Tsinghua-Berkeley Shenzhen Institute
LEARNING FROM DATA
Fall 2018

Homework 4

YOUR NAME Guoqing Zhang September 8, 2020

- **Acknowledgments:** This template takes some materials from course CSE 547/Stat 548 of Washington University: `https://courses.cs.washington.edu/courses/cse547/17sp/index.html`.

  If you refer to other materials in your homework, please list here.

- **Collaborators:** I finish this homework by myself.

4.1 (a)
If X and Y are independent, we could get that $p(xy) = p(x)p(y)$, hence we get $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. Because f,g are the function of X and Y, we could get that: $p(f(x)g(y) = p(f(x))p(g(y))$, which means $f(X)$ and $g(Y)$ are independent.
We know that:

$$\rho(X,Y) = \frac{\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]}{\sqrt{var(X)var(Y)}} \tag{1}$$

$$\begin{aligned}\mathbb{E}[(f(X)-\mathbb{E}[f(X)])(g(Y)-\mathbb{E}[g(Y)])] &= \mathbb{E}[f(X)g(Y) - f(X)\mathbb{E}[g(Y)] - g(Y)\mathbb{E}[f(X)] + \mathbb{E}[f(X)]\mathbb{E}[g(Y)]] \\ &= \mathbb{E}[f(X)g(Y)] - \mathbb{E}[g(Y)]\mathbb{E}[f(X)] \\ &= 0\end{aligned}$$

Put this into (1), we could get:

$$\rho(X,Y) = 0$$

(b) $\forall f, g, \rho(f(X), g(Y)) = 0$ means:

$$\forall f, g, \mathbb{E}[(f(X)-\mathbb{E}[f(X)])(g(Y)-\mathbb{E}[g(Y)])] = 0$$

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)]\mathbb{E}[g(Y)]$$

$$\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p(x,y)f(x)g(y) = \sum_{x\in\mathcal{X}} p(x)f(x) \sum_{y\in\mathcal{Y}} p(y)g(y)$$

$$\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p(x,y)f(x)g(y) = \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p(x)p(y)f(x)g(y)$$

$$\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} [p(x,y) - p(x)p(y)]f(x)g(y) = 0$$

Because f,g could be arbitrary, so we assume that f(x)g(y) to be $p(x,y) - p(x)p(y)$, so we get:

$$\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} [p(x,y) - p(x)p(y)]^2 = 0$$

Which means:

$$p(x, y) = p(x)p(y).$$

So we get X and Y are independent,$X \perp\!\!\!\perp Y$.

4.2

$$g(y) = \mathbb{E}(X|Y = y) = \sum_{x \in \mathcal{X}} P_{X|Y}(x|y)x$$

$$g^2(y) = (\sum_{x \in \mathcal{X}} P_{X|Y}(x|y)x)^2.$$

Because $f(x) = x^2$ is a convex function, from Jenson inequality, we get:

$$(\sum_{x \in \mathcal{X}} P_{X|Y}(x|y)x)^2 \leq \sum_{x \in \mathcal{X}} P_{X|Y}(x|y)x^2.$$

Which means:

$$\begin{aligned}
\mathbb{E}[g^2(y)] &= \sum_{y \in \mathcal{Y}} P_Y(y)(\sum_{x \in \mathcal{X}} P_{X|Y}(x|y)x)^2 \\
&\leq \sum_{y \in \mathcal{Y}} P_Y(y) \sum_{x \in \mathcal{X}} P_{X|Y}(x|y)x^2 \\
&= \sum_{y \in \mathcal{Y}} \sum_{x \in \mathcal{X}} P_{XY}(x, y)x^2 \\
&= \sum_{x \in \mathcal{X}} P_X(x)x^2 \sum_{y \in \mathcal{Y}} p(y|x) \\
&= \sum_{x \in \mathcal{X}} P_X(x)x^2 \\
&= \mathbb{E}[X^2]
\end{aligned}$$

So we get: $\mathbb{E}[g^2(Y)] \leq \mathbb{E}[x^2]$

4.3 (a) We know that:

$$AV = U\Sigma V^T V = U\Sigma$$

$$A \begin{bmatrix} \upsilon_1 & ... & \upsilon_r \end{bmatrix} = \begin{bmatrix} \mu_1 & ... & \mu_r \end{bmatrix} \begin{bmatrix} \sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_r \end{bmatrix}$$

$$\begin{bmatrix} A\upsilon_1 & ... & A\upsilon_r \end{bmatrix} = \begin{bmatrix} \sigma_1\mu_1 & ... & \sigma_r\mu_r \end{bmatrix}$$

So we could get:$A\upsilon_i = \sigma_i\mu_i$. $A^T = V\Sigma U^T$, then use the same way we could get:

$$A^T\mu_i = \sigma_i\upsilon_i$$

(b) First, let me extend the V to a basis:

$$\hat{V} = \begin{bmatrix} \upsilon_1 & \cdots & \upsilon_r & \cdots & \upsilon_n . \end{bmatrix}$$

Then any X could be written as:

$$X = \sum_{i=1}^{n} \lambda_i v_i = \hat{V}\hat{\lambda}$$

Where $\lambda = \begin{bmatrix} \lambda_1 & \cdots & \lambda_r & \cdots & \lambda_n \end{bmatrix}^T$.
And it's clear that $Av_j = U\Sigma V^T v_j = \mathbf{0}, j > r$, because $v_i \perp v_j$.
Then we could get:

$$\begin{aligned} Ax &= A(\sum_{i=1}^{n} \lambda_i v_i) \\ &= \sum_{i=1}^{r} \lambda_i \sigma_i \mu_i \\ &= U\Sigma\lambda \end{aligned}$$

Where $\lambda = \begin{bmatrix} \lambda_1 & \cdots & \lambda_r \end{bmatrix}^T$.
So $\|Ax\| = \|U\Sigma\lambda\| = \|\Sigma\lambda\|$, while $\|x\| = \|\hat{V}\hat{\lambda}\| = \|\hat{\lambda}\|$.

$$\frac{\|Ax\|}{\|x\|} = \frac{\|\Sigma\lambda\|}{\|\hat{\lambda}\|} \leq \frac{\|\Sigma\lambda\|}{\|\lambda\|} = \sqrt{\frac{\sigma_1^2\lambda_1^2 + ... + \sigma_r^2\lambda_r^2}{\lambda_1^2 + ... + \lambda_r^2}}$$

Because we know: $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_r$, so:

$$\sqrt{\frac{\sigma_1^2\lambda_1^2 + ... + \sigma_r^2\lambda_r^2}{\lambda_1^2 + ... + \lambda_r^2}} \leq \sqrt{\frac{\sigma_1^2\lambda_1^2 + ... + \sigma_1^2\lambda_r^2}{\lambda_1^2 + ... + \lambda_r^2}} \leq \sigma_1$$

And It's clear we could get this value, just let $x = (\lambda_1 + ... + \lambda_r)v_1 = Cv_1$. By now I finished the proof.

4.4 (a)
Because we know: $\phi_1 = (\sqrt{P_X(1)}, ..., \sqrt{P_X(|\mathcal{X}|)})^T$. So $\phi_1(x) = \sqrt{P_X(x)}$, which means $f(x) = 1$.
(b)

$$\mathbb{E}[f^2(X)] = \sum_{x\in\mathcal{X}} p(x)f^2(x) = \sum_{x\in\mathcal{X}} (\sqrt{p(x)}f(x))^2 = \sum_{x\in\mathcal{X}} \phi^2(x) = \|\phi\|^2$$

(c)

$$\begin{aligned} \langle\phi_1, \phi_2\rangle &= \sum_{x\in\mathcal{X}} \phi_1(x)\phi_2(x) \\ &= \sum_{x\in\mathcal{X}} \sqrt{p(x)}f_1(x)\sqrt{p(x)}f_2(x) \\ &= \sum_{x\in\mathcal{X}} p(x)f_1(x)f_2(x) \\ &= \mathbb{E}[f_1(x)f_2(x)] \end{aligned}$$

4.5 (a) i. Assume $m = |\mathcal{X}|$, $n = |\mathcal{Y}|$.

$$\psi^T B\phi = \begin{bmatrix}\sqrt{p(y)}g(y_1) & \cdots & \sqrt{p(y)}g(y_n)\end{bmatrix} \begin{bmatrix} \frac{P_{XY}(x_1,y_1)}{\sqrt{P_X(x_1)P_Y(y_1)}} & \cdots & \frac{P_{XY}(x_m,y_1)}{\sqrt{P_X(x_m)P_Y(y_1)}} \\ \vdots & \vdots & \vdots \\ \frac{P_{XY}(x_1,y_n)}{\sqrt{P_X(x_1)P_Y(y_n)}} & \cdots & \frac{P_{XY}(x_m,y_n)}{\sqrt{P_X(x_m)P_Y(y_n)}} \end{bmatrix} \begin{bmatrix} \sqrt{p(x)}f(x_1) \\ \vdots \\ \sqrt{p(x)}f(x_m) \end{bmatrix}$$

$$= \sum_{x,y} p(x,y)f(x)g(y)$$

$$= E[f(x)g(y)]$$

ii.

$$\begin{aligned} B\phi(\text{given } y) &= \sum_x B(y,x)\phi(x) \\ &= \sum_x \frac{p_{XY}(x,y)}{\sqrt{p_X(x)p_Y(y)}} \cdot \sqrt{p_X(x)f(x)} \\ &= \frac{1}{\sqrt{p_Y y}} \sum_x p_{XY}(x,y)f(x) \\ &= \sqrt{p_Y(y)} \sum_x \frac{p_{XY}(x,y)}{p_Y(y)} f(x) \\ &= \sqrt{p_Y(y)}\mathbb{E}[f(X)|Y=y] \end{aligned}$$

So, $B\phi = [\sqrt{p_Y(y_1)}\mathbb{E}[f(X)|Y=y_1], ..., \sqrt{p_Y(y_n)}\mathbb{E}[f(X)|Y=y_n]]^T$, Which means $B\phi \leftrightarrow \mathbb{E}[f(X)|Y]$.

iii.

$$\begin{aligned} B^T\psi(\text{given } x) &= \sum_y B(y,x)\psi(y) \\ &= \sum_y \frac{p_{XY}(x,y)}{\sqrt{p_X(x)p_Y(y)}} \cdot \sqrt{p_Y(y)g(y)} \\ &= \frac{1}{\sqrt{p_X x}} \sum_y p_{XY}(x,y)g(y) \\ &= \sqrt{p_X(x)} \sum_y \frac{p_{XY}(x,y)}{p_X(x)} g(y) \\ &= \sqrt{p_Y(y)}\mathbb{E}[g(Y)|X=x] \end{aligned}$$

So, $B^T\psi = [\sqrt{p_X(x_1)}\mathbb{E}[g(Y)|X=x_1], ..., \sqrt{p_X(x_m)}\mathbb{E}[g(Y)|X=x_m]]^T$, Which means $B^T\psi \leftrightarrow \mathbb{E}[g(Y)|X]$.

(b) From (a)ii we know:

$$\begin{aligned} B\phi_1(\text{given } y) &= \sum_x B(y,x)\phi(x) \\ &= \sqrt{p_Y(y)} \sum_x p(x|y) \\ &= \sqrt{p_Y(y)} \end{aligned}$$

So $B\phi_1 = [\sqrt{p_Y(y_1)}, ..., \sqrt{p_Y(y_n)}]^T = \psi_1$. Use the same way we could get that $B^T\psi_1 = \phi_1$.

Compare to (a)ii, we could get that $\mathbb{E}[f(X)|Y=y]=1$, $\mathbb{E}[g(Y)|X=x]=1$. In this condition, $f(x)=1, g(y)=1$, so $\mathbb{E}[1|Y=y]$, the conditional expectation, will also be a constant.

(c)

Because of 4.3(b), we know that $\|B\|_2=\sigma_1$. So now our goal is to find the largest singular vector. Because of 4.3(a):$Av_1=\sigma_1\mu_1$, $A^T\mu_1=\sigma_1 v_1$ and 4.5(a)i and iii:$B\phi_1=\psi_1, B^T\psi_1=\phi_1$, a rational guess is that the largest singular vectors are ϕ_1, ψ_1 and the corresponding singular value $\sigma_1=1$.

$B=U\Sigma V^T$, so $BB^T=U\Sigma V^T V\Sigma^T U^T=U\Sigma^2 U^T$. If σ_1 is the largest singular value, u_1 is the largest eigenvector of BB^T, use the same way we could get that v_1 is the largest eigenvector of B^TB.

$B^TB\phi_1=B^T\psi_1=\phi_1$, so we know that ϕ_1 is a eigenvector of B^TB where the eigenvalue is 1. Now we need to prove 1 the largest eigenvalue of B^TB, so we can get ϕ_1 is the largest eigenvector.

Because $v^TB^TBv=\|Bv\|^2$, if v is eigenvector,

$$v^TB^TBv=\lambda v^Tv=\lambda.$$

So the largest eigenvalue $\hat{\lambda}$:

$$\hat{\lambda}=\max(\|Bv\|^2), s.t. \|v\|=1$$

From 4.5(a)ii, we know that: $B\phi \leftrightarrow \mathbb{E}[f(X)|Y]=g(Y)$.

$$\|B\phi\|^2=\sum_{y\in\mathcal{Y}} P_Y(y)\mathbb{E}^2[f(X)|Y=y]=\sum_{y\in\mathcal{Y}} P_Y(y)g^2(y)=\mathbb{E}[g^2(Y)]$$

From 4.2, we know:

$$\|B\phi\|^2=\mathbb{E}[g^2(Y)]\le\mathbb{E}[f^2(x)]=\|\phi\|^2=1$$

Because the ϕ could be arbitrary(The only constrain is $\|\psi\|=\|\phi\|=1$ so that they could be unit vector), so we can get: $\hat{\lambda}=1$. So we know that the largest eigenvalue is 1. And the corresponding eigenvalue ϕ_1 is the largest eigenvector of B^TB, and we can use the same way to prove that ψ_1 is also the largest eigenvector of BB^T. So we know $\sigma_1=1$ is the largest singular value. And we have

$$\|B\|_2=1.$$