$, $\Pr\left[|X - E[X]| > c\sqrt{Var[X]}\right] \leq 1/c^2$.]

   (b) (5 points) Now suppose you know that exactly 50% of the population is over age 30, and the rest is under age 30. You know that roughly 75% of people over 30 support the tax, and roughly 25% of the people under 30 support the tax. How could you design a poll of $10,000$ people so as to improve the percentage error that you expect? Explain your answer with at most two sentences. [Hint: if you let n_{older} denote the number of random people over 30 that you poll, and $n_{young} = 10,000 - n_{old}$ denote the number of under-30's you poll, what is an estimate of the variance of the result of the poll, as a function of n_{old} and n_{young}?]

7. **Similarity and the SVD.** (16 points) Your aunt runs a small local news website, with articles falling into four different categories: *Politics, Sports, Crime, and Weather*. Fresh from CS168, you decide to help her out and spend a few weeks in the summer to add a 'Related Articles' feature to the website which would suggest articles similar to the article being currently read, so that people spend more time on the website. You build a large *document-word* matrix M, where the rows index articles and the columns index words. Each entry in the matrix $M[i, j]$ represents how many times word j appears in article i.

   (a) (2 points) We have studied at least 3 similarity metrics between vectors: cosine similarity, ℓ_2 distance, and Jaccard similarity. Name at least one similarity metric that you might expect to do a good job capturing the similarity between documents. Name at least one metric that you expect to be a poor choice. (Explain each decision with at most one sentence each.)

   (b) (2 points) Because there are so many distinct words, the vector representing each document is very long. Suppose you wish to *approximate* the ℓ_2 distance between all pairs of documents (i.e. between the columns of M). What technique can you employ as pre-processing that will reduce the amount of time it will take to compute these distances? (Just state the one word or phrase—no explanation is necessary.)

   (c) (2 points) Suppose you then wish to *approximate* the Jaccard similarity between all pairs of documents. What technique can you employ to reduce the amount of time it will take to compute these distances?

   We now consider applying the SVD to understand the structure of the document-word matrix M. Consider computing the SVD $M = USV^T$. We will try and understand what the columns of U and V represent.

   As a very simple example, suppose the document-word matrix M is the following:

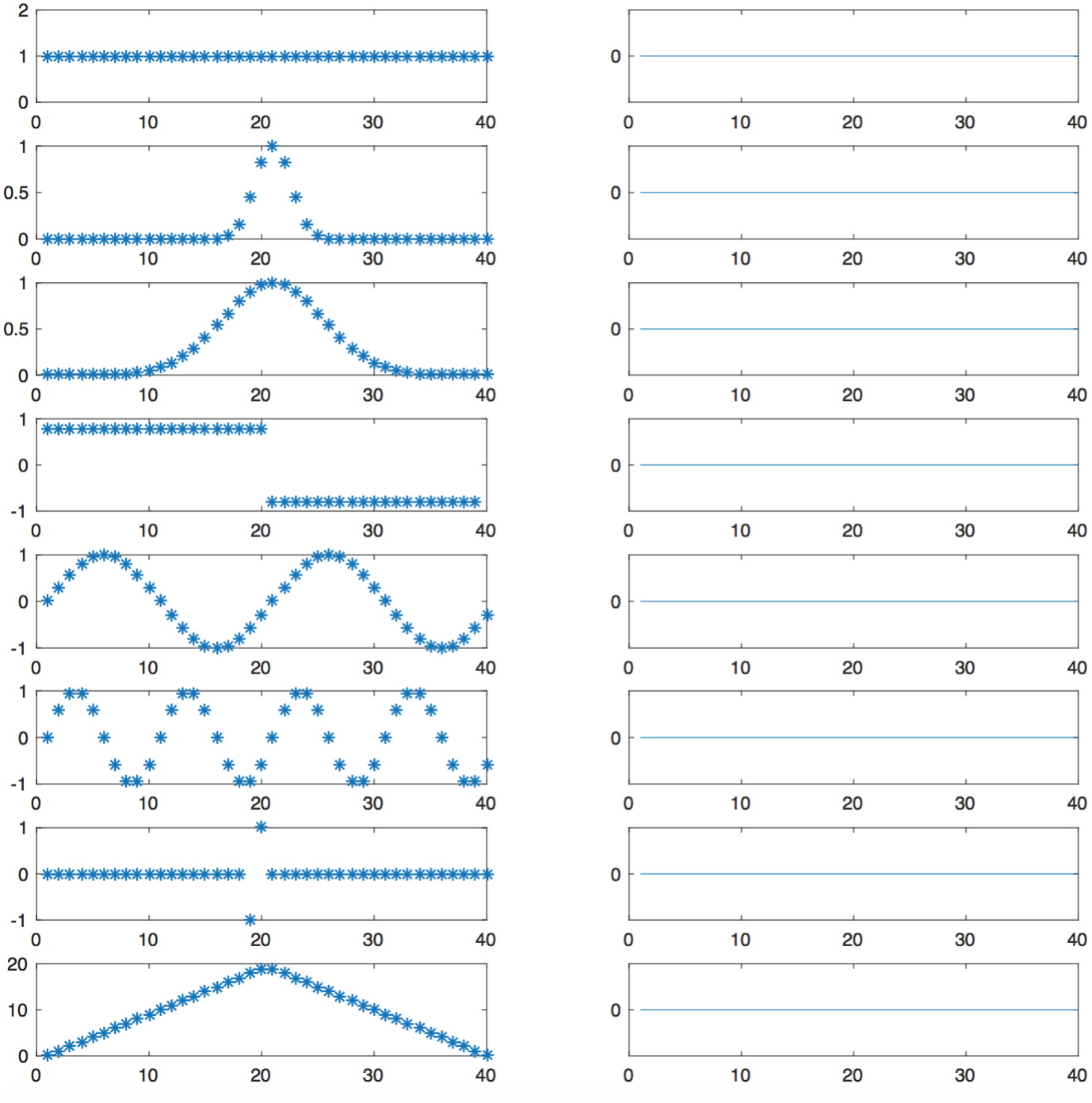
$$M = \begin{bmatrix} 5 & 5 & 0 & 0 & 0 \\ 5 & 5 & 0 & 0 & 0 \\ 0 & 0 & 5 & 5 & 5 \\ 0 & 0 & 5 & 5 & 5 \end{bmatrix}$$

   (d) (2 points) What is the rank of M?

(e) (4 points) Compute the SVD of the 4×5 matrix M given above. Specifically, if r is the rank of M, then write matrices U, S and V such that $M = USV^T$, where U is a $4 \times r$ matrix, D is a $r \times r$ matrix and V^T is a $r \times 5$ matrix.

(f) (4 points) Now imagine an actual (large) real-world document-word matrix. Suppose that, in general, articles on the same topic will use similar vocabularies, and that the different topics (Sports, Politics, Crime, Weather) will each use significantly different vocabularies. Of course, common words like "the" and "and" will occur often in all documents. What would you expect the columns of U corresponding to the largest singular values to represent for your news website articles? What you would expect the columns of V corresponding to the largest singular values to represent? (Write at most 2 sentences for each part. Do not just say that "they are the singular vectors" — you should give concrete and meaningful potential interpretations.)

8. **Fourier transformations.** (8 points) For each of eight signals in the left column, draw a plot in the corresponding space depicting the magnitude of the Fourier transform. Specifically, sketch a plot of the length 40 vector whose ith coordinate represents the magnitude of the (complex number) representing the ith index of the Fourier transform of the corresponding length 40 signal on the left.



9. **Sparse Recovery.** (15 points)

(a) (2 points) Suppose you want to recover an unknown signal z (of length n) from linear measurements $b = Az$, where A is an m by n matrix that you get to design. What's unsatisfactory about just taking $A = I$, where I is the $n \times n$ identity matrix?

(b) (2 points) Prove that, with no assumptions on z, there is no matrix A with fewer than n rows for which z can always be recovered from $b = Az$.

(c) (3 points) Consider the recovery of k-sparse signals with $k = \log n$. Explain why if A has $m = O(\log^2 n)$ rows, and each row has only 5 non-zero entries, then it is not be possible to recover every k-sparse signal z from $b = Az$.

(d) (4 points) Suppose you know a matrix A such that every signal z which is k-sparse in the standard basis can be recovered from $b = Az$. How can you then recover signals which are k-sparse in the Fourier basis?

(e) (4 points) State the ℓ_1-minimization problem for this setting where we wish to recover a signal z from measurements $b = Az$. Prove that this problem can be formulated as a linear program.

10. **The Algorithmic Toolbox.** (28 points) For each of the following scenarios, select the technique/tool from the list that is best suited to the problem. Note that some tools/techniques might be matched with several scenarios, and some tools might not be matched with any scenarios. For each question, identify a tool/technique, and **provide a one-sentence explanation of how or why the technique should be applied.**

    Reminder of tools/techniques: Consistent Hashing, Count-Min-Sketch, Dimension Reduction, k-d Trees and Nearest Neighbor Search, PCA, SVD and low-rank matrix approximation, Importance Sampling, Markov Chain Monte Carlo, Fourier Analysis/Convolution, Compressive Sensing, Linear/Convex Programming, Gradient Descent, Spectral Graph Theory.

    (a) (4 points) After a stellar track record, your database security startup begins to contract for a number of government agencies. In addition to maintaining secure datacenters, the agencies would also like to have a list of "suspicious" IP addresses to blacklist. Specifically, you want to compile a list of the IP addresses that are responsible for the largest number of suspicious access attempts, and you would like to implement an extremely fast/lightweight system that avoids keeping a log of all accesses.

    (b) (4 points) During your tenure as the head data scientist for the CDC (Center for Disease Control), you discover that many of the new antibiotic resistant bacteria originate from one reasonably small geographic area that contains 500 chicken farms. Currently, all 500 chicken farms use all 6 commonly available antibiotics. With the hopes of being able to contain the evolution and spread of new antibiotic resistant strains of bacteria, you want to assign 1 of the 6 antibiotics to each of the 500 farms, in such a way that nearly every farm uses an antibiotic that is different from the antibiotics used by the neighboring farms. Give TWO approaches to deciding this assignment of antibiotics to farms.

(c) (4 points) You are in charge of a large biology lab that is trying to figure out the regulatory mechanism of cells, and wish to map out the entire structure of which genes regulate which other genes (i.e. for which genes X and Y does the amount of X affect the amount of Y that the cell produces). You care about roughly $1,000$ genes, and have a fancy new technology that lets you selectively "turn off" different genes, and measure the amount of other genes. As a first step, you want to be able to predict, for every pair X and Y, how much of gene Y will be produced if you "turn off" gene X. It is way too expensive to directly run all 1000^2 tests—how can you hope to get around this?

(d) (4 points) After successfully reconstructing an accurate approximation of the gene interaction matrix, you now want to begin to tease apart the different interaction networks. As a first step, you would like to partition the set of genes into groups with the property that there are relatively few interactions between different groups. How might you do this?

(e) (4 points) After four years of taking classes at Stanford, and dutifully filling out class surveys, you are frustrated by the fact that you can not directly see how other students have rated a given class. You volunteer to help Stanford's course coordinator for a summer, and decide to go completely overboard: you plan to make a system that will let students enter ratings for courses they have taken, and then suggest courses that they might enjoy. What techniques might be at the core of this system?

(f) (4 points) You are tasked with designing and building the next generation space telescope. All goes well, and the telescope is deployed. At some point, a tiny piece of space dust hits the telescope, causing the entire module to continuously spin in a predictable (and measurable) way. Now, all the pictures show stars as arcs, and nebulous features (comets, etc.) as blurs. Rather than sending up a space shuttle to steady the telescope, you wish to solve the problem by simply post-processing the images. How might you consider doing this?

(g) (4 points) Your proposed fix to account for the spinning telescope garnered the admiration and respect of your colleagues, and saved NASA millions of dollars. Ten years later, some high-energy particles knock out all but one of the photo-receptors in the telescope's camera, essentially turning it into a camera that can only take single-pixel photos. Your colleagues are ready to retire this telescope, which has already outlived its initial specification and provided deep new insights into star formation and the atmospheric chemistry of extra-solar planets. You, however, realize that you might be able to leverage the fact that, while there is only one remaining photo-receptor, you have a large variety of different focal-lengths/lenses that you can use with the camera. How might you be able to save the day again?

[This page intentionally left blank]