CS4786/5786: Machine Learning for Data Science, Spring 2015
1/22/2015: Assignment 0: Diagnostic

---

We want you to see feedback we'll provide regarding your level of preparation for the class, but we don't want you to stress out about being "judged". So, for this assignment, which should take *at most* two hours, we'll give credit on an all-or-none basis, as follows. If you turn the assignment in on time with good-faith attempts to show how you approached each question (*whether or not* you arrive at a correct answer or even get to an answer at all), **and** pick up the hardcopy that we'll have written feedback on,[1] you'll get full credit. That's it!

Given that this is a diagnostic assignment, you should work *individually*. Don't discuss the problem set with other people. You may consult textbooks and other offline or online sources, but you should mention them and state what portions of your work they influenced.

Turn in your hardcopy answers at the beginning of class on Thursday, Jan 29th, 2015. At the top, write down your last/family name, then your first name, and then your Cornell NetId in all caps (e.g., LJL2).[2]

**Academic integrity policy** We distinguish between "merely" violating the rules for a given assignment and violating *academic integrity*. To violate the latter is to commit *fraud* by claiming credit for someone else's work. For this assignment, an example of the former would be getting an answer from person X but stating in your homework that X was the source of that particular answer. You would cross the line into fraud if you did not mention X. The worst-case outcome for the former is a grade penalty; the worst-case scenario in the latter is academic-integrity hearing procedures.

The way to avoid violating academic integrity is to always document any portions of work you submit that are due to or influenced by other sources, even if those sources weren't permitted by the rules.[3]

---

*Q1 (Linear Algebra).* Let $\overrightarrow{x}$ and $\overrightarrow{y}$ be two non-zero $n$-dimensional vectors that are orthogonal (i.e., $\overrightarrow{x} \cdot \overrightarrow{y} = 0$). Does there exist an $n \times n$ matrix $A$ such that the vectors $A\overrightarrow{x}$ and $A\overrightarrow{y}$ are **not** orthogonal? Prove your answer, showing all work.

Incidentally, this question relates to *dimensionality reduction*, the first topic of the class.

*Q2 (Partial Derivatives).* Let $f(x_1, x_2, \ldots, x_n) = -\sum_{i=1}^{n} x_i \log(x_i)$. Compute the second partial derivatives $\frac{f(x_1,x_2,\ldots,x_n)}{\partial x_i \partial x_j}$ for arbitrary $i, j \in 1, \ldots, n$. Show your work. (We're expecting that you might come up with something like "if $i \neq j$, then $3\pi x_i x_j$; otherwise, $e^{x_i}$".)

[1] Location and timeframe will be announced later.

[2] The reason for capital letters is because "ljl2" could look like either ell-jay-ell-two or ell-jay-twelve. We speak from experience.

[3] We make an exception for sources that can be taken for granted in the instructional setting, namely, the course materials. To minimize documentation effort, we also do not expect you to credit the course staff for ideas you get from them, although it's nice to do so anyway.

(PROBLEM(S) CONTINUE ON NEXT PAGE)

Incidentally, it's worth recalling that the matrix of second partial derivatives is called the *Hessian*, and the vector of first-order partial derivatives is called the *gradient*, both of which we'll see later in the semester.

*Q3 (Conditional Probability[4]).* Ms. Y lives next to a house with a big lawn. On any given day, the probability that it rains in Ms. Y's neighborhood is $0.1$. The probability that the sprinkler is turned on next door is $0.3$. When it rains, the probability that Ms. Y wears a poncho is $0.6$. When the sprinkler is on, the probability that Ms. Y wears a poncho is $0.2$. She never wears a poncho unless it either rained or the sprinkler was on.

You meet Ms. Y wearing a poncho. Is it more likely that the sprinkler was turned on that day or that it rained in Ms. Y's neighborhood that day? Show your work.

*Q4 (Linearity of Expectation).* In a group of $100$ sports fan, $90$ are football fans and $10$ are baseball fans. You have with you a box containing $90$ football jerseys and $10$ baseball jerseys. You distribute one jersey to each fan from this box in a random order. What is the expected number of baseball fans who get baseball jerseys? Show your work.

Note: you could imagine approaching this problem by first trying to work out the all the possible outcomes and how many baseball fans get baseball jerseys in each one. But instead, we want you to take a much simpler approach based on linearity of expectation. Start by defining an *indicator random variable* for each baseball fan, where the variable has value either 1 or 0 depending on whether or not that fan got a baseball jersey.

[4]Problem inspired by a classic example regarding Bayes Nets. (You don't need to know what a Bayes Net is.)