ORF 524: Statistical Theory and Methods
Homework 2

Lecturer: Samory Kpotufe
AI: Zhuoran Yang
Due: Oct. 14th

Exercise 1 (10 points)**.** *Let $T(X)$ be a sufficient statistic for $\mathcal{P}$. Consider the following experiment.*

- *Draw $X \sim P_\theta$, where $P_\theta \in \mathcal{P}$.*
- *Compute $T(X) = t$.*
- *Draw $X' \sim P_{X|t}$.*

Show that X' has the same (unconditional) distribution as X. For simplicity you can assume that all distributions are discrete.

Exercise 2 (Exponential families in natural form (10 points))**.** *Suppose that $\Theta \subset \mathbb{R}^d$. Let $\mathcal{P} = \{P_\theta, \theta \in \Theta\}$, where P_θ has probability density function (with respect to Lesbegue measure)*

$$f^\theta(x) = h(x)l(\theta)e^{\alpha(\theta)^T T(x)}, \qquad x \in \mathbb{R}^d,$$

where $T(x)$ is a r-vector. Show that $\mathcal{P}$ has an equivalent representation $\mathcal{P} = \{P_\alpha, \alpha \in \mathcal{A}\}$ for some set $\mathcal{A} \subset \mathbb{R}^r$, where the density of P_α is

$$f^\alpha(x) = h'(x)l'(\alpha)e^{\alpha^T T(x)}, \qquad x \in \mathbb{R}^d.$$

In other words, you need to verify that the following holds.

- *For each $\theta \in \Theta$, there exists an $\alpha \in \mathcal{A}$ such that $P_\theta = P_\alpha$.*
- *Parameter α uniquely determines P_α, that is, for any $\alpha \neq \alpha'$, P_α and $P_{\alpha'}$ are different.*

Exercise 3 (10 points)**.** *Let $\mathcal{P}$ be some family of distributions, and let $\mathcal{P}' \subseteq \mathcal{P}$ be a smaller family of distributions contained in $\mathcal{P}$. Suppose that T is sufficient for $\mathcal{P}$ and minimal sufficient for $\mathcal{P}'$. Show that T must also be minimal sufficient for $\mathcal{P}$.*

Exercise 4 (10 points)**.** *Let n random variables $X = \{X_i\}_1^n \sim \mathcal{N}^n(\mu, \mu)$, where μ is a positive real number. Here $\mathcal{N}(\mu, \mu)$ is the univariate Gaussian distribution with both mean and variance equal to μ and $\mathcal{N}^n(\mu, \mu)$ is its n-th product distribution. Consider the family of distributions $\mathcal{P} = \{\mathcal{N}^n(\mu, \mu), \mu > 0\}$.*

(1). Find a minimal sufficient statistic for $\mathcal{P}$.

(2). In the case of $n = 1$, consider another statistic $T_0(x) = x$. Is it sufficient? Is it minimal?

Exercise 5 (10 points)**.** *Suppose that $X_1, \ldots, X_n$ are i.i.d. d-dimensional Gaussian random vectors with mean μ and covariance Σ. Argue that $(\widehat{\mu}, \widehat{\Sigma})$ is a minimal sufficient statistic for $\mathcal{P} = \{\mathcal{N}^n(\mu, \Sigma)\}$, the family of Gaussian distributions with unknown Σ and μ.*

Exercise 6 (10points)**.** *Let $x = \{x_i\}_{i=1}^n$ be the realization of n i.i.d. Gaussian random variables $X = \{X_i\}_1^n \sim \mathcal{N}^n(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$, show that the maximum likelihood estimator for $\theta = (\mu, \sigma^2)$ is $(\bar{x}, S_n^2)$, where*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i, \text{ and } S_n^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})^2.$$

Exercise 7 (10 points)**.** *Let $x = \{x_i\}_1^n$ be i.i.d. realizations of a random variable $\xi \sim$ Uniform$([0, \theta])$, where $\theta > 0$. We have shown that the maximum likelihood estimator for θ is $\widehat{\theta} = \max_i x_i$. Show that*

(1). $\widehat{\theta}$ has a density with respect to Lebesgue measure.

(2). $\widehat{\theta}$ is biased.

Note*: In question (1), you only need to show that the cumulative distribution function of $\widehat{\theta}$ is absolutely continuous (easier than showing Lebesgue domination).*

Exercise 8 (10 points)**.** *This exercise is about the expectation-maximization (EM) algorithm. Derive an EM algorithm for mixture of K Gaussians with diagonal covariance matrices. In other words, suppose that we observe n i.i.d. observations $\{x_i\}_{i=1}^n$ of d-dimensional random vectors*

$$X \sim \sum_{\ell=1}^K \frac{1}{K} N(\mu_\ell, \Sigma_\ell), \text{ where } \Sigma_\ell = diag(\sigma_{\ell,1}^2, \ldots, \sigma_{\ell,d}^2),$$

derive the EM algorithm that estimates the parameters $\{\mu_\ell, \Sigma_\ell\}_{\ell=1}^K$.

Exercise 9 (10 points)**.** *This exercise relates maximum likelihood estimation with information theory. Let $\mathcal{P} = \{P_\theta \colon \theta \in \Theta\}$ be a family of probability measures on $\mathcal{X}$ with density f_θ w.r.t. to some measure σ. Assume for simplicity that all f_θ have the same support. Let $X = \{X_i\}_{i=1}^n$ be n i.i.d. random variables with $X_i \sim P_{\theta_0} \in \mathcal{P}$ for each i, where θ_0 is unknown. Let $x = \{x_i\}_{i=1}^n$ be the realization of X.*

(1). Let $L(\theta; x)$ be the likelihood function; express $\mathbb{E}\{n^{-1} \cdot \log L(\theta; X)\}$ in terms of information measures (entropy and/or K-L divergence).

(2). For any fixed θ, give a simple unbiased estimator of $\mathbb{E}\{n^{-1} \cdot \log L(\theta; X)\}$. Suppose this estimate is close to $\mathbb{E}\{n^{-1} \cdot \log L(\theta; X)\}$ (for instance for sufficiently large n), explain in simple terms (nothing rigorous here), how MLE might be interpreted as minimizing some notion of distance between distributions.

(3). Derive a simple form for the K-L divergence between two multivariate Gaussians $\mathcal{N}(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$. Here $\mu_1, \mu_2 \in \mathbb{R}^d$ and $\Sigma \in \mathbb{R}^{d \times d}$ is positive definite.

Suppose now that $\mathcal{P} = \{\mathcal{N}(\mu, \Sigma), \mu \in \mathbb{R}^d\}$, Σ fixed. Conclude that maximizing $\mathbb{E}\{n^{-1} \cdot \log L(\theta; X)\}$ is the same as minimizing some distance measure in parameter space*.*

Exercise 10 (10 points)**.** *Suppose we have data $(\mathbf{x}_i, y_i)_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$. The Ridge estimator of the linear model $Y = X^T\beta + \epsilon, \mathbb{E}\epsilon = 0, \epsilon \perp\!\!\!\perp X$, is defined as the minimizer of the following problem:*

$$\min_{\beta \in \mathbb{R}^d} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \beta)^2 + \lambda \|\beta\|^2, \qquad \lambda \geq 0,$$

where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^d$.

(1). Show that there exists some $\lambda \geq 0$, such that the above minimization problem has a unique minimizer.

(2). Derive the minimizer when the unique minimizer exists.

Exercise 11 (10 points)**.**

(1) For Ridge regression, derive a MAP interpretation. That is come up with a proper Bayesian setting where the MAP estimator corresponds to the Ridge estimator. You can consider a fixed design setting.

(2) Reduce the general polynomial model

$$Y = \mathrm{poly}_k(X) + \epsilon \qquad \mathbb{E}\epsilon = 0, \quad \epsilon \perp\!\!\!\perp X,$$

to the linear model and derive a solution. Note that $\mathrm{poly}_k(x)$, $x \in \mathbb{R}^d$ *is any polynomial of some degree* k *(*$k \geq 1$*), i.e.*

$$\mathrm{poly}_k(x) = \sum_{\ell \in \mathbb{N}^p : \sum \ell_i \leq k} w_\ell x^\ell.$$