CS 5350/6350: Machine Leariningm 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., $p(A)$.

   Solution: We can compute this probability by just calculating the chance of all coin tosses being tails, then subtracting this from 1. The chance of 10 tails is just $(\frac{1}{2})^{10} = \frac{1}{1024}$, so $p(A) = 1 - \frac{1}{1024} = \frac{1023}{1024}$.

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?

   Solution: This equality only holds when the two events A and B are mutually exclusive, that is, both events cannot occur simultaneously. This means that $p(A \cap B) = 0$. The probability of either event occurring is $p(A \cup B) = p(A) + p(B) - p(A \cap B)$, and because $p(A \cap B) = 0$ then we know that $p(A \cup B) = p(A) + p(B)$, satisfying the equality.

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)

Solution: This equality also only holds when all events are mutually exclusive, that is for any two events the probability of both happening is 0. If this is the case, then for $n = 2$ we have already proved the equality to hold in problem 2. Assume that the equality holds for $n = k$, then for $n = k + 1$ there is an additional event A_{k+1} and this event is also mutually exclusive for all other events, that is $p(A_{k+1} \cap (\cup_{i=1}^{k} A_i)) = 0$. Because the equality holds for $n = k$, then we know that $p(\cup_{i=1}^{k} A_i) = \sum_{i=1}^{k} p(A_i)$, and $p(\cup_{i=1}^{k+1} A_i) = \sum_{i=1}^{k} p(A_i) - p(A_{k+1} \cap (\cup_{i=1}^{k} A_i))$ which satisfies the equality since all events are presumed to be mutually exclusive. Since this is true for $n = k + 1$, then it must be true for all n.

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 $p(Y)$

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

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

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

v. the covariance between X and Y

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

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

Solution:

(a) i. $p(X = 0) = \frac{3}{10}$, $p(X = 1) = \frac{7}{10}$, $p(Y = 0) = \frac{4}{10}$, $p(Y = 1) = \frac{6}{10}$.

ii. $p(X = 0|Y = 0) = 25\%$, $p(X = 1|Y = 0) = 75\%$
$p(X = 0|Y = 1) = 33.33\%$, $p(X = 1|Y = 1) = 66.67\%$

iii. $\mathbb{E}(X) = \frac{7}{10}$, $\mathbb{E}(Y) = \frac{6}{10}$
$\mathbb{V}(X) = \frac{7}{10} - (\frac{7}{10})^2 = 0.21$, $\mathbb{V}(Y) = \frac{6}{10} - (\frac{6}{10})^2 = 0.24$

iv. $\mathbb{E}(Y|X = 0) = \frac{2}{3}$, $\mathbb{E}(Y|X = 1) = \frac{4}{7}$
$\mathbb{V}(Y|X = 0) = \frac{2}{3} - (\frac{2}{3})^2 = \frac{2}{9}$, $\mathbb{V}(Y|X = 1) = \frac{4}{7} - (\frac{4}{7})^2 = \frac{12}{49}$

v. $Cov(X, Y) = (1)(1)\frac{4}{10} - \frac{7}{10}\frac{6}{10} = -\frac{2}{100}$

(b) No, their covariance is not 0.

(c) Yes, the expectation and variance in this scenario is equivalent to the expectation and variance of the marginal distribution $p(Y)$ because even though X is not assigned a specific value, we know that some X must occur, either a 0 or a 1.

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) $\mathbb{E}(Y)$

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

Solution: The PDF $p(x)$ for this normal distribution is $\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$:

(a) $\mathbb{E}(Y) = \int_{-\infty}^{\infty} e^x p(x) = e^{\frac{1}{2}}$ (computed in MATLAB)

(b) $\mathbb{V}(Y) = \int_{-\infty}^{\infty} (e^x)^2 p(x) - \mathbb{E}(Y)^2 = e^2 - e^1$ (computed in MATLAB)

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

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

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

(Hints: using definition.)

Solution:

(a) Definition of expectation then reduction:

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

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{Y|X}(y|x) f_X(x) dy dx =$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{X,Y}(x, y) dx dy =$$

$$\int_{-\infty}^{\infty} y f_Y(y) dy = \mathbb{E}(Y)$$

(b) Definition of variance is:

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

Using the equality found in part a:

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

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

The terms $(\mathbb{E}(Y|X))^2) - (\mathbb{E}(\mathbb{E}(Y|X)))^2$ is the same as $\mathbb{V}(\mathbb{E}(Y|X))$, thus:

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

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

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

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

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

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

Solution:

(a) The partial derivative of the logistic function in terms of x_i is $\frac{a_i e^{-\mathbf{a}^T \mathbf{x}}}{(1+e^{-\mathbf{a}^T \mathbf{x}})^2} = a_i f(x)(1 - f(x))$, thus the gradient can be described as:

$$\nabla f(\mathbf{x}) = (a_0 f(\mathbf{x})(1 - f(\mathbf{x})), a_1 f(\mathbf{x})(1 - f(\mathbf{x})), ..., a_n f(\mathbf{x})(1 - f(\mathbf{x})))$$

(b) For the second partial derivative, we just take the derivative again in respect for each variable:

$$\frac{\partial}{\partial x_j} a_i f(\mathbf{x})(1-f(\mathbf{x})) = a_i[f'(\mathbf{x})(1-f(\mathbf{x})) - f(\mathbf{x})f'(\mathbf{x})] = a_i f'(\mathbf{x})(1-2f(\mathbf{x}))$$

Since the partial derivative in terms of x_j of $f(\mathbf{x})$ is really just $a_j f(\mathbf{x})(1-f(\mathbf{x}))$:

$$\frac{\partial^2}{\partial^2 x_i x_j} f(\mathbf{x}) = a_i a_j f(\mathbf{x})(1-f(\mathbf{x}))(1-2f(\mathbf{x}))$$

Meaning that the Hessian matrix can be described as:

$$\nabla^2 f(\mathbf{x}) = \mathbf{A} f(\mathbf{x})(1-f(\mathbf{x}))(1-2f(\mathbf{x}))$$

Where $\mathbf{A}$ is a 2x2 matrix populated by $a_i a_j$ where i and j are the row and column indices.

(c) We see that in this case $\mathbf{a}^T\mathbf{x} = 0$, thus $f(\mathbf{x}) = \frac{e^0}{1+e^0} = \frac{1}{2}$, and so the partial derivative is $a_i(\frac{1}{2})(\frac{1}{2}) = a_i\frac{1}{4}$, and thus the gradient is:

$$\nabla f(\mathbf{x}) = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$$

(d) Again $\mathbf{a}^T\mathbf{x} = 0$, so the second partial derivative is $a_i a_j(\frac{1}{2})(\frac{1}{2})(0) = 0$, meaning that:

$$\nabla^2 f(\mathbf{x}) = \mathbf{0}$$

Where the zero matrix is of size 5x5.

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

Solution: We can prove convexity by showing that the second derivative of g is greater than or equal to 0 for all x:

$$g'(x) = -\frac{1}{f(\mathbf{x})} f'(\mathbf{x}) = -\frac{f(\mathbf{x})(1-f(\mathbf{x}))}{f(\mathbf{x})} = f(\mathbf{x}) - 1$$

$$g''(x) = f'(\mathbf{x}) = f(\mathbf{x})(1-f(\mathbf{x})) \geq 0$$

$$f(\mathbf{x}) - (f(\mathbf{x}))^2 \geq 0$$

$$1 \geq f(\mathbf{x})$$

Which is true since $0 \leq f(\mathbf{x}) \leq 1$.