CS 5350/6350: Machine Learining Fall 2021

Homework 0

Handed out: 24 Aug, 2021
Due: 11:59pm, 3 Sep, 2021

- 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 to 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., $p(A)$.

   Let A^c be the complement of A i.e. A^c is the event "getting 0 heads". There is only one way of getting 0 heads, which is if you got all tails, and we also know that there are 2^{10} possible combinations of heads and tails when a coin is flipped 10 times. Therefore by the complement rule:

   $$p(A) = 1 - p(A^c)$$

   $$p(A) = 1 - \frac{1}{2^{10}} = .9990$$

2. [10 points] Given two events A and B, prove that

   $$p(A \cup B) \leq p(A) + p(B).$$

We are given the equation $p(A \cup B) = p(A) + p(B) - p(A \cap B)$.

$$p(A) + p(B) - p(A \cap B) \leq p(A) + p(B)$$

Subtract $p(A) + p(B)$ from both sides gives you

$$-p(A \cap B) \leq 0$$

We know that $p(A \cap B) \geq 0$ so the comparison holds.

When does the equality hold?

The equality holds when $p(A \cap B) = 0$ i.e. when the two events are independent.

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)

We have solved the base case above in #2. We assume $p(\cup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} p(A_i)$ is correct. We must prove $p(\cup_{i=1}^{n+1} A_i) \leq \sum_{i=1}^{n+1} p(A_i)$

$$p(\cup_{i=1}^{n} A_i \cup A_{n+1}) \leq \sum_{i=1}^{n+1} p(A_i)$$

$$p(\cup_{i=1}^{n} A_i) + p(A_{n+1}) - p(\cup_{i=1}^{n} A_i \cap A_{n+1}) \leq \sum_{i=1}^{n} p(A_i) + p(A_{n+1})$$

Since we know $p(\cup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} p(A_i)$ and $p(A_{n+1}) - p(\cup_{i=1}^{n} A_i \cap A_{n+1}) \leq p(A_{n+1})$ it must be true that $p(\cup_{i=1}^{n+1} A_i) \leq \sum_{i=1}^{n+1} p(A_i)$. The equality holds when all events are independent.

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 $p(Y)$

$$p(X = 0) = \frac{1}{10} + \frac{2}{10} = \frac{3}{10}$$

$$p(X = 1) = \frac{3}{10} + \frac{4}{10} = \frac{7}{10}$$

$$p(Y = 0) = \frac{1}{10} + \frac{3}{10} = \frac{4}{10}$$

$$p(Y=1) = \frac{2}{10} + \frac{4}{10} = \frac{6}{10}$$

ii. the conditional distributions $p(X|Y)$ and $p(Y|X)$

$$p(X=0|Y=0) = \frac{p(X=0 \cap Y=0)}{p(Y=0)} = \frac{1}{4}$$

$$p(X=1|Y=0) = \frac{p(X=1 \cap Y=0)}{p(Y=0)} = \frac{3}{4}$$

$$p(X=0|Y=1) = \frac{p(X=0 \cap Y=1)}{p(Y=1)} = \frac{1}{3}$$

$$p(X=1|Y=1) = \frac{p(X=1 \cap Y=1)}{p(Y=1)} = \frac{2}{3}$$

$$p(Y=0|X=0) = \frac{p(Y=0 \cap X=0)}{p(X=0)} = \frac{1}{3}$$

$$p(Y=1|X=0) = \frac{p(Y=1 \cap X=0)}{p(X=0)} = \frac{2}{3}$$

$$p(Y=0|X=1) = \frac{p(Y=0 \cap X=1)}{p(X=1)} = \frac{3}{7}$$

$$p(Y=1|X=1) = \frac{p(Y=1 \cap X=1)}{p(X=1)} = \frac{4}{7}$$

iii. $\mathbb{E}(X)$, $\mathbb{E}(Y)$, $\mathbb{V}(X)$, $\mathbb{V}(Y)$

$$\mathbb{E}(X) = 0 * \frac{3}{10} + 1 * \frac{7}{10} = \frac{7}{10}$$

$$\mathbb{E}(Y) = 0 * \frac{4}{10} + 1 * \frac{6}{10} = \frac{6}{10}$$

$$\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = \frac{7}{10} - \frac{49}{100} = \frac{21}{100}$$

$$\mathbb{V}(Y) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2 = \frac{6}{10} - \frac{36}{100} = \frac{24}{100}$$

iv. $\mathbb{E}(Y|X=0)$, $\mathbb{E}(Y|X=1)$, $\mathbb{V}(Y|X=0)$, $\mathbb{V}(Y|X=1)$

$$\mathbb{E}(Y|X=0) = 0 * \frac{1}{3} + 1 * \frac{2}{3} = \frac{2}{3}$$

$$\mathbb{E}(Y|X=1) = 0 * \frac{3}{7} + 1 * \frac{4}{7} = \frac{4}{7}$$

$$\mathbb{V}(Y|X=0) = \mathbb{E}(Y^2|X=0) - \mathbb{E}(Y|X=0)^2 = \frac{2}{3} - \frac{4}{9} = \frac{2}{9}$$

$$\mathbb{V}(Y|X=1) = \mathbb{E}(Y^2|X=1) - \mathbb{E}(Y|X=1)^2 = \frac{4}{7} - \frac{16}{49} = \frac{12}{49}$$

v. the covariance between X and Y

$$Cov(X,Y) = \mathbb{E}(XY) - E(X)E(Y)$$

$$= \frac{4}{10} - \frac{7}{10} * \frac{6}{10} = -\frac{2}{100} = -\frac{1}{50}$$

(b) [5 points] Are X and Y independent? Why?
$p(X|Y) = \frac{2}{3}$ and $p(X) = \frac{7}{10}$ are not equal therefore they are not independent.

(c) [5 points] When X is not assigned a specific value, are $\mathbb{E}(Y|X)$ and $\mathbb{V}(Y|X)$ still constant? Why?
No. The R.V. X and Y are dependent so $\mathbb{E}(Y|X)$ and $\mathbb{V}(Y|X)$ will be different based on the value of X.

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) $\mathbb{E}(Y)$

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

The above integral takes the form of the Guassian integral: $\int_{-\infty}^{\infty} e^{-a(x+b)^2} dx = \sqrt{\frac{\pi}{a}}$

$$= \frac{e^{1/2}}{\sqrt{2\pi}} \sqrt{2\pi} = \sqrt{e}$$

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

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

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{4x-x^2}{2}+\frac{4}{2}-\frac{4}{2}} dx$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{-(x^2-4x+4)}{2}} e^2 dx$$

$$= \frac{e^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{\frac{-(x-2)^2}{2}} dx = \frac{e^2}{\sqrt{2\pi}} \sqrt{2\pi} = e^2$$

$$\mathbb{V}(Y) = E(Y^2) - E(Y)^2 = e^2 - e$$

6. Given two random variables X and Y, show that

(a) [20 points] $\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y)$
$f_X(x)$ is the pdf for random variable X, $f_Y(y)$ is the pdf for random variable Y, etc.

$$\mathbb{E}(\mathbb{E}(Y|X)) = \int_x E(Y|X) f_X(x) dx$$

$$= \int_x f_X(x) \int_y y f_{Y|X}(y|x) dy dx$$

$$= \int_x f_X(x) \int_y y \frac{f_{X,Y}(x,y)}{f_X(x)} dy dx$$

$$= \int_x \frac{f_X(x)}{f_X(x)} \int_y y f_{X,Y}(x,y) dy dx$$

$$= \int_x \int_y y f_{X,Y}(x,y) dy dx = \int_y y \int_x f_{X,Y}(x,y) dx dy$$

$$= \int_y y f_Y(y) dy = E(Y)$$

(b) [**Bonus question** 20 points] $\mathbb{V}(Y) = \mathbb{E}(\mathbb{V}(Y|X)) + \mathbb{V}(\mathbb{E}(Y|X))$

$$\mathbb{V}(Y) = \mathbb{E}(Y^2) - \mathbb{E}(Y)^2$$

Applying the proof from above to each term gives:

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

We know that $\mathbb{V}(Y|X) = \mathbb{E}(Y^2|X) - \mathbb{E}(Y|X)^2$ therefore:

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

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

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

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

(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) $\nabla f(\mathbf{x}) = [\frac{a_1 exp(-\mathbf{a}^\top \mathbf{x})}{(1+exp(-\mathbf{a}^\top \mathbf{x}))^2}, \frac{a_2 exp(-\mathbf{a}^\top \mathbf{x})}{(1+exp(-\mathbf{a}^\top \mathbf{x}))^2}, \ldots, \frac{a_n exp(-\mathbf{a}^\top \mathbf{x})}{(1+exp(-\mathbf{a}^\top \mathbf{x}))^2}]$

(b) $\nabla^2 f(\mathbf{x})$ = n x n matrix H where each entry in the matrix is:

$$H_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{2a_i a_j exp(-\mathbf{a}^\top \mathbf{x})}{(1+\exp(-\mathbf{a}^\top \mathbf{x}))^3} - \frac{a_i a_j exp(-\mathbf{a}^\top \mathbf{x})}{(1+\exp(-\mathbf{a}^\top \mathbf{x}))^2}$$

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

$$= [\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}]$$

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

$$H = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

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

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

$$\nabla g(\mathbf{x}) = -(1+\exp(-\mathbf{a}^\top \mathbf{x}))\frac{\mathbf{a}\exp(-\mathbf{a}^\top \mathbf{x})}{(1+\exp(-\mathbf{a}^\top \mathbf{x}))^2}$$

$$= -\mathbf{a}\frac{1+\exp(-\mathbf{a}^\top \mathbf{x}) - 1}{1+\exp(-\mathbf{a}^\top \mathbf{x})}$$

$$= -\mathbf{a}(\frac{1+\exp(-\mathbf{a}^\top \mathbf{x})}{1+\exp(-\mathbf{a}^\top \mathbf{x})} - \frac{1}{1+\exp(-\mathbf{a}^\top \mathbf{x})})$$

$$= -\mathbf{a}(1-f)$$

$\nabla^2 g$ = n x n matrix H where each entry in the matrix is:

$$H_{ij} = \frac{\partial^2 g}{\partial x_i \partial x_j} = \frac{\partial}{\partial x_j}(-a_i(1-f))$$

$$= \frac{\partial}{\partial x_j}(\frac{a_i}{1+\exp(-\mathbf{a}^\top \mathbf{x})})$$

$$= a_i a_j \frac{\exp(-\mathbf{a}^\top \mathbf{x})}{(1+\exp(-\mathbf{a}^\top \mathbf{x}))^2}$$

$$= a_i a_j \frac{1+\exp(-\mathbf{a}^\top \mathbf{x}) - 1}{(1+\exp(-\mathbf{a}^\top \mathbf{x}))^2}$$

$$= a_i a_j \frac{1}{1+\exp(-\mathbf{a}^\top \mathbf{x})} - \frac{1}{(1+\exp(-\mathbf{a}^\top \mathbf{x}))^2}$$

$$= a_i a_j (f - f^2)$$

$$= a_i a_j f(1-f)$$

Therefore the matrix $H = \mathbf{a}\mathbf{a}^\top f(1-f)$. To prove that g is convex we must prove that $\mathbf{x}^\top H \mathbf{x} \geq 0$ where $\mathbf{x} \in \mathbf{R}^n$

$$\mathbf{x}\top H\mathbf{x} = \mathbf{x}^\top \mathbf{a}\mathbf{a}^\top f(1-f)\mathbf{x}$$

$$= \mathbf{x}^\top \mathbf{a}\mathbf{a}^\top \mathbf{x} f(1-f)$$

$$= (\mathbf{x}^\top \mathbf{a})^2 f(1-f)$$

$(\mathbf{x}^\top \mathbf{a})^2$ and $f(1-f)$ are positive therefore g is convex.