CS446: Machine Learning **Spring 2017**

Problem Set 0 Solutions

Handed Out: January 25^{th}, 2017 *Handed In: NONE*

1. [Probability] Assume that the probability of obtaining heads when tossing a coin is λ.

   a. What is the probability of obtaining the first head at the $(k+1)$-th toss?

   b. What is the expected number of tosses needed to get the first head?

Solution:

a.
$$\Pr(k \text{ tails in the first } k \text{ tosses, then 1 head}) = (1-\lambda)^k\lambda$$

b. Let X be the number of the tosses required to get the first head and let $S = E[X]$. There are many ways to solve this problem.

Solution (1): Using the definition of expectation.
By definition,

$$S = E[X] = \sum_{i=1}^{\infty} i \times \lambda \times (1-\lambda)^{i-1} = \lambda \sum_{i=1}^{\infty} i \times (1-\lambda)^{i-1}. \tag{1}$$

Let
$$f(\lambda) = \sum_{i=1}^{\infty}(1-\lambda)^i = \frac{1-\lambda}{\lambda}.$$

Then, we have
$$\sum_{i=1}^{\infty} i \times (1-\lambda)^{i-1} = -f'(\lambda) = \frac{1}{\lambda^2}. \tag{2}$$

With substitution of Eq. (2) into Eq. (1), we find the expectation of number of tosses is $\frac{1}{\lambda}$.

Solution (2): Using properties of expectation.
Given that tosses are independent, and expectation is additive:

$$S = \lambda \times 1 + (1-\lambda) \times (S+1)$$

Solving for S gives $S = \frac{1}{\lambda}$.

Solution (3): Using moment generating function.
The moment generating function $\phi(t))$ of X is

$$\phi(t) = E[e^{tX}] = \sum_{x=1}^{\infty} e^{tx} P(X=x) = \sum_{x=1}^{\infty} e^{tx}\lambda \times (1-\lambda)^{x-1} = \frac{\lambda e^t}{1-e^t(1-\lambda)}.$$

Then,

$$E[X] = \frac{d}{dt}E[e^{tX}] \mid_{t=0} = \frac{d}{dt}\phi(t) \mid_{t=0} = \frac{\lambda e^t}{(1-e^t(1-\lambda))^2} \mid_{t=0} = \frac{1}{\lambda}.$$

2. [Probability] Assume X is a random variable.

   a. We define the variance of X as: $Var(X) = E[(X - E[X])^2]$. Prove that $Var(X) = E[X^2] - E[X]^2$.

   b. If $E[X] = 0$ and $E[X^2] = 1$, what is the variance of X? If $Y = a + bX$, what is the variance of Y?

   Solution:

   a. Directly from the definition of variance:

   $$\begin{aligned} E[(X - E[X])^2] &= E[X^2 - 2XE[X] + E[X]^2] = E[X^2] - 2E[XE[X]] + E[X]^2 \\ &= E[X^2] - 2E[X]^2 + E[X]^2 = E[X^2] - E[X]^2 \qquad (3) \end{aligned}$$

   where the second equality makes use of the additivity of expectations and the third makes use of the fact that $E[X]$ is a constant.

   b. Substituting the values for $E[X]$ and $E[X^2]$ in Eq. 3, we get

   $$Var(X) = E[X^2] - E[X]^2 = 1$$

   If $Y = a + bX$,

   $$\begin{aligned} E[Y^2] &= E[(a + bX)^2] = E[a^2 + 2abX + b^2X^2] \\ &= a^2 + 2abE[X] + b^2E[X^2] = a^2 + b^2 \\ E[Y] &= E[a + bX] = a + bE[X] = a \\ Var(Y) &= E[Y^2] - E[Y]^2 = a^2 + b^2 - a^2 = b^2 \end{aligned}$$

3. [Probability] John is a great fortune teller. Assume that we know three facts: 1) If John tells you that a lottery ticket will win, it will win with probability 0.99. 2) If John tells you that a lottery ticket will not win, it will not win with probability 0.99999. 3) With probability 10^{-5}, John predicts that a ticket as a winning ticket. This also means that with probability $1 - 10^{-5}$, John predicts that a ticket will not win.

   a. Given a ticket, what is the probability that it wins?

   b. What is the probability that John correctly predicts a winning ticket?

   Solution: Let T be the event "John predicts that a given ticket is a winning ticket". Let $\neg T$ be the event "John predicts that a given ticket is not a winning ticket". Similarly, let W be the event that the given ticket wins and $\neg W$ be the event that the given ticket does not win. Then:

   a. Given a ticket, the probability that it wins is:

   $$\begin{aligned} P(W) &= P(W, T) + P(W, \neg T) = P(W \mid T)P(T) + P(W \mid \neg T)P(\neg T) \\ &= 0.99 \times 10^{-5} + (1 - 0.99999) \times (1 - 10^{-5}) \\ &\approx 1.99 \times 10^{-5} \end{aligned}$$

b. The probability that John correctly predicts a winning ticket is:

$$\begin{aligned}
P(T|W) &= \frac{P(T,W)}{P(W)} = \frac{P(W \mid T)P(T)}{P(W)} \\
&= \frac{0.99 \times 10^{-5}}{0.99 \times 10^{-5} + (1 - 0.99999) \times (1 - 10^{-5})} \\
&\approx 0.497
\end{aligned}$$

4. [Calculus] Let $f(x, y) = 3x^2 + y^2 - xy - 11x$

   a. Find $\frac{\partial f}{\partial x}$, the partial derivative of f with respect to x. Find $\frac{\partial f}{\partial y}$.

   b. Find $(x, y) \in \mathbb{R}^2$ that minimizes f.

   Solution: This question serves as a review of multivariate calculus.

   a. $\frac{\partial f}{\partial x} = 6x - y - 11 \qquad \frac{\partial f}{\partial y} = 2y - x$

   b. Recall from basic calculus that a function attains its maxima and minima at points where the derivative is zero. Setting the derivative from (a.) to zero, we see that f is maximized or minimized at $(x, y) = (2, 1)$.

      One approach to show that this point is a minimizer is to consider the matrix of second derivatives, the Hessian, and show that it is positive definite. In our case, the Hessian is

$$H_f = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} = \begin{bmatrix} 6 & -1 \\ -1 & 2 \end{bmatrix}.$$

      This matrix is positive definite for all (x, y) because the principal minors are positive. Another way to check if H_f is positive definite is to verify if $z^T H_f z$ is positive for every non-zero column vector $z = [z_1 \ z_2]^T$:

$$z^T H_f z = 6z_1^2 - 2z_1 z_2 + 2z_2^2 = 5z_1^2 + z_2^2 + (z_1 - z_2)^2 > 0$$

5. [Linear Algebra] Assume that $w \in \mathbb{R}^n$ and b is a scalar. A hyper-plane in $\mathbb{R}^n$ is the set, $\{x : x \in \mathbb{R}^n, w^T x + b = 0\}$.

   a. For $n = 2$ and 3, find two example hyper-planes (say, for $n = 2$, $w^T = \begin{bmatrix} 1 & 1 \end{bmatrix}$ and $b = 2$ and for $n = 3$, $w^T = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$ and $b = 3$) and draw them on a paper.

   b. The distance between a point $x_0 \in \mathbb{R}^n$ and the hyperplane $w^T x + b = 0$ can be described as the solution of the following optimization problem:

$$\begin{aligned} &\min_x \|x_0 - x\|^2 \\ &\text{s.t. } w^T x + b = 0 \end{aligned}$$

      However, it turns out that the distance between x_0 and $w^T x + b = 0$ has an analytic solution. Derive the solution. (*Hint: you may be familiar with another way of deriving this distance; try your way too*)

c. Assume that we have two hyper-planes, $w^T x + b_1 = 0$ and $w^T x + b_2 = 0$. What is the distance between these two hyperplanes?

Solution: This problem will also appear in Hw1. We will release the solution later.

6. [Linear Algebra] One way to define a convex function is as follows. A function $f(x)$ is convex if

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$$

for all x, y and $0 < \lambda < 1$.

a. Prove that $f(x) = x^2$ is a convex function. (Prove by applying the definition.)

b. A n-by-n matrix A is a positive semi-definite matrix if $x^T Ax \geq 0$, for any $x \in \mathbb{R}^n$ s.t $x \neq 0$.

Prove that the function $f(x) = x^T Ax$ is convex if A is a positive semi-definite matrix. Note that x is a vector here. (*Hint: the solution is somewhat similar to the solution of part (a.)*)

Solution:

a. We use the definition of convex function :

$$\begin{aligned}
&f(\lambda x + (1-\lambda)y) - \lambda f(x) - (1-\lambda)f(y)\\
=&(\lambda x + (1-\lambda)y)^2 - \lambda x^2 - (1-\lambda)y^2\\
=&\lambda^2 x^2 + (1-\lambda)^2 y^2 + 2\lambda(1-\lambda)xy - \lambda x^2 - (1-\lambda)y^2\\
=&(\lambda^2 - \lambda)x^2 + ((1-\lambda)^2 - (1-\lambda))y^2 + 2\lambda(1-\lambda)xy\\
=&(\lambda - 1)\lambda(x^2 + y^2 - 2xy)\\
=&(\lambda - 1)\lambda(x - y)^2 \leq 0
\end{aligned}$$

Note that the last inequality comes from the fact $0 < \lambda < 1$.

b. As in (a.), using the definition:

$$\begin{aligned}
&f(\lambda x + (1-\lambda)y) - \lambda f(x) - (1-\lambda)f(y)\\
=&(\lambda x + (1-\lambda)y)^T A(\lambda x + (1-\lambda)y) - \lambda x^T Ax - (1-\lambda)y^T Ay\\
=&\lambda^2 x^T Ax + (1-\lambda)^2 y^T Ay + \lambda(1-\lambda)x^T Ay + \lambda(1-\lambda)y^T Ax - \lambda x^T Ax - (1-\lambda)y^T Ay\\
=&(\lambda^2 - \lambda)x^T Ax + ((1-\lambda)^2 - (1-\lambda))y^T Ay + \lambda(1-\lambda)x^T Ay + \lambda(1-\lambda)y^T Ax\\
=&(\lambda - 1)\lambda(x^T Ax + y^T Ay - x^T Ay - y^T Ax)\\
=&(\lambda - 1)\lambda(x - y)^T A(x - y) \leq 0
\end{aligned}$$

The last inequality holds because A is positive semi-definite.

7. [CNF and DNF] Consider the following Boolean function written in a conjunctive normal form

$$(x_1 \vee x_2) \wedge (x_3 \vee x_4) \wedge \ldots (x_{15} \vee x_{16})$$

If no new variable is introduced, how many clauses do you need to write down the same function in disjunctive normal form ?

Solution: Each clause in the final disjunctive normal form (DNF) should be of the form $x_{i_1} \wedge x_{i_2} \wedge \ldots \wedge x_{i_8}$, where i_1 can be 1 or 2, i_2 can be 3 or 4, and so on. Therefore, 2^8 clauses are needed to write out the final DNF.