Tsinghua-Berkeley Shenzhen Institute
LEARNING FROM DATA
Fall 2018

Problem Set 4

Issued: Monday 3rd December, 2018 **Due:** Monday 10th December, 2018

Notations: We will abuse some notations in this homework (as you may see in most lecture notes/papers). Capital letters can represent either random variables or matrices, and lowercase letters may represent vectors or scalars. The meaning of these notations can be understood from the context.
Specifically, X, Y denote random variables, and take values from $\mathcal{X}$ and $\mathcal{Y}$, respectively. Here we assume $\mathcal{X}$ and $\mathcal{Y}$ are both finite sets. Without loss of generality, let $\mathcal{X} = \{1, \dots, |\mathcal{X}|\}, \mathcal{Y} = \{1, \dots, |\mathcal{Y}|\}$.
The joint distribution $P_{X,Y}(x, y)$ indicates the probability $\mathbb{P}(X = x, Y = y)$. The conditional distributions $P_{Y|X}(y|x), P_{X|Y}(x|y)$, and marginal distributions $P_X(x), P_Y(y)$ are defined in the same way.

4.1. The Pearson correlation coefficient $\rho(X, Y)$ of two random variables X and Y is defined as

$$\rho(X, Y) \triangleq \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}}.$$

Prove that

$$X \perp\!\!\!\perp Y \iff \forall\, f, g,\ \rho(f(X), g(Y)) = 0,$$

where $f\colon \mathcal{X} \to \mathbb{R}, g\colon \mathcal{Y} \to \mathbb{R}$.

4.2. Given random variables X, Y, define $g(y)$ as the conditional expectation

$$g(y) \triangleq \mathbb{E}[X|Y = y] = \sum_{x\in\mathcal{X}} P_{X|Y}(x|y)x, \forall\, y \in \mathcal{Y}.$$

Then $\mathbb{E}[X|Y] = g(Y)$ is also a random variable. Prove that

$$\mathbb{E}[g^2(Y)] \le \mathbb{E}[X^2].$$

4.3. Suppose a rank-r matrix $A \in \mathbb{R}^{m\times n}$ has the SVD: $A = U\Sigma V^{\mathrm{T}}$, where $U = [u_1, \dots, u_r] \in \mathbb{R}^{m\times r}, \Sigma = \operatorname{diag}(\sigma_1, \dots, \sigma_r), V = [v_1, \dots, v_r] \in \mathbb{R}^{n\times r}, U^{\mathrm{T}}U = V^{\mathrm{T}}V = I_r, \sigma_1 \ge \cdots \ge \sigma_r > 0$.

(a) Show that $Av_i = \sigma_i u_i, A^{\mathrm{T}}u_i = \sigma_i v_i, i = 1, \dots, r$.

(b) The 2-norm of A is defined as

$$\|A\|_2 \triangleq \max_{x\in\mathbb{R}^n\colon \|x\|>0} \frac{\|Ax\|}{\|x\|}.$$

Prove that $\|A\|_2 = \sigma_1$. (*Hint: If* $U^{\mathrm{T}}U = I$, *then* $\|Ux\| = \|x\|$.)

4.4. *Information Vectors* Given a function $f\colon \mathcal{X} \to \mathbb{R}$, we can define the corresponding information vectors $\phi \in \mathbb{R}^{|\mathcal{X}|}$ with elements $\phi(x) = f(x)\sqrt{P_X(x)}$. This correspondence between function f and information vector ϕ is denoted by $\phi \leftrightarrow f(X)$. Show that

(a) $\phi_1 \leftrightarrow 1(X)$, where $\phi_1 = \left(\sqrt{P_X(1)}, \dots, \sqrt{P_X(|\mathcal{X}|)}\right)^{\mathrm{T}}$, and $1(x)$ is a constant function: $1(x) : x \mapsto 1$.

(b) $\mathbb{E}[f^2(X)] = \|\phi^2\|$, where $\phi \leftrightarrow f(X)$.

(c) $\langle \phi_1, \phi_2 \rangle = \mathbb{E}[f_1(X) f_2(X)]$, where $\phi_1 \leftrightarrow f_1(X), \phi_2 \leftrightarrow f_2(X)$.

4.5. Given two random variables $X \in \mathcal{X}, Y \in \mathcal{Y}$ with the joint distribution $P_{X,Y}(x, y)$, the corresponding B matrix ($B \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{X}|}$) is defined as

$$B(y, x) \triangleq \frac{P_{X,Y}(x, y)}{\sqrt{P_X(x)}\sqrt{P_Y(y)}}.$$

Suppose $\phi \leftrightarrow f(X)$, $\psi \leftrightarrow g(Y)$, where $f\colon \mathcal{X} \to \mathbb{R}, g\colon \mathcal{Y} \to \mathbb{R}$.

(a) Show that

i. $\mathbb{E}[f(X)g(Y)] = \psi^{\mathrm{T}} B \phi$.

ii. $B\phi \leftrightarrow \mathbb{E}[f(X)|Y]$.

iii. $B^{\mathrm{T}}\psi \leftrightarrow \mathbb{E}[g(Y)|X]$.

(b) Suppose $\phi_1 = \left(\sqrt{P_X(1)}, \dots, \sqrt{P_X(|\mathcal{X}|)}\right)^{\mathrm{T}}$, $\psi_1 = \left(\sqrt{P_Y(1)}, \dots, \sqrt{P_Y(|\mathcal{Y}|)}\right)^{\mathrm{T}}$. Show that $B\phi_1 = \psi_1, B^{\mathrm{T}}\psi_1 = \phi_1$, and interpret their meanings from the perspective of conditional expectation.

(c) Prove that $\|B\|_2 = 1$. (*Hint*: All the results you have proven can be useful.)