Stat 300A Theory of Statistics

Homework 2

Andrea Montanari *Due on October 10, 2018*

- Solutions should be complete and concisely written. Please, use a separate sheet (or set of sheets) for each problem.
- We will be using Gradescope (`https://www.gradescope.com`) for homework submission (you should have received an invitation) - no paper homework will be accepted. Handwritten solutions are still fine though, just make a good quality scan and upload it to Gradescope.
- You are welcome to discuss problems with your colleagues, but should write and submit your own solution.

1: Tweedie's formula

(*a*) Consider the normal mean model $\mathsf{P}_\theta = \mathsf{N}(\theta, \sigma^2)$, $\theta \in \Theta = \mathbb{R}$, and assume the variance σ^2 to be known. Let Q be a prior distribution for the parameter θ. Show that the posterior expectation (which is Bayes optimal for the loss $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$) is given by

$$\hat{\theta}_{\text{Bayes}}(x) = x + \sigma^2 \frac{\mathrm{d}}{\mathrm{d}x} \log \mathsf{p}(x)\,, \tag{1}$$

where $\mathsf{p}(x) = \int \mathsf{p}_\theta(x)\, \mathsf{Q}(\mathrm{d}\theta)$ is the marginal distribution of x.

(*b*) Generalize the above formula to the case of an exponential family in canonical form defined by the following density in $\mathbb{R}^d$:

$$\mathsf{p}_{\boldsymbol{\theta}}(\boldsymbol{x}) = \frac{1}{Z(\boldsymbol{\theta})}\, e^{\langle \boldsymbol{\theta}, \boldsymbol{x} \rangle}\, h(\boldsymbol{x})\,. \tag{2}$$

(Here $h : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ can be assumed to be differentiable.)

2: Estimating a single bit

Consider a statistical model with two elements $\mathscr{P} = \{\mathsf{P}_0, \mathsf{P}_1\}$, on a common space $\mathcal{X} \subseteq \mathbb{R}^n$. Hence the parameter space is $\Theta = \{0, 1\}$. We consider a prior distribution Q on Θ, which is completely specified by $\mathsf{Q}(\{1\}) = \mathsf{q}$, whence $\mathsf{Q}(\{0\}) = \bar{\mathsf{q}} = 1 - \mathsf{q}$. We will assume that P_0, P_1 have densities, denoted respectively by p_0, p_1. Finally, we use the decision space $\mathcal{A} = \{0, 1\}$, and the loss

$$L(\hat{\theta}, \theta) = c_1\, \mathbf{1}_{\{\theta=0, \hat{\theta}=1\}} + c_2\, \mathbf{1}_{\{\theta=1, \hat{\theta}=0\}}\,. \tag{3}$$

(*a*) Derive an expression for the Bayes optimal estimator.

(*b*) Derive an expression for the Bayes risk $R_{\mathrm{B}}(\mathsf{Q})$.

(*c*) Assume $c_1 = c_2 = 1$, and consider the case of a uniform prior $\mathsf{Q} = \mathsf{Q}_{\mathsf{unif}}$. Show that the Bayes risk is given by $R_{\mathrm{B}}(\mathsf{Q}_{\mathsf{unif}}) = (1 - \|\mathsf{P}_0 - \mathsf{P}_1\|_{\mathrm{TV}})/2$ where the total variation distance of two probability distributions with densities $\mathsf{p}_0, \mathsf{p}_1$ is defined as

$$\big\|\mathsf{P}_0 - \mathsf{P}_1\big\|_{\mathrm{TV}} = \frac{1}{2} \int \big|\mathsf{p}_0(\boldsymbol{x}) - \mathsf{p}_1(\boldsymbol{x})\big|\, \mathrm{d}\boldsymbol{x}\,. \tag{4}$$

(*d*) Always assume $c_1 = c_2 = 1$. Provide an example of distributions $\{\mathsf{P}_0, \mathsf{P}_1\}$ such that $R_{\rm B}(\mathsf{Q}) > R_{\rm B}(\mathsf{Q}_{\sf unif})$ for some non-uniform prior Q. For this point, you do not need to limit yourself to $\mathsf{P}_0, \mathsf{P}_1$ with a density, it might be easier to consider a finite sample space $\mathcal{X}$.

3: Stochastic block model

The objective of this problem is to derive a lower bound on the risk in estimating the community structure in a stochastic block model (SBM). We will be concerned with the two-groups symmetric SBM, which is a distribution over graphs defined as follows. The parameter is a vector $\boldsymbol{\theta} \in \Theta = \{+1, -1\}^n$. For each $\boldsymbol{\theta} \in \Theta$, $\mathsf{P}_{\boldsymbol{\theta}}$ is a probability distribution over undirected graphs $G = (V, E)$, with vertex set $V = [n] \equiv \{1, \dots, n\}$ and independent edges with edge probabilities

$$\mathsf{P}_{\boldsymbol{\theta}}\Big((i,j) \in E\Big) = \begin{cases} a/n & \text{if } \theta_i = \theta_j, \\ b/n & \text{if } \theta_i \neq \theta_j. \end{cases} \tag{5}$$

Equivalently, we can regard $\mathsf{P}_{\boldsymbol{\theta}}$ as a probability distribution over symmetric $0-1$ matrices $\boldsymbol{X}$ (the adjacency matrix of G).

We consider the loss function (for $\hat{\boldsymbol{\theta}} \in \mathcal{A} = \{+1, -1\}^n$):

$$L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = 1 - \left|\frac{1}{n}\langle \hat{\boldsymbol{\theta}}, \boldsymbol{\theta}\rangle\right|^2 . \tag{6}$$

Note that $0 \le L(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) \le 1$, with $L(\boldsymbol{\theta}, \boldsymbol{\theta}) = 0$. We further assume a uniform prior $\mathsf{Q}(\{\boldsymbol{\theta}\}) = 1/2^n$ for all $\boldsymbol{\theta}$. (Equivalently, under Q, $\boldsymbol{\theta}$ has i.i.d. components $\theta_i \sim \mathsf{Unif}(\{+1, -1\})$.)

Our objective in this homework is to derive a lower bound on the Bayes risk $R_{\rm B}(\mathsf{Q})$.

(*a*) Does the loss function (6) seem a reasonable choice to you? Why not use something simpler, such as $\tilde{L}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) = \{1 - (\langle \hat{\boldsymbol{\theta}}, \boldsymbol{\theta}\rangle / n)\}/2$, which counts the number of incorrectly estimated vertex labels?

(*b*) Let $\hat{\boldsymbol{\theta}}_{\sf B}$ denote the Bayes optimal estimator. For any fixed permutation $\pi \in S_n$ (a permutation over n objects), and a vector $\boldsymbol{v} \in \mathbb{R}^n$, denote by $\boldsymbol{v}^\pi$ the vector obtained by permuting the entries of $\boldsymbol{v}$ according to π. Show that there exists a (possibly randomized) Bayes optimal estimator such that, for any fixed permutation π, $(\boldsymbol{\theta}^\pi, \hat{\boldsymbol{\theta}}_{\sf B}^\pi)$ has the same distribution as $(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}_{\sf B})$.

[Hint: Let G^π the graph obtained by permuting the vertices of G. Show that given a Bayes optimal estimator $\hat{\boldsymbol{\theta}}_{0,\sf B}$, you can construct a new (possibly randomized) estimator $\hat{\boldsymbol{\theta}}_{\sf B}$, that has the same risk as $\hat{\boldsymbol{\theta}}_{0,\sf B}$ and such that $\hat{\boldsymbol{\theta}}_{\sf B}(G^\pi)$ has the same distribution as $\hat{\boldsymbol{\theta}}_{\sf B}(G)^\pi$.]

(*c*) Prove the lower bound

$$R_{\rm B}(\mathsf{Q}) \ge \frac{1}{2}\Big(1 - \frac{1}{n}\Big)\, \Big\{1 - \mathbb{E}\Big(\hat{\theta}_{{\sf B},1}(G)\hat{\theta}_{{\sf B},2}(G)\theta_1\theta_2\Big)\Big\} . \tag{7}$$

(*d*) Let $\boldsymbol{\theta}_{\sim 1} = (\theta_2, \dots, \theta_n)$ be the vector of vertex labels, except θ_1. Derive the following lower bounds

$$\begin{aligned} R_{\rm B}(\mathsf{Q}) &\ge \frac{1}{2}\Big(1 - \frac{1}{n}\Big) \inf_{\hat{\theta}_1(\,\cdot\,), \hat{\theta}_2(\,\cdot\,)} \Big\{1 - \mathbb{E}\Big(\hat{\theta}_1(G; \boldsymbol{\theta}_{\sim 1})\hat{\theta}_2(G; \boldsymbol{\theta}_{\sim 1})\theta_1\theta_2\Big)\Big\} & (8) \\ &\ge \frac{1}{2}\Big(1 - \frac{1}{n}\Big) \left\{1 - \sup_{\hat{\theta}_1(\,\cdot\,)} \mathbb{E}\Big(\hat{\theta}_1(G; \boldsymbol{\theta}_{\sim 1})\theta_1\Big)\right\} . & (9) \end{aligned}$$

In other words, we reduced the problem of lower bounding $R_{\rm B}(\mathsf{Q})$ to the problem of lower bounding the Bayes risk in estimating θ_1 given observations G, $\boldsymbol{\theta}_{\sim 1}$.

(*e*) Let $N_+ = \#\{i \in \{2, \dots, n\} : \theta_i = +1\}$ be the number of vertices among $\{2, \dots, n\}$ with label $+1$ and $N_- = \#\{i \in \{2, \dots, n\} : \theta_i = -1\}$ the number of vertices with label -1. Further, define the number of edges that connect vertex 1 with these two sets of vertices:

$$X_+ = \#\big\{j \in \{2, \dots, n\} : \theta_j = +1,\ (1, j) \in E\big\}\,, \tag{10}$$

$$X_- = \#\big\{j \in \{2, \dots, n\} : \theta_j = -1,\ (1, j) \in E\big\}\,. \tag{11}$$

Prove that (N_+, N_-, X_+, X_-) is a sufficient statistic in the problem of estimating θ_1 from observations $(G, \boldsymbol{\theta}_{\sim 1})$. Write the conditional distribution of (N_+, N_-, X_+, X_-), given θ_1.

At this point, we reduced the original problem to a much simpler one, namely the problem of estimating a uniformly random bit $\theta_1 \sim \mathsf{Unif}(\{+1, -1\})$from observations (N_+, N_-, X_+, X_-). This problem further simplifies as $n \to \infty$ with a, b fixed. We will assume, for the sake of simplicity, $n = 2m + 1$. First of all $N_+ = n - 1 - N_-$ is a binomial random variable $\text{Binom}(n - 1, 1/2)$ and hence concentrates around $m = (n - 1)/2$. We can make the simplifying assumption that $N_+ = N_- = m$. Second, conditional on N_+, N_-, we see that X_+, X_- are independent with

$$\begin{aligned} \theta_1 = +1 \quad &\Rightarrow \quad X_+ \sim \text{Binom}(a/n, N_+),\ \ X_- \sim \text{Binom}(b/n, N_-)\,, \\ \theta_1 = -1 \quad &\Rightarrow \quad X_+ \sim \text{Binom}(b/n, N_+),\ \ X_- \sim \text{Binom}(a/n, N_-)\,, \end{aligned}$$

For large n, we can approximate the above binomials by Poisson distributions.

This suggests the following model. We deed to estimate a single bit $\theta \sim \mathsf{Unif}(\{+1, -1\})$ from observations (Z_+, Z_-), whereby

$$\begin{aligned} \theta_1 = +1 \quad &\Rightarrow \quad Z_+ \sim \text{Poisson}(a/2),\ \ Z_- \sim \text{Poisson}(b/2)\,, \\ \theta_1 = -1 \quad &\Rightarrow \quad Z_+ \sim \text{Poisson}(b/2),\ \ Z_- \sim \text{Poisson}(a/2)\,, \end{aligned}$$

Denote by $R^s_{\rm B}(a, b)$ the Bayes risk in this simplified one-bit problem (with loss $L^s(\hat{\theta}, \theta) = \mathbf{1}_{\{\hat{\theta} \neq \theta\}}$). It is possible to make the above argument rigorous, thus getting:

$$R_{\rm B}(\mathsf{Q}) \geq 2\, R^s_{\rm B}(a, b) + o_n(1)\,. \tag{12}$$

Here $o_n(1)$ denotes a term vanishing as $n \to \infty$.

(*f*) Derive an expression for $R^s_{\rm B}(a, b)$ as a function of a, b and show that it vanishes as $a - b \to \infty$.

[Hint: Use the results of the previous problem.]