CS 5350/6350: Machine Learning Spring 2019

Homework 0

Handed out: 7 January, 2019
Due: 11:59pm, 16 January, 2019

- You are welcome to talk to other members of the class about the homework. I am more concerned that you understand the underlying concepts. However, you should write down your own solution. Please keep the class collaboration policy in mind.
- Feel free discuss the homework with the instructor or the TAs.
- Your written solutions should be brief and clear. You need to show your work, not just the final answer, but you do *not* need to write it in gory detail. Your assignment should be **no more than 10 pages**. Every extra page will cost a point.
- Handwritten solutions will not be accepted.
- The homework is due by **midnight of the due date**. Please submit the homework on Canvas.
- Some questions are marked **For 6350 students**. Students who are registered for CS 6350 should do these questions. Of course, if you are registered for CS 5350, you are welcome to do the question too, but you will not get any credit for it.

Basic Knowledge Review

1. [5 points] We use sets to represent events. For example, toss a fair coin 10 times, and the event can be represented by the set of "Heads" or "Tails" after each tossing. Let a specific event A be "at least one head". Calculate the probability that event A happens, i.e., $p(A)$.
   $p(A) = 1 - \binom{10}{0}(.5)^0(.5)^{10} = 1 - 0.0009765625 = 0.9990234375$
2. [10 points] Given two events A and B, prove that
   $$p(A \cup B) \leq p(A) + p(B).$$
   When does the equality hold?
   $p(A \cup B) = p(A) + p(B \cap A')$
   $= p(A) + [p(B) - p(A \cap B)]$
   $= p(A) + p(B) - p(A \cap B)$
   and $p(A \cap B) \geq 0$
   Therefore $p(A \cup B) \leq p(A) + p(B)$
   This is true for any two events that satisfy the basic properties of probability.
3. [10 points] Let $\{A_1, \ldots, A_n\}$ be a collection of events. Show that
   $$p(\cup_{i=1}^n A_i) \leq \sum_{i=1}^n p(A_i).$$

When does the equality hold? (Hint: induction)
$p(\cup_{i=1}^{n} A_i) = p(A_1) + ... + p(A_n) - p(A_1 \cap A_2) - ... - p(\cap_{i=1}^{n} A_i)$
$p(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} p(A_i) - p(A_1 \cap A_2) - ... - p(\cap_{i=1}^{n} A_i)$
Therefore $p(\cup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} p(A_i)$.
This equality will also hold for all events that satisfy the three axioms of probability.

4. [20 points] We use $\mathbb{E}(\cdot)$ and $\mathbb{V}(\cdot)$ to denote a random variable's mean (or expectation) and variance, respectively. Given two discrete random variables X and Y, where $X \in \{0, 1\}$ and $Y \in \{0, 1\}$. The joint probability $p(X, Y)$ is given in as follows:

	$Y = 0$	$Y = 1$
$X = 0$	1/10	2/10
$X = 1$	3/10	4/10

(a) [10 points] Calculate the following distributions and statistics.

i. the the marginal distributions $p(X)$ and $p(Y)$

$$p_X(x) = \begin{cases} 3/10 & x = 0 \\ 7/10 & x = 1 \\ 0 & otherwise \end{cases}$$

$$p_Y(y) = \begin{cases} 4/10 & y = 0 \\ 6/10 & y = 1 \\ 0 & otherwise \end{cases}$$

ii. the conditional distributions $p(X|Y)$ and $p(Y|X)$
$P(A|B) = \frac{P(A \cap B)}{P(B)}$

$$p(X|Y) = \begin{cases} 1/4 & x = 0, y = 0 \\ 3/4 & x = 1, y = 0 \\ 1/3 & x = 0, y = 1 \\ 2/3 & x = 1, y = 1 \\ 0 & otherwise \end{cases}$$

$$p(Y|X) = \begin{cases} 1/3 & x = 0, y = 0 \\ 2/3 & x = 1, y = 0 \\ 2/3 & x = 0, y = 1 \\ 1/3 & x = 1, y = 1 \\ 0 & otherwise \end{cases}$$

iii. $\mathbb{E}(X)$, $\mathbb{E}(Y)$, $\mathbb{V}(X)$, $\mathbb{V}(Y)$
$\mathbb{E}(X) = (0 * 3/10) + (1 * 7/10) = 7/10$
$\mathbb{E}(Y) = (0 * 4/10) + (1 * 6/10) = 6/10$
$\mathbb{V}(X) = (0^2 * 3/10) + (1^2 * 7/10) - (7/10)^2 = .21$
$\mathbb{V}(Y) = (0^2 * 4/10) + (1^2 * 6/10) - (6/10)^2 = .24$

iv. $\mathbb{E}(Y|X = 0)$, $\mathbb{E}(Y|X = 1)$, $\mathbb{V}(Y|X = 0)$, $\mathbb{V}(Y|X = 1)$
$\mathbb{E}(Y|X = 0) = (0 * 1/3) + (1 * 2/3) = 2/3$
$\mathbb{E}(Y|X = 1) = (0 * 2/3) + (1 * 1/3) = 1/3$
$\mathbb{V}(Y|X = 0) = \mathbb{E}((Y - \mathbb{E}(Y|X = 0))^2|X = 0) = \mathbb{E}((Y - 2/3)^2|X = 0) = (4/9 * 1/3) + (1/9 * 2/3) = 0.074$

$\mathbb{V}(Y|X=1) = \mathbb{E}((Y-\mathbb{E}(Y|X=1))^2|X=1) = \mathbb{E}((Y-1/3)^2|X=1) = (1/9*1/3)+(4/9*2/3) = 0.296$

v. the covariance between X and Y
$Cov(X,Y) = (-7/10*-6/10*1/10) + (-7/10*4/10*2/10) + (3/10*-6/10*3/10) + (3/10*4/10*4/10) = -0.02$

(b) [5 points] Are X and Y independent? Why?

X and Y are not independent, this is because $p(x,y) \neq p_X(x) * p_Y(y)$ for all (x,y) pairs. $p(0,0) = 1/10, p_X(0) * p_Y(0) = 3/10 * 4/10 = 0.12$.

(c) [5 points] When X is not assigned a specific value, are $\mathbb{E}(Y|X)$ and $\mathbb{V}(Y|X)$ still constant? Why?

$\mathbb{E}(Y|X)$ is a function of the random variable X, and is therefore a random variable itself. Additionally, $\mathbb{V}(Y|X)$ is a function of $\mathbb{E}(Y|X)$, making it a random variable as well.

5. [10 points] Assume a random variable X follows a standard normal distribution, i.e., $X \sim \mathcal{N}(X|0,1)$. Let $Y = e^X$. Calculate the mean and variance of Y.

(a) $\mathbb{E}(Y)$

$\mathbb{E}(Y) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} * e^x dx = \sqrt{e}$

(b) $\mathbb{V}(Y)$

$\mathbb{V}(Y) = \int_{-\infty}^{\infty} (e^x - \sqrt{e})^2 * (\frac{1}{\sqrt{2\pi}} e^{-x^2/2}) dx = e^2 - e$

6. [20 points] Given two random variables X and Y, show that

(a) $\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y)$

Let $g(x) = \mathbb{E}(Y|X=x)$. Then $\mathbb{E}(\mathbb{E}(Y|X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx$.

Additionally, $g(x) = \int_{-\infty}^{\infty} y * f_{Y|X}(y|x) dy$, or $\int_{-\infty}^{\infty} y * \frac{f(x,y)}{f_X(x)} dy$.

Substituting this into the first integral gives $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y * \frac{f(x,y)}{f_X(x)} * f_X(x) dy dx$,

or $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y * f(x,y) dy dx$, the definition of $\mathbb{E}(Y)$.

(b) $\mathbb{V}(Y) = \mathbb{E}(\mathbb{V}(Y|X)) + \mathbb{V}(\mathbb{E}(Y|X))$

First we start by using the following formula for conditional variance:

$\mathbb{V}(Y|X) = \mathbb{E}(Y^2|X) - [\mathbb{E}(Y|X)]^2$

Taking the expectation of this gives $\mathbb{E}(\mathbb{V}(Y|X)) = \mathbb{E}(\mathbb{E}(Y^2|X)) - \mathbb{E}([\mathbb{E}(Y|X)]^2)$, which can be simplified to $\mathbb{E}(Y^2) - \mathbb{E}([\mathbb{E}(Y|X)]^2)$.

Next we take the variance of conditional expected value, $\mathbb{V}(\mathbb{E}(Y|X)) = \mathbb{E}([\mathbb{E}(Y|X)]^2) - [\mathbb{E}(\mathbb{E}(Y|X))]^2$. Because $\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y)$, this can be simplified to $\mathbb{E}([\mathbb{E}(Y|X)]^2) - [\mathbb{E}(Y)]^2$.

Finally these are combined, giving $\mathbb{E}(\mathbb{V}(Y|X)) + \mathbb{V}(\mathbb{E}(Y|X)) = \mathbb{E}(Y^2) - \mathbb{E}([\mathbb{E}(Y|X)]^2) + \mathbb{E}([\mathbb{E}(Y|X)]^2) - [\mathbb{E}(Y)]^2 = \mathbb{E}(Y^2) - [\mathbb{E}(Y)]^2 = \mathbb{V}(Y)$.

Therefore $\mathbb{V}(Y) = \mathbb{E}(\mathbb{V}(Y|X)) + \mathbb{V}(\mathbb{E}(Y|X))$

(Hints: using definition.)

7. [15 points] Given a logistic function, $f(\mathbf{x}) = 1/(1+\exp(-\mathbf{a}^\top \mathbf{x}))$ ($\mathbf{x}$ is a vector), derive/calculate the following gradients and Hessian matrices.

(a) $\nabla f(\mathbf{x})$

$f(\mathbf{x}) = 1/(1 + \mathbf{e}^{-\mathbf{a}_1\mathbf{x}_1 - \ldots - \mathbf{a}_n\mathbf{x}_n})$

$$\nabla f(\mathbf{x}) = \left\{ \begin{matrix} -\frac{1+\mathbf{e}^{-\mathbf{a}_1\mathbf{x}_1-\ldots-\mathbf{a}_n\mathbf{x}_n}}{(1+\mathbf{e}^{-\mathbf{a}_1\mathbf{x}_1-\ldots-\mathbf{a}_n\mathbf{x}_n})^2} \\ \vdots \\ -\frac{1+\mathbf{e}^{-\mathbf{a}_1\mathbf{x}_1-\ldots-\mathbf{a}_n\mathbf{x}_n}}{(1+\mathbf{e}^{-\mathbf{a}_1\mathbf{x}_1-\ldots-\mathbf{a}_n\mathbf{x}_n})^2} \end{matrix} \right\}$$

(b) $\nabla^2 f(\mathbf{x})$

$$\nabla^2 f(\mathbf{x}) = \begin{Bmatrix} \frac{1+\mathbf{e}^{-\mathbf{a}_1\mathbf{x}_1-\ldots-\mathbf{a}_n\mathbf{x}_n}}{(1+\mathbf{e}^{-\mathbf{a}_1\mathbf{x}_1-\ldots-\mathbf{a}_n\mathbf{x}_n})^2} & \cdots & \frac{1+\mathbf{e}^{-\mathbf{a}_1\mathbf{x}_1-\ldots-\mathbf{a}_n\mathbf{x}_n}}{(1+\mathbf{e}^{-\mathbf{a}_1\mathbf{x}_1-\ldots-\mathbf{a}_n\mathbf{x}_n})^2} \\ \vdots & & \vdots \\ \frac{1+\mathbf{e}^{-\mathbf{a}_1\mathbf{x}_1-\ldots-\mathbf{a}_n\mathbf{x}_n}}{(1+\mathbf{e}^{-\mathbf{a}_1\mathbf{x}_1-\ldots-\mathbf{a}_n\mathbf{x}_n})^2} & \cdots & \frac{1+\mathbf{e}^{-\mathbf{a}_1\mathbf{x}_1-\ldots-\mathbf{a}_n\mathbf{x}_n}}{(1+\mathbf{e}^{-\mathbf{a}_1\mathbf{x}_1-\ldots-\mathbf{a}_n\mathbf{x}_n})^2} \end{Bmatrix}$$

(c) $\nabla f(\mathbf{x})$ when $\mathbf{a} = [1,1,1,1,1]^\top$ and $\mathbf{x} = [0,0,0,0,0]^\top$

$$\begin{Bmatrix} -.5 \\ -.5 \\ -.5 \\ -.5 \\ -.5 \end{Bmatrix}$$

(d) $\nabla^2 f(\mathbf{x})$ when $\mathbf{a} = [1,1,1,1,1]^\top$ and $\mathbf{x} = [0,0,0,0,0]^\top$

$$\begin{Bmatrix} .5 & .5 & .5 & .5 & .5 \\ .5 & .5 & .5 & .5 & .5 \\ .5 & .5 & .5 & .5 & .5 \\ .5 & .5 & .5 & .5 & .5 \\ .5 & .5 & .5 & .5 & .5 \end{Bmatrix}$$

Note that $0 \le f(\mathbf{x}) \le 1$.

8. [10 points] Show that $g(x) = -\log(f(\mathbf{x}))$ where $f(\mathbf{x})$ is a logistic function defined as above, is convex.

$$\mathbf{A} = \nabla^2 g(x) = \begin{Bmatrix} \frac{1+\mathbf{e}^{-\mathbf{a}_1\mathbf{x}_1-\ldots-\mathbf{a}_n\mathbf{x}_n}}{ln(10)(1+\mathbf{e}^{-\mathbf{a}_1\mathbf{x}_1-\ldots-\mathbf{a}_n\mathbf{x}_n})^2} & \cdots & \frac{1+\mathbf{e}^{-\mathbf{a}_1\mathbf{x}_1-\ldots-\mathbf{a}_n\mathbf{x}_n}}{ln(10)(1+\mathbf{e}^{-\mathbf{a}_1\mathbf{x}_1-\ldots-\mathbf{a}_n\mathbf{x}_n})^2} \\ \vdots & & \vdots \\ \frac{1+\mathbf{e}^{-\mathbf{a}_1\mathbf{x}_1-\ldots-\mathbf{a}_n\mathbf{x}_n}}{ln(10)(1+\mathbf{e}^{-\mathbf{a}_1\mathbf{x}_1-\ldots-\mathbf{a}_n\mathbf{x}_n})^2} & \cdots & \frac{1+\mathbf{e}^{-\mathbf{a}_1\mathbf{x}_1-\ldots-\mathbf{a}_n\mathbf{x}_n}}{ln(10)(1+\mathbf{e}^{-\mathbf{a}_1\mathbf{x}_1-\ldots-\mathbf{a}_n\mathbf{x}_n})^2} \end{Bmatrix}$$

For $g(x)$ to be convex, then $\mathbf{v}^T\mathbf{A}\mathbf{v}$ must be positive for all non-negative $\mathbf{v} \in \mathbb{R}^n$, aka semi-definite positive.

All the second partial derivatives in the Hessian of $g(x)$ are strictly positive, therefore $\mathbf{v}^T\mathbf{A}\mathbf{v}$ must always be positive, and $g(x)$ must be convex.