Homework 1
Background Test Solution

CMU 10601: Machine Learning (Spring 2016)
OUT: Jan. 13, 2016
DUE: 5:30 pm, Jan. 21, 2016
TAs: William Herlands and Han Zhao

Guidelines

The goal of this homework is for you to determine whether you have the mathematical background needed to take this class, and to do some background work to fill in any areas in which you may be weak. Although most students find the machine learning class to be very rewarding, it does assume that you have a basic familiarity with several types of math: calculus, matrix and vector algebra, and basic probability. You do not need to be an expert in all these areas, but you will need to be conversant in each, and to understand:

- Basic calculus (at the level of a first undergraduate course). For example, we rely on you being able to take derivatives. During the class you might be asked, for example, to calculate derivatives (gradients) of functions with several variables.
- Linear algebra (at the level of a first undergraduate course). For example, we assume you know how to multiply vectors and matrices, and that you understand matrix inversion.
- Basic probability and statistics (at the level of a first undergraduate course). For example, we assume you know how to find the mean and variance of a set of data, and that you understand basic notions such as conditional probabilities and Bayes rule. During the class, you might be asked to calculate the probability of a data set with respect to a given probability distribution.
- Basic tools concerning analysis and design of algorithms, including the big-O notation for the asymptotic analysis of algorithms.

For each of these mathematical topics, this homework provides (1) a minimum background test, and (2) a medium background test. If you pass the medium background tests, you are in good shape to take the class. If you pass the minimum background, but not the medium background test, then you can still successfully take and pass the class but you should expect to devote some extra time to fill in necessary math background as the course introduces it. If you cannot pass the minimum background test, we suggest you fill in your math background before taking the class. Here are some useful resources for brushing up on, and filling in this background.

Probability

- Lecture notes: http://www.cs.cmu.edu/~aarti/Class/10701/recitation/prob_review.pdf.

Linear Algebra:

- Short video lectures by Prof. Zico Kolter: http://www.cs.cmu.edu/~zkolter/course/linalg/outline.html.
- Handout associated with above video: http://www.cs.cmu.edu/~zkolter/course/linalg/linalg_notes.pdf.
- Book: Gilbert Strang. Linear Algebra and its Applications. HBJ Publishers.

Matlab tutorial

- http://www.math.mtu.edu/~msgocken/intro/intro.pdf.
- http://ubcmatlabguide.github.io/.

Big-O notation:

- http://www.stat.cmu.edu/~cshalizi/uADA/13/lectures/app-b.pdf
- http://www.cs.cmu.edu/~avrim/451f13/recitation/rec0828.pdf
- See ASYMPTOTIC ANALYSIS (Week 1) in the following: https://class.coursera.org/algo-004/lecture/preview

Instructions

- **Submit your homework** on time **BOTH** electronically by submitting to Autolab **AND** by dropping off a hardcopy in the bin outside Gates 8221 by 5:30 pm, Thursday, January 21, 2016.
- **Late homework policy**: Homework is worth full credit if submitted before the due date, half credit during the next 48 hours, and zero credit after that. Additionally, you are permitted to drop 1 homework.
- **Collaboration policy**: For this homework only, you are welcome to collaborate on any of the questions with anybody you like. However, you *must* write up your own final solution, and you must list the names of anybody you collaborated with on this assignment. The point of this homework is not really for us to evaluate you, but instead for *you* to determine whether you have the right background for this class, and to fill in any gaps you may have.

Minimum Background Test [80 pts]

Vectors and Matrices [20 pts] (HZ)

Consider the matrix X and the vectors $\mathbf{y}$ and $\mathbf{z}$ below:

$$X = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \qquad \mathbf{z} = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$$

1. What is the inner product of the vectors $\mathbf{y}$ and $\mathbf{z}$? (this is also sometimes called the *dot product*, and is sometimes written as $\mathbf{y}^T\mathbf{z}$)
   Solution:
   $$\mathbf{y}^T\mathbf{z} = 1 \times 3 + 2 \times 4 = 11$$
2. What is the product $X\mathbf{y}$?
   Solution:
   $$X\mathbf{y} = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} = \begin{pmatrix} 5 \\ 11 \end{pmatrix}$$
3. Is X invertible? If so, give the inverse, and if no, explain why not.
   Solution:
   $$det(X) = 1 \times 4 - 2 \times 3 = -2 \neq 0$$

   X is invertible as its determinant is not 0.

   The inverse matrix can be computed by:
   $$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}^{-1} = \frac{1}{det(X)} \begin{pmatrix} 4 & -2 \\ -3 & 1 \end{pmatrix} = \begin{pmatrix} -2 & 1 \\ \frac{3}{2} & -\frac{1}{2} \end{pmatrix}$$
4. What is the rank of X?
   Solution: The rank of X is 2 since it is invertible.

Calculus [20 pts] (HZ)

1. If $y = x^4 + 2x^2 - 1$ then what is the derivative of y with respect to x?
   Solution:
   $$\frac{dy}{dx} = 4x^3 + 4x$$
2. If $y = \log(\frac{x^7}{10x}) + \sin(z)x^{z-8}$, what is the partial derivative of y with respect to x?
   Solution:
   $$\frac{\partial y}{\partial x} = \frac{6}{x} + (z-8)\sin(z)x^{z-9}$$

Probability and Statistics [20 pts] (HZ)

Consider a sample of data $S = \{0, 1, 1, 0, 0\}$ created by flipping a coin x five times, where 0 denotes that the coin turned up heads and 1 denotes that it turned up tails.

1. What is the sample mean for this data?
   Solution: The sample mean is
   $$\hat{p} = \frac{1+1}{5} = 0.4$$

2. What is the sample variance for this data?
   Solution: The sample variance is
   $$\hat{S} = \frac{2}{5} - \left(\frac{2}{5}\right)^2 = \frac{6}{25}$$

3. What is the probability of observing this data, assuming it was generated by flipping a biased coin with $p(x = 1) = 0.6, p(x = 0) = 0.4$.
   Solution:
   $$p(\{0, 1, 1, 0, 0\}) = 0.4 \times 0.6 \times 0.6 \times 0.4 \times 0.4 = 0.02304$$

4. Note that the probability of this data sample would be greater if the value of $p(x = 1)$ was not 0.6, but instead some other value. What is the value that maximizes the probability of the sample S? Please justify your answer.
   Solution: The value that maximizes the probability of the sample is
   $$\hat{p}(x = 1) = 0.4, \quad \hat{p}(x = 0) = 0.6$$
   To see why this is the case, define $p = \hat{p}(x = 1)$ and consider the likelihood of the sample as a function of p:
   $$\mathcal{L}(p) = p^2(1-p)^3, \qquad 0 \le p \le 1$$
   By the geometric-arithmetic mean inequality, we have
   $$\begin{aligned}\mathcal{L}(p) &= \frac{4}{9}\frac{9}{4}p^2(1-p)^3 \\ &\le \frac{4}{9}\left(\frac{1.5p + 1.5p + 1 - p + 1 - p + 1 - p}{5}\right)^5 \\ &= \frac{2^2 \times 3^3}{5^5}\end{aligned}$$
   where the equality holds when $1.5p = 1 - p$, i.e., $p = 0.4$.

5. Consider the following joint probability table where both A and B are binary random variables:

A	B	$P(A, B)$
0	0	0.5
0	1	0.1
1	0	0.3
1	1	0.1

(a) What is $P(A = 0, B = 0)$?
Solution:

$$P(A = 0, B = 0) = 0.5$$

(b) What is $P(A = 1)$?
Solution:

$$P(A = 1) = P(A = 1, B = 0) + P(A = 1, B = 1) = 0.4$$

(c) What is $P(A = 0|B = 1)$?
Solution:

$$P(A = 0|B = 1) = \frac{P(A = 0, B = 1)}{P(B = 1)} = \frac{P(A = 0, B = 1)}{P(A = 0, B = 1) + P(A = 1, B = 1)} = \frac{0.1}{0.1 + 0.1} = 0.5$$

(d) What is $P(A = 0 \vee B = 0)$?
Solution:

$$P(A = 0 \vee B = 0) = 1 - P(A = 1, B = 1) = 0.9$$

Big-O Notation [20 pts] (HZ)

For each pair (f, g) of functions below, list which of the following are true: $f(n) = O(g(n))$, $g(n) = O(f(n))$, or both. Briefly justify your answers.

1. $f(n) = 2^n$, $g(n) = e^n$.
Solution: We have $f(n) \in O(g(n))$ because $2^n \leq e^n, \forall n$. On the other direction, there does not exist $c > 0$ such that $2^n \geq ce^n$ for sufficiently large n as $\left(\frac{e}{2}\right)^n$ is unbounded from above as $n \to \infty$.

2. $f(n) = n^2$, $g(n) = n^4 + 2n + 3$.
Solution: We have $f(n) \in O(g(n))$ because $n^2 \leq n^4 + 2n + 3, \forall n$. Similarly, the other direction does not hold since $\lim_{n\to\infty} \frac{g(n)}{f(n)} = \infty$.

3. $f(n) = n$, $g(n) = \log_{10} n$.
Solution: We have $g(n) \in O(f(n))$ because $\log_{10} n \leq n, \forall n$. The other direction does not hold since $\lim_{n\to\infty} \frac{f(n)}{g(n)} = \infty$.

Medium Background Test [20 pts]

Algorithm [5 pts] (WH)

Divide and Conquer: Assume that you are given a sorted array with n numbers ranging from $(-\infty, +\infty)$. Additionally, you are told that somewhere in the array is the number 0. Provide an algorithm to locate the 0 which runs in $O(\log(n))$. Explain your algorithm in words, describe why the algorithm is correct, and justify its running time.
Solution:
Classic binary search would solve this problem. An example implementation is:

1. Start value = 0, low = 0, high = n

2. BinarySearch(array, value, low, high) {
```
mid = (low + high) / 2
if (array[mid] > value): return BinarySearch(A, value, low, mid-1)
else if (array[mid] < value): return BinarySearch(array, value, mid+1, high)
else: return mid
}
```

The algorithm is correct because the array is known to be sorted. Thus by moving in the direction that is closer to zero than the current point, the algorithm ensures that it will always converge to zero while continually shirking the search space

The algorithm runs in $O(\log(n))$ because at each iteration it bisects the remaining possible search space. Thus the maximum number of times the algorithm can run is $\log_2 n$

Probability and Random Variables [5 pts] (WH)

Probability

State true or false. Here Ω denotes the sample space and A^c denotes the complement of the event A.

1. Assume $P(B) > 0$, then $P(A|B) = \dfrac{P(B|A)P(A)}{P(B)}$.
   Solution: True.

2. For any $A, B \subseteq \Omega$ such that $P(B) > 0, P(A^c) > 0$, $P(A|B) + P(B|A^c) = 1$.
   Solution: False.

3. For any $A, B \subseteq \Omega$ such that $0 < P(B) < 1$, $P(A|B) + P(A|B^c) = 1$.
   Solution: False.

4. For any $A, B \subseteq \Omega$, $P(B^c \cup (A \cup B)) + P(B \cap (A \cup A^c)) = 1$.
   Solution: False.

5. Let $\{A_i\}_{i=1}^n$ be mutually independent. Then, $P(\bigcap_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$
   Solution: False.

Discrete and Continuous Distributions

Write down the formula of the probabilistic density/mass functions of random variable X.

1. 1d Gaussian distribution. $X \sim \mathcal{N}(x; \mu, \sigma^2)$.
Solution: $\frac{1}{\sigma\sqrt{2\pi}} exp(\frac{-(x-\mu)^2}{2\sigma^2})$

2. Bernoulli distribution. $X \sim \text{Bernoulli}(p), 0 < p < 1$.
Solution: $p^x(1-p)^{1-x}$

3. Uniform distribution. $X \sim \text{Unif}(a, b), a < b$.
Solution: $\frac{I(a<x<b)}{b-a}$

4. Exponential distribution. $X \sim \text{Exp}(\lambda), \lambda > 0$.
Solution: $exp(-\frac{x}{\lambda})/\lambda$

5. Poisson distribution. $X \sim \text{Poisson}(\lambda), \lambda > 0$.
Solution: $\dfrac{\lambda^x e^{-\lambda}}{x!}$

Mean and Variance

1. Given the following Poisson distribution,

$$p(x; \lambda) = \frac{\frac{\lambda}{3}^x e^{-\lambda/3}}{x!}, \quad \lambda > 0 \tag{1}$$

   (a) What is the mean of the distribution?
   Solution: $\frac{\lambda}{3}$

   (b) What is the variance of the distribution?
   Solution: $\frac{\lambda}{3}$

2. Let X be a random variable and $\mathbb{E}[X] = 1, \text{Var}(X) = 1$. Compute the following values:

   (a) $\mathbb{E}[2X]$.
   Solution: 2

   (b) $\text{Var}(2X)$ and $\text{Var}(X + 1)$.
   Solution: 4 and 1

Mutual and Conditional Independence

1. If X and Y are independent random variables, show that $\text{Var}(X + Y) = \text{Var}X + \text{Var}Y$. (Hint: $\text{Var}(X + Y) = \text{Var}X + 2\text{Cov}(X, Y) + \text{Var}Y$)
Solution:
Since X and Y are independent their covariance is zero.
$\text{Var}(X + Y) = \text{Var}X + 2\text{Cov}(X, Y) + \text{Var}Y$
$\text{Var}(X + Y) = \text{Var}X + 0 + \text{Var}Y$
$\text{Var}(X + Y) = \text{Var}X + \text{Var}Y$

2. Consider rolling two fair, six-sided dice.

(a) If you observe the first die to be 1, what do you know about the second die?
Solution: Nothing.

(b) If you observe the first die to be 1, and then your friend Friedrich tells you the sum of the two dice is even, is the result of the second die independent of the first die?
Solution: Yes.

Law of Large Numbers and the Central Limit Theorem

Provide one line justification.

1. If a fair die is rolled 60000 times, the number of times 1 shows up is close to 10000.
Solution:
The Law of Large Numbers states that the sample average, $\bar{X}$ converges in probability to the expectation $\mu = E[X]$. In this case, $\mu = 1/6$ and the sample average should be near $60000 \times 1/6 = 10000$.

2. Let $X_i \sim \mathcal{N}(0,1)$ and $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, then the distribution of $\bar{X}$ satisfies

$$\sqrt{n}\bar{X} \overset{n\to\infty}{\longrightarrow} \mathcal{N}(0,1)$$

Some useful background material: http://www.cs.cmu.edu/~aarti/Class/10701/recitation/prob_review.pdf
Solution:
The Central Limit Theorem states that is X_i are iid generated from a distribution with mean μ and variance σ then, $\dfrac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \rightsquigarrow N(0,1)$. Since in our case $\mu = 0$ and $\sigma = 1$ the result follows.

Linear Algebra [5 pts] (HZ)

Vector Norms

Draw the regions corresponding to vectors $\mathbf{x} \in \mathbb{R}^2$ with the following norms:

1. $||\mathbf{x}||_1 \leq 1$ (Recall that $||\mathbf{x}||_1 = \sum_i |x_i|$)
2. $||\mathbf{x}||_2 \leq 1$ (Recall that $||\mathbf{x}||_2 = \sqrt{\sum_i x_i^2}$)
3. $||\mathbf{x}||_\infty \leq 1$ (Recall that $||\mathbf{x}||_\infty = \max_i |x_i|$)

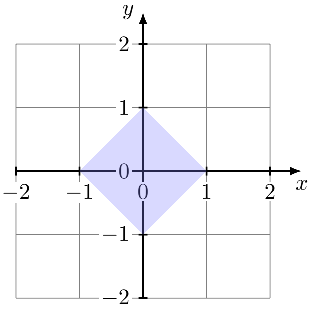
Solution: The three unit balls under the corresponding distance metrics are shown as follows:

Geometry

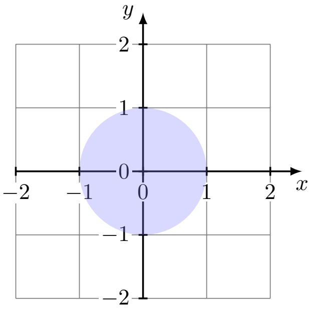
Prove that these are true or false. Provide all steps!

1. The Euclidean distance from the origin to the hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$ is $\frac{|b|}{||\mathbf{w}||_2}$.

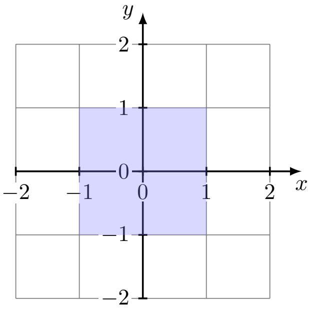
Proof. Let $\mathbf{x}$ be any point on the hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$. The Euclidean distance from the origin to the hyperplane can be computed by projecting $\mathbf{x}$ onto the normal vector of the hyerplane, which is given by $\mathbf{w}$. Hence the distance is given by $\frac{|\mathbf{w}^T\mathbf{x}|}{||\mathbf{w}||_2} = \frac{|-b|}{||\mathbf{w}||_2} = \frac{|b|}{||\mathbf{w}||_2}$, which completes the proof. ■



(a) Unit ℓ_1 ball.



(b) Unit ℓ_2 ball.



(c) Unit ℓ_∞ ball.

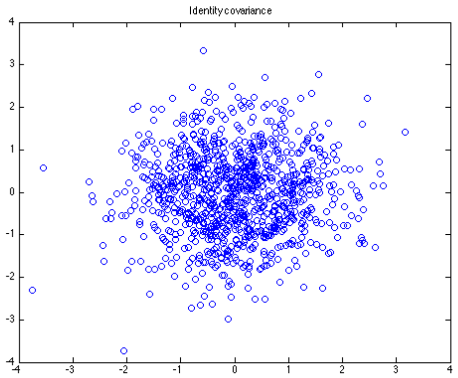
2. The Euclidean distance between two parallel hyperplane $\mathbf{w}^T\mathbf{x} + b_1 = 0$ and $\mathbf{w}^T\mathbf{x} + b_2 = 0$ is $\frac{|b_1 - b_2|}{||\mathbf{w}||_2}$ (Hint: you can use the result from the last question to help you prove this one).

   Proof. Without loss of generality we can translate both hyperplanes along the y-axis such that one of them go through the origin. The equations for the translated hyperplanes are given by $\mathbf{w}^T\mathbf{x} = 0$ and $\mathbf{w}^T\mathbf{x} + b_2 - b_1 = 0$. Note that since we translate both hyperplanes along y-axis by the same distance, i.e., b_1, the distance between these two hyperplanes are kept. Using the result from the last question we know that the Euclidean distance from the origin to $\mathbf{w}^T\mathbf{x} + b_2 - b_1 = 0$ is given by $\frac{|b_2 - b_1|}{||\mathbf{w}||_2}$. Further note that the origin is on the hyperplane $\mathbf{w}^T\mathbf{x} = 0$. We reach the conclusion that the distance between these two hyperplanes is given by $\frac{|b_2 - b_1|}{||\mathbf{w}||_2}$. ■

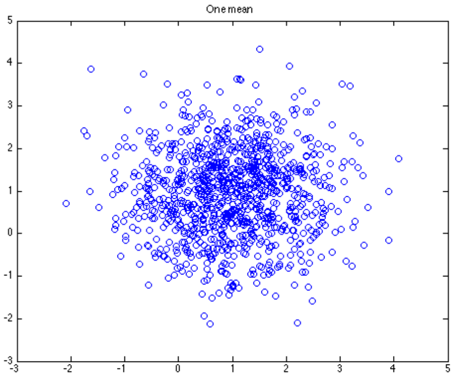
Programming Skills - Matlab [5pts] (WH)

Sampling from a distribution.

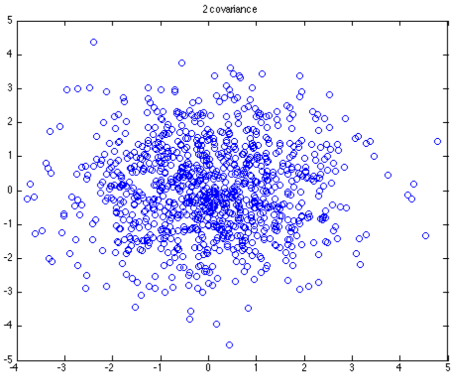
1. Draw 1000 samples $\mathbf{x} = [x_1, x_2]^T$ from a 2-dimensional Gaussian distribution with mean $(0,0)^T$ and identity covariance matrix, i.e., $p(\mathbf{x}) = \frac{1}{2\pi} \exp\left(-\frac{||\mathbf{x}||^2}{2}\right)$, and make a scatter plot (x_1 vs. x_2). Submit your plot.
   Solution:
   x = mvnrnd([0,0], [1,0 ; 0,1], 1000); figure; plot(x(:,1),x(:,2),'o'); title('Identity covariance');



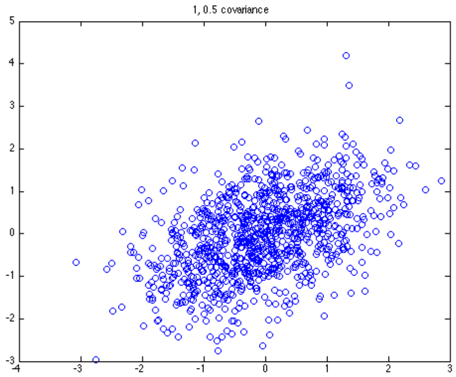
2. How does the scatter plot change if the mean is $(1, 1)^T$ (and identity variance)? Submit your plot.
Solution:
x = mvnrnd([1,1], [1,0 ; 0,1], 1000); figure; plot(x(:,1),x(:,2),'o'); title('One mean');
Shifts up and to the left by 1.



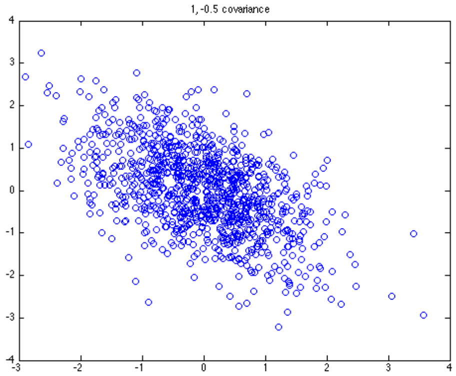
3. How does the scatter plot change if the covariance matrix is $\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$ (and zero mean)? Submit your plot.
Solution:
x = mvnrnd([0,0], [2,0 ; 0,2], 1000); figure; plot(x(:,1),x(:,2),'o'); title('2 covariance');
More spread out.



4. How does the scatter plot change if the covariance matrix is changed to $\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$ (and zero mean)? Submit your plot.
Solution:
x = mvnrnd([0,0], [1,0.5 ; 0.5,1], 1000); figure; plot(x(:,1),x(:,2),'o'); title('1, 0.5 covariance');
skewed so that they stretch from the lower left to the upper right.



5. How does the scatter plot change if the covariance matrix is changed to $\begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$ (and zero mean)? Submit your plot.
Solution: x = mvnrnd([0,0], [1,-0.5 ; -0.5,1], 1000); figure; plot(x(:,1),x(:,2),'o'); title('1, -0.5 covariance');
skewed so that they stretch from the lower right to the upper left.


1, -0.5 covariance