CS761 Spring 2017 Homework 2 Solution

Assigned Mar. 13, due Mar. 27 before class

Instructions:

- Homeworks are to be done individually.
- Typeset your homework in latex using this file as template (e.g. use pdflatex). Show your derivations.
- Hand in the compiled pdf (not the latex file) online. Instructions will be provided. We do not accept hand-written homeworks.
- Homework will no longer be accepted once the lecture starts.
- Let the TA know if you have any questions about the solution:

Name: Xuezhou Zhang
Email: zhangxz1123@cs.wisc.edu

1. Let $X_0, X_1, \ldots, X_{M-1}$ denote a random sample of N-dimensional random vectors X_n, each of which has mean value m and covariance matrix R. Show that the sample mean

$$\hat{m}_t = \frac{1}{t+1} \sum_{n=0}^{t} X_n$$

and the sample covariance

$$S_t(\hat{m}_t) = \frac{1}{t+1} \sum_{n=0}^{t} (X_n - \hat{m}_t)(X_n - \hat{m}_t)^\top$$

may be written recursively as

$$\hat{m}_t = \frac{t}{t+1} \hat{m}_{t-1} + \frac{1}{t+1} X_t, \quad \hat{m}_0 = X_0,$$

and

$$S_t(\hat{m}_t) = Q_t - \hat{m}_t \hat{m}_t^\top,$$

where

$$Q_t = \frac{t}{t+1} Q_{t-1} + \frac{1}{t+1} X_t X_t^\top.$$

Proof:

a. By definition we have

$$\hat{m}_t = \frac{1}{t+1} \sum_{n=0}^{t} X_n = \frac{1}{t+1} \left(\sum_{n=0}^{t-1} X_n + X_t \right) = \frac{1}{t+1} (t m_{t-1} + X_t)$$

as needed.

b. Again by definition we have

$$\begin{aligned} S_t(\hat{m}_t) &= \frac{1}{t+1} \sum_{n=0}^{t} (X_n - \hat{m}_t)(X_n - \hat{m}_t)^\top \\ &= \frac{1}{t+1} \left[\sum_{n=0}^{t} X_n X_n^\top - \sum_{n=0}^{t} X_n \hat{m}_t{}^\top - \sum_{n=0}^{t} \hat{m}_t X_n^\top + \sum_{n=0}^{t} \hat{m}_t \hat{m}_t{}^\top \right] \\ &= \frac{1}{t+1} \sum_{n=0}^{t} X_n X_n^\top - \hat{m}_t \hat{m}_t{}^\top \end{aligned}$$

Let

$$Q_t = \frac{1}{t+1} \sum_{n=0}^{t} X_n X_n^\top,$$

then

$$S_t(\hat{m}_t) = Q_t - \hat{m}_t \hat{m}_t^\top,$$

and

$$
\begin{aligned}
Q_t &= \frac{1}{t+1}\sum_{n=0}^{t} X_n X_n^\top \\
&= \frac{1}{t+1}\left[\frac{t}{t}\sum_{n=0}^{t-1} X_n X_n^\top + X_t X_t^\top\right] \\
&= \frac{t}{t+1} Q_{t-1} + \frac{1}{t+1} X_t X_t^\top .
\end{aligned}
$$

Base cases are easy to verify.

2. Suppose we roll a fair 6-sided die 100 times. Let X be the sum of the outcomes. Bound $P(|X - 350| \geq 100)$ using Chebyshev and Hoeffding, respectively.

   Solution:

   Let $X_1, ..., X_{100}$ be the random variables representing the 100 die rolls, then they are all linearly independent, and each of them is a uniform distribution over the set $\{1, 2, 3, 4, 5, 6\}$. Therefore, we have $\mathbb{E}(X_i) = 3.5$ and $\mathrm{Var}(X_i) = \frac{35}{12}$. X is the sum of X_i's, and therefore we have $\mathbb{E}(X) = 350$ and $\mathrm{Var}(X) = \frac{875}{3}$. Chebyshev tells us

$$
\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}
$$

   In our case, $k = 100/\sqrt{875/3}$, and so

$$
\mathbb{P}(|X - 350| \geq 100) \leq \frac{1}{k^2} \approx 0.029
$$

   Hoeffding inequality states that if X_i's are random variables bounded by interval $[a_i, b_i]$, then

$$
\mathbb{P}(|\bar{X} - \mathbb{E}(\bar{X})| \geq t) \leq 2\exp\left(-\frac{2n^2t^2}{\sum(a_i - b_i)^2}\right)
$$

   where $\bar{X}$ is the mean of X_i's. Now, plug it in our problem, we get

$$
\mathbb{P}(|X - 350| \geq 100) = \mathbb{P}(|\bar{X} - 3.5| \geq 1) \leq 2\exp\left(-\frac{2 \times 100^2 \times 1^2}{\sum(6-1)^2}\right) \approx 0.000671
$$

3. Let $\mathcal{X}$ be the vector space of *finitely* nonzero sequences $X = (x_1, x_2, \ldots, x_n, 0, 0, \ldots)$. Define the norm on $\mathcal{X}$ as $\|X\| = \max |x_i|$. Let X_n be a point in $\mathcal{X}$ (a sequence) defined by

$$
X_n = \left(1, \frac{1}{2}, \frac{1}{3}, \ldots, \frac{1}{n}, 0, 0, \ldots\right).
$$

- Show that the sequence X_n is a Cauchy sequence.

  Proof:

  First, notice that by the definition of $\{X_n\}$, $||X_m - X_n|| = \frac{1}{m+1}$, for any $m < n \in \mathbb{N}$. Therefore, let $\epsilon > 0$ be given, let N be the smallest integer, s.t. $N > \frac{1}{\epsilon}$. Then for any $n > m > N$, $||X_m - X_n|| = \frac{1}{m+1} \leq \frac{1}{N} \leq \epsilon$. Therefore, X_n is a Cauchy sequence.

- Show that $\mathcal{X}$ is not complete.

  Proof: Since X_n is a Cauchy sequence, and the number of nonzero entries of X_n is monotonically increasing, X_n does not converge in $\mathcal{X}$. Therefore, $\mathcal{X}$ is not complete.

4. Determine the range and nullspace of the following linear operators (matrices):

$$A = \begin{bmatrix} 1 & 0 \\ 5 & 4 \\ 2 & 4 \end{bmatrix} \quad B = \begin{bmatrix} 1 & 0 & 1 \\ 5 & 4 & 9 \\ 2 & 4 & 6 \end{bmatrix}$$

Solution:

The range of A is the span of its two column vector, the nullspace is $\{0\}$.

The range of B is the span of its first two column vector, as the third one is a linear combination of the first two. The nullspace is therefore the span of vector (1,1,-1).

5. Let

$$A = \begin{bmatrix} 1 & 4 & 5 & 6 \\ 6 & 7 & 2 & 1 \end{bmatrix} \quad b = \begin{bmatrix} 48 \\ 30 \end{bmatrix}.$$

One solution to $Ax = b$ is $x = [1, 2, 3, 4]^\top$. Compute the least-squares solution using the SVD (explain how), and compare. Why was the solution chosen?

Solution:

Let the SVD of A be

$$A = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}.$$

Then the linear system $Ax = b$ can be written as

$$\begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} x = b,$$

$$\begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_1^\top x_1 \\ V_2^\top x_2 \end{bmatrix} = \begin{bmatrix} U_1^\top b_1 \\ U_2^\top b_2 \end{bmatrix}.$$

Let $\tilde{b_1} = U_1^\top b_1, \tilde{b_2} = U_2^\top b_2$, and $\tilde{x_1} = V_1^\top x_1$, $\tilde{x_2} = V_2^\top x_2$. Then $x = V_1 \tilde{x_1} + V_2 \tilde{x_2}$. We also know that in SVD V_1 spans $R(A^\top)$ and V_2 spans $N(A)$. These two spaces perpendicularly decomposed the domain. As a result, the least square solution should have zero component in the $N(A)$ space, this implies that $x_{ls} = V_1\tilde{x_1} = V_1\Sigma_1 U_1^\top b_1$. In our case, the least square solution is given by

$$x_{ls} = \begin{bmatrix} 0.54 \\ 2.40 \\ 3.09 \\ 3.73 \end{bmatrix}$$

Comparing the norm, $||x_{ls}|| = 9.77$, while $||x|| = 10$.

6. Consider the following process. A probability vector $p = (p_1, \ldots, p_d)$ is drawn from a Dirichlet distribution with parameter vector α. Then, a vector of category counts $x = (x_1, \ldots, x_d)$ is drawn from a multinomial distribution with probability vector p and number of trials N. Give an analytic form of $P(x \mid \alpha)$.

   Solution:

   Since $x \perp \alpha | p$, we have

$$\mathbb{P}(x|\alpha) = \int_{\triangle^d} f_p(x) f_\alpha(p) dp = \int_{\triangle^d} \frac{N!}{x_1!...x_d!} \prod_{i=1}^d p_i^{x_i} \frac{1}{B(\alpha)} \prod_{i=1}^d p_i^{\alpha_i - 1} dp.$$

7. Let $X_1, X_2, \ldots, X_m$ be a random sample, where $X_i \sim U(0, \theta)$ the uniform distribution.

   - Show that $\hat{\theta}_{ML} = \text{argmax}_\theta = \max X_i$.
     Clearly $\theta \geq \max X_i$. Now,

$$\hat{\theta}_{ML} = \text{argmax}_{\theta \geq \max X_i} \prod_{i=1}^m p(X_i|\theta) = \text{argmax}_{\theta \geq \max X_i} \frac{1}{\theta^m} = \max X_i.$$

   - Show that the density of $\hat{\theta}_{ML}$ is $f_\theta(x) = \frac{m}{\theta^m} x^{m-1}$.
     The CDF of $\hat{\theta}_{ML}$ is $F_\theta(x) = \left(\frac{x}{\theta}\right)^m$. Taking the derivative gives the expected answer.
   - Find the expected value of $\hat{\theta}_{ML}$.

$$\mathbb{E}(\hat{\theta}_{ML}) = \int_0^\theta x \frac{m}{\theta^m} x^{m-1} dx = \frac{m}{m+1}\theta.$$

   - Find the variance of $\hat{\theta}_{ML}$.

$$\text{Var}(\hat{\theta}_{ML}) = \mathbb{E}(\hat{\theta}_{ML}^2) - \mathbb{E}(\hat{\theta}_{ML})^2 = \int_0^\theta x^2 \frac{m}{\theta^m} x^{m-1} dx - \left(\frac{m}{m+1}\theta\right)^2 = \frac{m}{m+2}\theta^2 - \left(\frac{m}{m+1}\theta\right)^2.$$

8. Let $X_1, \ldots, X_n$ be a sample from $N(\mu, \sigma^2)$.

   - Show that the MLE of σ^2 is

     $$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

     Proof:
     The log-likelihood function is

     $$L(\sigma) = -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

     To compute the MLE, differentiate the log-likelihood function w.r.t σ:

     $$L'(\sigma) = \frac{1}{2\sigma^2} \left[\frac{1}{2\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X})^2 - n \right]$$

     Setting derivative to zero gives us the expected answer.

   - Show that $\hat{\sigma}^2$ has a smaller mean squared error than

     $$(n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

     Proof:
     Denote $S = (n-1)^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$. Then, we can compute $\mathbb{E}(\hat{\sigma}^2) = \frac{n-1}{n} \sigma^2$, $\mathrm{Var}(\hat{\sigma}^2) = \frac{2(n-1)\sigma^4}{n^2}$, and $\mathbb{E}(S) = \sigma^2$, $\mathrm{Var}(S) = \frac{2\sigma^4}{n-1}$. Then, we have

     $$\mathrm{MSE}(\hat{\sigma}^2) = \mathbb{E}((\hat{\sigma}^2 - \theta^2)^2) = E(\hat{\sigma}^4) - 2\theta^2 \mathbb{E}(\hat{\sigma}^2) + \theta^4 = \frac{2n-1}{n^2} \sigma^4.$$

     Similarly,

     $$\mathrm{MSE}(S) = \frac{2\sigma^4}{n-1}.$$

     We therefore have

     $$\mathrm{MSE}(\hat{\sigma}^2) = \frac{2n-1}{n^2} \sigma^4 < \frac{2n}{n^2} \sigma^4 = \frac{2}{n} \sigma^4 < \frac{2}{n-1} \sigma^4 = \mathrm{MSE}(S).$$

     as needed.

9. Consider the directed graphical model in which none of the variables is observed.

   $$\begin{array}{lll} a \searrow & & \\ & c \rightarrow d & \\ b \nearrow & & \end{array}$$

Show that $a \perp b | \emptyset$ by using a probability argument. Suppose we now observe the variable d. Show that in general $a \not\perp b | d$ (you can use a counterexample).

a.

$$\mathbb{P}(a|b) = \frac{\mathbb{P}(a,b)}{\mathbb{P}(b)} = \frac{\int_{C\times D} \mathbb{P}(a)\mathbb{P}(b)\mathbb{P}(c|a,b)\mathbb{P}(d|c)\mathrm{d}c\mathrm{d}d}{\mathbb{P}(b)} = \frac{\mathbb{P}(a)\mathbb{P}(b)}{\mathbb{P}(b)} = \mathbb{P}(a).$$

Therefore, $a \perp b | \emptyset$.

b. Since

$$\mathbb{P}(a,b,d) = \int_C \mathbb{P}(a)\mathbb{P}(b)\mathbb{P}(c|a,b)\mathbb{P}(d|c)\mathrm{d}c = \frac{\mathbb{P}(a)\mathbb{P}(b)}{\mathbb{P}(d)} \int_C \mathbb{P}(c|a,b)\frac{\mathbb{P}(c|d)}{\mathbb{P}(C)}$$

which can not be written in a form of

$$f(a,d)g(b,d)$$

because of the integral term, which implies $a \not\perp b|d$. Alternatively, you can simply find a counter-example.

10. Consider two discrete random variables $x, y \in \{A, B, C\}$. Construct a joint distribution $p(x, y)$ with the following properties:

    - $\hat{x}$ is the maximizer of the marginal $p(x)$
    - $\hat{y}$ is the maximizer of the marginal $p(y)$
    - $p(\hat{x}, \hat{y}) = 0$.

    Let $\mathbb{P}(A,B) = \mathbb{P}(A,C) = \mathbb{P}(B,A) = \mathbb{P}(C,A) = \frac{1}{4}$, with all other combination being zero.

11. Logistic regression for $y \in \{-1, 1\}$ is defined by

$$p(y \mid x; w, b) = \frac{1}{1 + e^{-y(x^\top w + b)}}.$$

Show that logistic regression is in the exponential family, that is, the probability distribution can be written in the form

$$p(y \mid x; \tilde{w}) = \frac{1}{Z(x, \tilde{w})} e^{\phi(y,x)^\top \tilde{w}}.$$

Note the mapping ϕ depends only on y, x, but not on w or b.

Proof:

$$
\begin{aligned}
p(y \mid x; w, b) &= \frac{1}{1 + e^{-y(x^\top w + b)}} \\
&= \frac{e^{y(x^\top w+b)/2}}{e^{y(x^\top w+b)/2}} \frac{1}{1 + e^{-y(x^\top w+b)}} \\
&= \frac{e^{y(x^\top w+b)/2}}{e^{y(x^\top w+b)/2} + e^{-y(x^\top w+b)/2}} \\
&= \frac{1}{e^{(x^\top w+b)/2} + e^{-(x^\top w+b)/2}} e^{y(x^\top w+b)/2} \\
&= \frac{1}{e^{(x^\top w+b)/2} + e^{-(x^\top w+b)/2}} e^{\phi(y,x)^\top \tilde{w}}
\end{aligned}
$$

where

$$\phi(y, x) = y \begin{bmatrix} x \\ 1 \end{bmatrix}$$

and

$$\tilde{w} = \begin{bmatrix} w \\ b \end{bmatrix}.$$