Tsinghua-Berkeley Shenzhen Institute
Information Theory and Statistical Learning
Fall 2020

Problem Set 6

Issued: Monday 7th December, 2020 **Due:** Sunday 20th December, 2020

Notations: We use $\mathsf{x}, \mathsf{y}, \mathsf{w}$ and $\underline{\mathsf{x}}, \underline{\mathsf{y}}, \underline{\mathsf{w}}$ to denote random variables and random vectors.

6.1. Suppose that y_1 and y_2 are independent random variables each uniformly distributed between x and $x+1$. Let $\mathsf{s} = \max(\mathsf{y}_1, \mathsf{y}_2)$ and $\mathsf{r} = \mathsf{y}_1 - \mathsf{y}_2$.

(a) Show that s is not a sufficient statistic for $p_{\mathsf{y}_1\mathsf{y}_2}(y_1, y_2; x)$.

(b) An ancillary statistic is one whose distribution does not depend on the parameters of the model. Show that r is an ancillary statistic for $p_{\mathsf{y}_1\mathsf{y}_2}(y_1, y_2; x)$.

(c) Is $\mathbf{u} = \begin{bmatrix} \mathsf{s} \\ \mathsf{r} \end{bmatrix}$ a sufficient statistic for $p_{\mathsf{y}_1\mathsf{y}_2}(y_1, y_2; x)$?

6.2. Please verify that Pearson correlation coefficient does not satisfy property (5), i.e., for any one-to-one mapping $\xi(\cdot)$ and $\eta(\cdot)$, $\rho(\xi(\mathsf{x}), \eta(\mathsf{y})) = \rho(\mathsf{x}, \mathsf{y})$.

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

$$\mathbf{B}(y, x) \triangleq \frac{P_{\mathsf{xy}}(x, y)}{\sqrt{P_{\mathsf{x}}(x)}\sqrt{P_{\mathsf{y}}(y)}}.$$

In the derivation of HGR maximal correlation analysis, given feature functions $f \colon \mathcal{X} \to \mathbb{R}, g \colon \mathcal{Y} \to \mathbb{R}$, we defined the corresponding *information vectors* as the vectors $\boldsymbol{\phi} \in \mathbb{R}^{|\mathcal{X}|}$, $\boldsymbol{\psi} \in \mathbb{R}^{|\mathcal{Y}|}$ with elements $\boldsymbol{\phi}(x) = f(x)\sqrt{P_{\mathsf{x}}(x)}$, $\boldsymbol{\psi}(y) = g(y)\sqrt{P_{\mathsf{y}}(y)}$. We denote them as $\boldsymbol{\phi} \leftrightarrow f(\mathsf{x})$ and $\boldsymbol{\psi} \leftrightarrow g(\mathsf{y})$.

(a) Show that

i. $\mathbb{E}[f(\mathsf{x})g(\mathsf{y})] = \boldsymbol{\psi}^{\mathrm{T}}\mathbf{B}\boldsymbol{\phi}$.

ii. $\mathbf{B}\boldsymbol{\phi} \leftrightarrow \mathbb{E}[f(\mathsf{x})|\mathsf{y}]$.

iii. $\mathbf{B}^{\mathrm{T}}\boldsymbol{\psi} \leftrightarrow \mathbb{E}[g(\mathsf{y})|\mathsf{y}]$.

(b) Suppose $\boldsymbol{\phi}_1 = \left[\sqrt{P_{\mathsf{x}}(1)}, \dots, \sqrt{P_{\mathsf{x}}(|\mathcal{X}|)}\right]^{\mathrm{T}}$, $\boldsymbol{\psi}_1 = \left[\sqrt{P_{\mathsf{y}}(1)}, \dots, \sqrt{P_{\mathsf{y}}(|\mathcal{Y}|)}\right]^{\mathrm{T}}$. Show that $\mathbf{B}\boldsymbol{\phi}_1 = \boldsymbol{\psi}_1, \mathbf{B}^{\mathrm{T}}\boldsymbol{\psi}_1 = \boldsymbol{\phi}_1$, and interpret their meanings from the perspective of conditional expectation.

(c) Prove that $\|\mathbf{B}\|_2 = 1$. The definition of L-p matrix norm is $\|\boldsymbol{A}\|_p \triangleq \sup_{\boldsymbol{x} \neq \mathbf{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_p}{\|\boldsymbol{x}\|_p}$.

6.4. We are now using softmax regression under a discriminative model of the form

$$Q_{\mathsf{y}|\mathsf{x}}(y|x) = \frac{\exp(\boldsymbol{f}(x)^{\mathrm{T}}\boldsymbol{g}(y) + b(y))}{\sum_{y' \in \mathcal{Y}} \exp(\boldsymbol{f}^{\mathrm{T}}(x)\boldsymbol{g}(y') + b(y'))},$$

to address the classification problems. Note that here x and y is the sample and label with alphabet $\mathcal{X}$ and $\mathcal{Y}$ repectively. Now we assume that feature $\boldsymbol{f}(\cdot)$ is decided. We hope to derive $\boldsymbol{g}(\cdot)$ and $b(\cdot)$ by

$$(\boldsymbol{g}^\star, b^\star) = \arg\min_{(\boldsymbol{g}, b)} D(P_{\mathsf{xy}} \| P_{\mathsf{x}} Q_{\mathsf{y}|\mathsf{x}}),$$

where P_{xy} is the distribution corresponding to the dataset.

(a) Explain when we have data $\{(x_i, y_i)\}_{i=1}^n$, the softmax regression optimal solution $(\boldsymbol{g}^\star, b^\star)$ is minimizing the empirical mean

$$(\boldsymbol{g}^\star, b^\star) = \arg\max_{(\boldsymbol{g}, b)} \frac{1}{n} \sum_{i=1}^{n} \log Q_{\mathsf{y}|\mathsf{x}}(y_i|x_i)$$

(b) We assume a weak dependency here, which means the true distribution P_{xy} satisfies $P_{\mathsf{xy}}(x, y) - P_{\mathsf{x}}(x)P_{\mathsf{y}}(y) = O(\epsilon)$. We define $d^\star(y) \triangleq b^\star(y) - \log P_{\mathsf{y}}(y)$. We can derive an equivalent expression of the weak dependency, which also needs some mathematical tricks to prove, $\boldsymbol{f}^{\mathrm{T}}(x)\boldsymbol{g}^\star(y) + d^\star(y) = O(\epsilon)$. Please prove that the first order approximation of $D(P_{\mathsf{xy}} \| P_{\mathsf{x}} Q_{\mathsf{y}|\mathsf{x}})$ is $\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} \frac{(P_{\mathsf{xy}}(x,y) - P_{\mathsf{x}}(x)Q_{\mathsf{y}|\mathsf{x}}(y|x))^2}{P_{\mathsf{x}}(x)Q_{\mathsf{y}|\mathsf{x}}(y|x)}$.