Homework 1: Background Test

10-601 Machine Learning
Due 5 p.m. Friday, January 16, 2015

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 don't 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.

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`

0 Instructions

- **Submit your homework** by dropping off a hardcopy in the bin outside Gates 8215 by 5 p.m. Friday, January 16, 2015.
- **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.
- **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 Points]

Vectors and Matrices [20 Points]

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

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

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 $\mathbf{y^T z}$)
   Solution:

$$\mathbf{y}^\top \mathbf{z} = (1 \times 2) + (3 \times 3) = 11 \tag{1}$$

2. What is the product $\mathbf{Xy}$?
   Solution:

$$\mathbf{Xy} = \begin{bmatrix} (2 \times 1) + (4 \times 3) \\ (1 \times 1) + (3 \times 3) \end{bmatrix} = \begin{bmatrix} 14 \\ 10 \end{bmatrix} \tag{2}$$

3. Is $\mathbf{X}$ invertible? If so, give the inverse, and if no, explain why not.
   Solution: Yes.

$$\mathbf{X}^{-1} = \frac{1}{(2 \times 3) - (1 \times 4)} \begin{bmatrix} 3 & -4 \\ -1 & 2 \end{bmatrix} = \begin{bmatrix} \frac{3}{2} & -2 \\ -\frac{1}{2} & 1 \end{bmatrix} \tag{3}$$

4. What is the rank of $\mathbf{X}$?
   Solution: The rank of $\mathbf{X}$ is 2, since the column rank is 2, since $\begin{bmatrix} 4 \\ 3 \end{bmatrix} \neq c \begin{bmatrix} 2 \\ 1 \end{bmatrix}$ for all $c \in \mathbb{R}$.

Calculus [20 Points]

1. If $y = x^3 + x - 5$ then what is the derivative of y with respect to x?

   Solution:

$$\frac{dy}{dx} = 3x^2 + 1 \tag{4}$$

2. If $y = x\sin(z)e^{-x}$ then what is the partial derivative of y with respect to x?

Solution:

$$\begin{aligned} \frac{dy}{dx} &= \sin(z)e^{-x} - x(\sin(z)e^{-x}) && (5) \\ &= (1-x)\sin(z)e^{-x} && (6) \end{aligned}$$

Probability and Statistics [20 Points]

Consider a sample of data $S = \{1, 1, 0, 1, 0\}$ 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:

$$\text{Sample mean} = \frac{1+1+0+1+0}{5} = \frac{3}{5} \tag{7}$$

2. What is the sample variance for this data?

Solution:

$$\begin{aligned} \text{Sample variance} &= \frac{1}{5}\left[\left(\frac{2}{5}\right)^2 + \left(\frac{2}{5}\right)^2 + \left(-\frac{3}{5}\right)^2 + \left(\frac{2}{5}\right)^2 + \left(-\frac{3}{5}\right)^2\right] && (8) \\ &= \frac{1}{5}\left[\frac{3\times 4}{25} + \frac{2\times 9}{25}\right] = \frac{6}{25} && (9) \end{aligned}$$

3. What is the probability of observing this data, assuming it was generated by flipping a coin with an equal probability of heads and tails (i.e. the probability distribution is $p(x = 1) = 0.5$, $p(x = 0) = 0.5$).

Solution:

$$\text{Probability of } S = 0.5^5 = \frac{1}{32} \tag{10}$$

4. Note that the probability of this data sample would be greater if the value of $p(x = 1)$ was not 0.5, but instead some other value. What is the value that maximizes the probability of the sample S. Please justify your answer.

   Solution: Let p be the probability of 1 (i.e. $p(x = 1)$). We want to find the value of the p that maximizes the probability of the sample S. Note that the probably of the sample S can be written

$$\prod_{i=1}^{5} p^{x_i}(1-p)^{(1-x_i)} = p^{\sum_{i=1}^{5} x_i}(1-p)^{n-\sum_{i=1}^{5} x_i}. \tag{11}$$

   We want to maximize the above as a function of p. We first take the log of the above and call this function $\ell(p)$, which we can write as

$$\ell(p) = \left(\sum_{i=1}^{5} x_i\right)\log(p) + \left(n - \sum_{i=1}^{5} x_i\right)\log(1-p) \tag{12}$$

   To find the p that maximizes the probability of $p(x = 1)$, we can find the p that maximizes the probability of $\ell(p)$. To do this, we take the derivative of $\ell(p)$ with respect to p, set this to zero, and solve for p:

$$\frac{d\ell(p)}{dp} = \frac{1}{p}\sum_{i=1}^{5} x_i - \frac{1}{1-p}\left(n - \sum_{i=1}^{5} x_i\right) = 0 \tag{13}$$

$$\implies 0 = \frac{\sum_{i=1}^{5} x_i - pn}{p(1-p)} \tag{14}$$

$$\implies pn = \sum_{i=1}^{5} x_i \tag{15}$$

$$\implies p = \frac{1}{n}\sum_{i=1}^{5} x_i \tag{16}$$

   Plugging in our values for $x_1, \ldots, x_5$ into the above formula, we find that the best $p = \frac{3}{5}$.

5. Consider the following joint probability table over variables y and z, where y takes a value from the set {a,b,c}, and z takes a value from the set {T,F}:

		y		
		a	b	c
z	T	0.2	0.1	0.2
	F	0.05	0.15	0.3

   - What is $p(\text{z} = \text{T} \ \ \text{AND} \ \ \text{y} = \text{b})$?
     Solution: The answer is 0.1.

   - What is $p(\text{z} = \text{T}|\text{y} = \text{b})$?
     Solution: Using the definition of conditional probability, we see that

$$p(\mathsf{z} = \mathsf{T}|\mathsf{y} = \mathsf{b}) = \frac{p(\mathsf{z} = \mathsf{T} \ \ \mathsf{AND} \ \ \mathsf{y} = \mathsf{b})}{p(\mathsf{y} = \mathsf{b})} = \frac{0.1}{0.1 + 0.15} = 0.4 \tag{17}$$

Big-O Notation [20 Points]

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. $f(n) = \ln(n)$, $g(n) = \lg(n)$. Note that ln denotes log to the base e and lg denotes log to the base 2.
   Solution: Both, since the functions are equivalent up to a multiplicative constant.

2. $f(n) = 3^n$, $g(n) = n^{10}$
   Solution: $g(n) = O(f(n))$, since $f(n)$ grows much more rapidly as n becomes large.

3. $f(n) = 3^n$, $g(n) = 2^n$
   Solution: $g(n) = O(f(n))$, since $f(n)$ grows much more rapidly as n becomes large.

Medium Background Test [20 Points]

Algorithms [5 Points]

Divide and Conquer: Assume that you are given an array with n elements all entries equal either to 0 or $+1$ such that all 0 entries appear before $+1$ entries. You need to find the index where the transition happens, i.e. you need to report the index with the last occurrence of 0. Give an algorithm that runs in time $O(\log n)$. Explain your algorithm in words, describe why the algorithm is correct, and justify its running time.

Solution:

We give an algorithm below, called find-transition, for this problem:

```
find-transition (i , j)

    let mid = i + floor ( (j-i) /2 )

    let a = array [ mid ] and b = array [ mid + 1 ]

    if (a == b == 1) return find-transition ( i , mid )

    else if (a == b == 0) return find-transition ( mid + 1 , j )

    else /* a == 0 and b == 1 */

    return mid (last occurrence of 0)
```

This algorithm returns the correct results: note that for each recursive call we know that array[i]==0 and array[j]==1. When we stop, we know that a==0 and b==1, so we array[mid] is the last entry with 0.

The running time can be analyzed via the recurrence: $T(n) = T(n/2) + O(1)$, $T(n) = c$ for $n \leq 4$ which solves to $O(\log n)$. This algorithm is based on the idea of binary search and hence the running time is as expected.

Probability and Random Variables [5 Points]

Probability

State true or false. Here A^c denotes complement of the event A.

(a) $P(A \cup B) = P(A \cap (B \cap A^c))$

(b) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

(c) $P(A) = P(A \cap B) + P(A^c \cap B)$
(d) $P(A|B) = P(B|A)$
(e) $P(A_1 \cap A_2 \cap A_3) = P(A_3|(A_2 \cap A_1))P(A_2|A_1)P(A_1)$

Solutions: (a) False, (b) True, (c) False, (d) False, (e) True

Discrete and Continuous Distributions
Match the distribution name to its probability density function (pdf).

(a) Multivariate Gaussian	(f) $p^x(1-p)^{1-x}$
(b) Exponential	(g) $\frac{1}{b-a}$ when $a \leq x \leq b$; 0 otherwise
(c) Uniform	(h) $\binom{n}{x}p^x(1-p)^{n-x}$
(d) Bernoulli	(i) $\lambda e^{-\lambda x}$ when $x \geq 0$; 0 otherwise
(e) Binomial	(j) $\frac{1}{\sqrt{(2\pi)^d\|\Sigma\|}} \exp\left(-\frac{1}{2} - (\mathbf{x}-\mu)^\top \Sigma^{-1}(\mathbf{x}-\mu)\right)$

Solutions: (a) with (j), (b) with (i), (c) with (g), (d) with (f), (e) with (h)

Mean, Variance and Entropy
(a) What is the mean, variance, and entropy of a Bernoulli(p) random variable?

Solution: The mean is p, the variance is $p(1-p)$, and the entropy is $-(1-p)\log(1-p) - p\log(p)$.

(b) If the variance of a zero-mean random variable x is σ^2, what is the variance of $2x$? What about the variance of $x+2$?

Solution: The variance of $2x$ is $4\sigma^2$, while the variance of $x+2$ remains σ^2.

Mutual and Conditional Independence
(a) If X and Y are independent random variables, show that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

Solution: Using the definition of expectations,

$$\mathbb{E}[XY] = \int xyp(x,y)dxdy = \int xp(x)yp(y)dxdy = \int xp(x)dx \int yp(y)dy = \mathbb{E}[X]\mathbb{E}[Y]$$

where $p(x,y)$ is the joint pdf of X and Y, $p(x)$ is the marginal pdf of X and $p(y)$ is the marginal pdf of Y.

(b) Alice rolls a die and calls up Bob and Chad to tell them the outcome A. Due to disturbance in the phones, Bob and Chad think the roll was B and C, respectively. Is B independent of C? Is B independent of C given A?

Solution: B is not independent of C, but it is independent of C given A.

Law of Large Numbers and Central Limit Theorem
Provide one line justifications.
(a) If a die is rolled 6000 times, the number of times 3 shows up is close to 1000.

Solution: Assuming a fair die, the number of times 3 shows up should be close to 1000 due to the Law of Large Numbers.

(b) If a fair coin is tossed n times and $\bar{X}$ denotes the average number of heads, then distribution of $\bar{X}$ satisfies

$$\sqrt{n}(\bar{X} - 1/2)] \overset{n\rightarrow\infty}{\rightarrow} \mathcal{N}(0, 1/4)$$

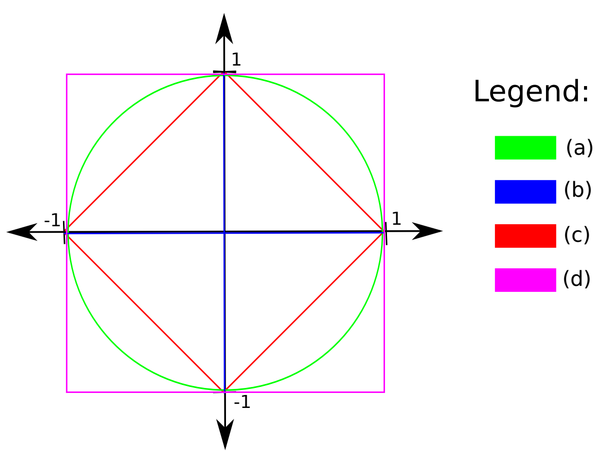
Solution: The expression on the left-hand-side should tend to the expression on the right-hand-side as $n \to \infty$ due to the Central Limit Theorem.

Some useful background reading material:
http://www.cs.cmu.edu/∼aarti/Class/10701/recitation/prob_review.pdf

Linear Algebra [5 Points]

Vector norms
Draw the regions corresponding to vectors $\mathbf{x} \in \mathbb{R}^2$ with following norms:
(a) $\|\mathbf{x}\|_2 \leq 1$ (Recall $\|x\|_2 = \sqrt{\sum_i x_i^2}$)
(b) $\|\mathbf{x}\|_0 \leq 1$ (Recall $\|x\|_0 = \sum_{i:x_i \neq 0} 1$)
(c) $\|\mathbf{x}\|_1 \leq 1$ (Recall $\|x\|_1 = \sum_i |x_i|$)
(d) $\|\mathbf{x}\|_\infty \leq 1$ (Recall $\|x\|_\infty = \max_i |x_i|$)

Solution:



Geometry
(a) Show that the vector $\mathbf{w}$ is orthogonal to the line $\mathbf{w}^\top \mathbf{x} + b = 0$. *(Hint: Consider two points*

$\mathbf{x}_1, \mathbf{x}_2$ *that lie on the line. What is the inner product* $\mathbf{w}^\top(\mathbf{x}_1 - \mathbf{x}_2)$*?)*

Solution: This line is all $\mathbf{x}$ such that $\mathbf{w}^\top \mathbf{x} + b = 0$. Consider two such $\mathbf{x}$, called $\mathbf{x}_1$ and $\mathbf{x}_2$. Note that $\mathbf{x}_1 - \mathbf{x}_2$ is a vector parallel to our line. Also note that

$$\mathbf{w}^\top \mathbf{x}_1 + b = 0 = \mathbf{w}^\top \mathbf{x}_2 + b \implies \mathbf{w}^\top \mathbf{x}_1 = \mathbf{w}^\top \mathbf{x}_2 \implies \mathbf{w}^\top(\mathbf{x}_1 - \mathbf{x}_2) = 0 \tag{18}$$

which shows that the vector $\mathbf{w}$ is orthogonal to our line.

(b) Argue that the distance from the origin to the line $\mathbf{w}^\top \mathbf{x} + b = 0$ is $\frac{b}{\|\mathbf{w}\|}$.

Solution: We can show this by first finding the closest point to the origin that lies on this line, and then finding the distance to this point. Let $\mathbf{a}^*$ be the closest point to the origin the lies on the line. We can write $\mathbf{a}^*$ as

$$\mathbf{a}^* = \min \mathbf{a}^\top \mathbf{a} \tag{19}$$

$$\text{s.t. } \mathbf{w}^\top \mathbf{a} + b = 0 \tag{20}$$

So we first solve this constrained optimization problem and find $\mathbf{a}^*$. We start by taking the derivative of the objective, setting it to zero, and using Lagrange multipliers (i.e. setting the derivative of the Lagrangian $\mathbf{a}^\top \mathbf{a} - \lambda(\mathbf{w}^\top a + b)$ to zero). We can write

$$2\mathbf{a}^* = \lambda \mathbf{w} \implies \mathbf{a}^* = \frac{\lambda}{2}\mathbf{w} \tag{21}$$

Hence, plugging this value for $\mathbf{a}$ into the constraint, we can write:

$$\mathbf{w}^\top \mathbf{a} + b = 0 \tag{22}$$

$$\implies \mathbf{w}^\top \left(\frac{\lambda}{2}\mathbf{w}\right) + b = 0 \tag{23}$$

$$\implies \frac{\lambda}{2}\mathbf{w}^\top \mathbf{w} + b = 0 \tag{24}$$

$$\implies \lambda = \frac{-2b}{\mathbf{w}^\top \mathbf{w}} \tag{25}$$

$$\implies \mathbf{a}^* = \frac{-b}{\mathbf{w}^\top \mathbf{w}}\mathbf{w} \tag{26}$$

once we have $\mathbf{a}*$, we can compute the distance between $\mathbf{a}^*$ and the origin to get

$$\text{distance} = \sqrt{(\mathbf{a}^*)^\top \mathbf{a}^*} = \sqrt{\left(\frac{-b}{\mathbf{w}^\top \mathbf{w}}\right)^2 \mathbf{w}^\top \mathbf{w}} \tag{27}$$

$$= \frac{b}{\mathbf{w}^\top \mathbf{w}}\sqrt{\mathbf{w}^\top \mathbf{w}} = \frac{b}{\sqrt{\mathbf{w}^\top \mathbf{w}}} = \frac{b}{\|\mathbf{w}\|} \tag{28}$$

Some useful background reading material:
http://www.cs.cmu.edu/∼aarti/Class/10701/recitation/LinearAlgebra_Matlab_Review.ppt
http://www.cs.cmu.edu/∼zkolter/course/15-884/linalg-review.pdf
Wikipedia: http://en.wikipedia.org/wiki/Eigenvalues_and_eigenvectors
Gilbert Strang. Linear Algebra and its Applications, Ch 5. HBJ Publishers.

Programming skills - MATLAB/R/C [5 Points]

Sampling from a distribution

(a) Draw 100 samples $\mathbf{x} = [x_1 \; x_2]$ from a 2-dimensional Gaussian distribution with mean $[0, 0]$ and identity covariance matrix i.e. $p(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^d}} \exp\left(-\frac{\|\mathbf{x}\|^2}{2}\right)$, and make a scatter plot (x_1 vs. x_2).
(b) How does the scatter plot change if the mean is $[-1, 1]$?
(c) How does the scatter plot change if you double the variance of each component?
(d) How does the scatter plot change if the covariance matrix is changed to the following?

$$\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

(e) How does the scatter plot change if the covariance matrix is changed to the following?

$$\begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$$

Solutions:
(a) See attached code.
(b) See attached code. The data moves up and to the left. Namely, the center of the data moves from roughly $[0, 0]$ to roughly $[-1, 1]$.
(c) See attached code. The data become more "spread out".
(d) See attached code. The data become skewed so that they stretch from the lower left to the upper right.
(d) See attached code. The data become skewed so that they stretch from the upper left to the bottom right.

Some useful background reading material:
Matlab tutorial - http://www.math.mtu.edu/~msgocken/intro/intro.pdf
R tutorial - http://math.illinoisstate.edu/dhkim/rstuff/rtutor.html