Spring 2016

ECE 598

Information-theoretic methods in high-dimensional statistics

Due: Mar 3, 2016

Prof. Yihong Wu

Rules:

- It is mandatory to type your solutions in LaTeX. Email your solution in pdf by midnight of the due date to `yihongwu@illinois.edu` with subject line `Homework XX: your name`.
- Justify your work rigorously. As long as you are able to prove the result or a stronger version, there is no need to follow the hints.

1. (Coin flips) Consider the experiment where we observe $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Bern}(\theta)$ with $\theta \in \Theta = [0,1]$ and estimate the bias θ. Consider the quadratic loss function $\ell(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$ and denote the minimax risk by R^*.

   (a) Use the empirical frequency $\hat{\theta}_{\mathsf{emp}} = \bar{X}$ to estimate θ. Compute and plot the risk $R_\theta(\hat{\theta})$ and show that
   $$R^* \leq \frac{1}{4n}.$$

   (b) Compute the Fisher information of $P_\theta = \text{Bern}(\theta)^{\otimes n}$ and $Q_\theta = \text{Binom}(n, \theta)$. Explain why they are equal.

   (c) Invoke the Bayesian Cramér-Rao lower bound to show that
   $$R^* = \frac{1 + o(1)}{4n}.$$

   (d) Notice that the risk of $\hat{\theta}_{\mathsf{emp}}$ is maximized at $1/2$ (fair coin), which suggests that it might be possible to hedge against this situation by the following randomized estimator
   $$\hat{\theta}_{\mathsf{rand}} = \begin{cases} \hat{\theta}_{\mathsf{emp}}, & \text{with probability } \delta \\ \frac{1}{2} & \text{with probability } 1 - \delta \end{cases}$$
   Find the worst-case risk of $\hat{\theta}_{\mathsf{rand}}$ as a function of δ. Choose the best δ and show that this leads to a better upper bound:
   $$R^* \leq \frac{1}{4(n+1)}.$$

   (e) Randomization is always improvable when the loss is convex; so we should always average out the randomness by considering the estimator
   $$\hat{\theta}^* = \mathbb{E}[\hat{\theta}_{\mathsf{rand}} | X] = \bar{X}\delta + \frac{1}{2}(1 - \delta).$$
   Optimizing over δ to minimize the worst-case risk, find the resulting estimator $\hat{\theta}^*$ and its risk, show that it is constant (independent of θ), and conclude
   $$R^* \leq \frac{1}{4(1 + \sqrt{n})^2}.$$

(f) (Equalizer) Prove the following general fact: Given an experiment $\{P_\theta : \theta \in \Theta\}$ and a loss function $\ell(\theta, \hat{\theta})$, if for some prior π the corresponding Bayes estimator $\hat{\theta}$ has constant risk, namely, $R_\theta(\hat{\theta})$ is the same for all $\theta \in \Theta$, then $\hat{\theta}$ is minimax.

(g) Next we show $\hat{\theta}^*$ found in part (e) is exactly minimax and hence

$$R^* = \frac{1}{4(1+\sqrt{n})^2}.$$

Consider the following prior Beta(a, b) with density:

$$\pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1}, \quad \theta \in [0, 1],$$

where $\Gamma(a) \triangleq \int_0^\infty x^{a-1} e^{-x} dx$. Show that if $a = b = \frac{\sqrt{n}}{2}$, $\hat{\theta}^*$ coincides with the Bayes estimator for this prior, which is therefore least favorable. (Hint: work with the sufficient statistic $S = X_1 + \ldots + X_n$.)

(h) Show that the least favorable prior is not unique; in fact, there is a continuum of them. (Hint: consider the Bayes estimator $\mathbb{E}[\theta|X]$ and show that it only depends on the first $n+1$ moments of π.)

(i) (Nonparametric extension) Consider the following nonparametric model $\mathcal{P} = \mathcal{M}([0, 1])$, the set of all probability distributions on $[0, 1]$. The data are $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P \in \mathcal{P}$ and the goal is to estimate the mean of P under the quadratic loss. Show that the minimax risk is

$$R^* = \frac{1}{4(1+\sqrt{n})^2}.$$

(Hint: for any $[0, 1]$-valued random variable Z, show that $\text{var}(Z) \le \mathbb{E}[Z](1 - \mathbb{E}[Z])$.)

2. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P_\theta$ and $\theta \in [-a, a]$.

(a) State appropriate regularity conditions and prove the Chernoff-Rubin-Stein lower bound on the minimax risk:

$$\inf_{\hat{\theta}} \sup_{\theta \in [-a,a]} \mathbb{E}_\theta[(\theta - \hat{\theta})^2] \ge \min_{0<\epsilon<1} \max \left\{ \epsilon^2 a^2, \frac{(1-\epsilon)^2}{n\bar{I}} \right\},$$

where $\bar{I} = \frac{1}{2a}\int_{-a}^{a} I(\theta) d\theta$ is the average Fisher information. (Hint: You can proceed as in the classical proof of Bayesian Cramér-Rao by expanding $\int_{-a}^{a} (\theta - \hat{\theta}(x)) \frac{\partial p_\theta}{\partial \theta} d\theta$.)

(b) Simplify the above bound and show that

$$\inf_{\hat{\theta}} \sup_{\theta \in [-a,a]} \mathbb{E}_\theta[(\theta - \hat{\theta})^2] \ge \left(\frac{1}{a^{-1} + \sqrt{n\bar{I}}} \right)^2$$

(c) Assuming the continuity of $\theta \mapsto I(\theta)$, show that the above result also leads to the optimal local minimax lower bound which was obtained in class from Bayesian Cramér-Rao:

$$\inf_{\hat{\theta}} \sup_{\theta \in [\theta_0 \pm n^{-1/4}]} \mathbb{E}_\theta[(\theta - \hat{\theta})^2] \ge \frac{1+o(1)}{nI(\theta_0)}.$$

3. (More properties of f-divergences)

   (a) (Invariance) For any one-to-one transformation $g : \mathcal{X} \to \mathcal{Y}$, show that

   $$D_f(P_{g(X)} \| Q_{g(X)}) = D_f(P_X \| Q_X).$$

   Hence f-divergences are invariant under translation, dilation or rotation.

   (b) (Sufficiency) Let Y be a sufficient statistic of X for testing P_X and Q_X. Show that

   $$D_f(P_Y \| Q_Y) = D_f(P_X \| Q_X).$$

   (c) Show that

   $$D_f(P_0 \otimes Q \| P_1 \otimes Q) = D_f(P_0 \| P_1).$$

   (d) Show that

   $$d_{\mathrm{TV}}\left(\prod_{i=1}^k P_i, \prod_{i=1}^k Q_i\right) \le \sum_{i=1}^k d_{\mathrm{TV}}(P_i, Q_i).$$

   (Hint: use the coupling characterization of d_{TV}).

4. (f-divergences for Gaussian distributions) Let Σ be a positive semidefinite matrix.

   (a) Show that $d_{\mathrm{TV}}(\mathcal{N}(\theta, \Sigma), \mathcal{N}(0, \Sigma)) = 1 - 2Q(\|\Sigma^{-1/2}\theta\|_2/2)$, where $Q(a) \triangleq \int_a^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$ denotes standard normal tail probability. (Hint: first prove for $p = 1$ then; for general p, apply whitening and use 3(a) and 3(c).)

   (b) Compute $\chi^2(\mathcal{N}(\theta, \Sigma), \mathcal{N}(0, \Sigma))$.

   (c) Compute $H^2(\mathcal{N}(\theta, \Sigma), \mathcal{N}(0, \Sigma))$.

5. (Joint range) Consider $L(P\|Q) = \int \frac{(P-Q)^2}{P+Q}$ and squared Hellinger distance $H^2(P, Q) = \int(\sqrt{P} - \sqrt{Q})^2$.

   (a) (10%) Show that L is an f-divergence.

   (b) (20%) Find and plot the joint range of H^2 versus L.

   (c) (70%) Find the close-form (not parametric form) expressions of the lower and upper boundary (if they exist) and *rigorously* prove your results are in fact the boundaries.