Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.437 Inference and Information
Spring 2015

Problem Set 9

Issued: Tuesday, May 5, 2015 **Due:** Never

Problem 9.1

Let $\mathsf{y} = [y_1, \ldots, y_N]$ be a binary string that represents the outcomes of N coin tosses, where $y_n = 1$ corresponds to "heads" on the nth toss, and $y_n = 0$ corresponds to "tails" on the nth toss. For the first k tosses, we use a biased coin with probability of head $q > 1/2$, while for the remaining $N - k$ tosses, we use a fair coin. In this problem, we investigate the behavior of y for large N.

(a) In this part, let $k = \rho N$ where ρ is a rational number and N only takes values such that ρN is an integer. Show that for all $\gamma < 1/2$,

$$\mathbb{P}\left(\frac{1}{N}\sum_{i=1}^{N} y_i \leq \gamma\right) \dot{\leq} e^{-N\,E_*(\gamma)} \quad \text{as } N \to \infty,$$

where

$$E_*(\gamma) = \min_{(p_1,p_2)\in\mathcal{S}} \left[\beta D_{\mathrm{B}}(p_1\|q) + (1-\beta)\left(\ln(2) - H_{\mathrm{B}}(p_2)\right)\right],$$

with $H_{\mathrm{B}}(\delta)$ denoting the entropy of a Bernoulli distribution with parameter δ and $D_{\mathrm{B}}(\delta_1\|\delta_2)$ denoting the divergence between two Bernoulli distributions with parameters δ_1 and δ_2, respectively. Express the constant β and the set $\mathcal{S}$ in terms of q and ρ.

(b) In this part, number of heads in the first N tosses is a variable denoted as k and takes values from $\{0, 1, ..., N\}$. Show that the normalized model capacity C_N/N vanishes as $N \to \infty$.

Problem 9.2

Let $y_1, \ldots, y_N$ be a sequence of iid discrete random variables with a known alphabet of size $M < \infty$. We do not know anything else about the distribution.

(a) Determine an appropriate parameterization of the pdf for y. How many parameters do you have?

Generalizing the asymptotic result from lecture to multidimensional $\mathbf{x} \in \mathbb{R}^d$, we have

$$I(\mathsf{x}; \mathsf{y}) = \mathbb{E}_{\mathsf{x}}[I(\mathsf{x} = \mathbf{x}; \mathsf{y})]$$

where

$$I(\mathsf{x} = \mathbf{x}; \mathsf{y}) = \frac{d}{2} \ln \frac{N}{2\pi e} + \frac{1}{2} \ln |\mathbf{J}_{\mathsf{y}}(\mathbf{x})| + \ln \frac{1}{p_{\mathsf{x}}(\mathbf{x})} + o(1)$$

with $\mathbf{J}_{\mathsf{y}}(\mathbf{x})$ denoting the Fisher Information matrix

$$[\mathbf{J}_{\mathsf{y}}(\mathbf{x})]_{ij} = -\mathbb{E}\left[\frac{\partial^2}{\partial x_i \partial x_j} \log p_{\mathsf{y}}(y; \mathbf{x})\right],$$

and $p_{\mathsf{x}}(\mathbf{x})$ the mixture weights.

(b) What does the expression for $I(\mathsf{x} = \mathbf{x}; \mathsf{y})$ above suggest might be a good first order approximation for the normalized model capacity C_N/N as N gets very large?

(c) Using the expression for $I(\mathsf{x} = \mathbf{x}; \mathsf{y})$ above, determine the form of Jeffrey's prior as a function of $|\mathbf{J}_{\mathsf{y}}(\mathbf{x})|$.

(d) (**practice**) Compute $|\mathbf{J}_{\mathsf{y}}(\mathbf{x})|$.

Problem 9.3

Consider the regression model H_K described in lecture:

$$\mathsf{y}_i = \sum_{k=0}^{K} \mathsf{w}_k x_i^k + \mathsf{z}_i, \quad 1 \le i \le N.$$

The z_i represent independent additive Gaussian noise, with distribution $\mathcal{N}(0, \sigma^2 I)$. The experimenter decides to make N measurements $\mathsf{y}_1, \ldots, \mathsf{y}_N$ and chooses the x_i to be evenly spaced: $x_i = i/N$. Assume independent Gaussian priors for the parameters w_k, so that they are distributed as $\mathcal{N}(0, I)$. Also assume that $N > K$.

(a) Find the exact expression for the evidence of the data under model H_K.

(b) Find the ML estimator, $\hat{\mathsf{w}}_{ML}$, under H_K, and determine the corresponding log-likelihood.

Generate 750 samples $\mathbf{y} = (y_1, \ldots, y_{750})$ under H_2 with $\sigma^2 = 1$, $\mathsf{w}_0 = 0, \mathsf{w}_1 = 1, \mathsf{w}_2 = 1/3$. Use this vector and σ^2 for the rest of the question.

(c) Let $N = 20$. Get the first N samples from $\mathbf{y}$. Use these samples to plot, on the same plot, as a fuction of K, the normalized logarithm of the evidence, i.e. $1/N$ times the logarithm of the evidence, the normalized log-likelihood achieved by the ML estimator, and the normalized BIC without higher order terms as seen below. Vary K from $0 \le K \le 10$.

$$\text{Under BIC:} \quad \frac{1}{N} \log\ p_{\mathsf{y}}^N(y) \approx \frac{1}{N} \log\ p^N\left(y; \hat{\mathsf{w}}_{ML}^{H_k}\right) - \frac{K}{2N} \log\left(\frac{N}{2\pi}\right)$$

(d) Repeat part (c) for $N = 100$ and $N = 750$. You should now have a total of 3 plots. Looking over the 3 plots, how does the relationship between the three estimators change as N increases? This is a system with $K = 2$, does the ML estimator produce a sharp peak at $K = 2$, why, why not? Is there peaking noticed for other values of K?

(e) Now, plot, against N, the difference between the normalized log evidence and the normalized log likelihood. Do this on the same figure for $K = 1, 2, 3, 4$ each against $N = 10, 20, 50, 100, 250, 500, 750$.

(f) On the same plot as that produced in part (e), repeat part (e), but this time plot the difference between the normalized log evidence and the normalized BIC.

Problem 9.4

(a) Suppose that $\mathbf{y}^N = (\mathsf{y}_1, \ldots, \mathsf{y}_N)$ is a set of i.i.d. binary random variables drawn from a common distribution $p_{\mathsf{y}} = [0.9, 0.1]$. As usual, we let

$$\hat{p}(a; \mathbf{y}^N) \triangleq \frac{1}{N} \sum_{n=1}^{N} \mathbb{1}_{\{y_n = a\}}$$

be the empirical distribution of the sequence $\mathbf{y}^N$. Find the probability to first order in the exponent such that the entropy of the empirical distribution satisfies

$$H(\hat{p}(\cdot; \mathbf{y}^N)) \geq 0.5.$$

You can see that exponent is given by a convex program. Solve for the exponent numerically (e.g., in MATLAB).

(b) Let $p_{\mathsf{x},\mathsf{y}} \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ be a joint probability mass function where $\mathcal{X}$ and $\mathcal{Y}$ are finite sets. Further assume that $p_{\mathsf{x},\mathsf{y}} = p_{\mathsf{x}} p_{\mathsf{y}}$ is a product distribution. Let $(\mathbf{x}^N, \mathbf{y}^N) = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ be a set of N i.i.d. samples drawn from $p_{\mathsf{x},\mathsf{y}}$. Denote

$$\hat{p}_{\mathsf{x},\mathsf{y}}(a, b; \mathbf{x}^N, \mathbf{y}^N) \triangleq \frac{1}{N} \sum_{l=1}^{N} \mathbb{1}_{\{x_l = a, y_l = b\}}.$$

as the joint type, which will be abbreviated as $\hat{p}_{\mathsf{x},\mathsf{y}}(a, b)$. Also define

$$I(\hat{p}_{\mathsf{x},\mathsf{y}}) \triangleq \sum_{(a,b) \in \mathcal{X} \times \mathcal{Y}} \hat{p}_{\mathsf{x},\mathsf{y}}(a, b) \log \frac{\hat{p}_{\mathsf{x},\mathsf{y}}(a, b)}{\hat{p}_{\mathsf{x}}(a) \hat{p}_{\mathsf{y}}(b)}$$

as the mutual information of the empirical distribution. Fix $\epsilon > 0$. Find the probability to first order in the exponent so that

$$I(\hat{p}_{\mathsf{x},\mathsf{y}}) \geq \epsilon.$$

You may leave the exponent as an optimization problem.

Problem 9.5

Let $\mathsf{y}_1, \mathsf{y}_2, \ldots, \mathsf{y}_N$ be i.i.d. samples from the pdf $q(y)$, where $q(\cdot)$ and its derivative $q'(\cdot)$ is bounded and continuous on $(0, 1]$. Moreover, we assume the second derivative $q''(\cdot)$ exists and is also bounded on $(0, 1]$. Given a realization $(\mathsf{y}_1, \ldots, \mathsf{y}_N) = (y_1, \ldots y_N) = \mathbf{y}$, we wish to model q, which does not admit a parametric form whose parameters can be estimated.

Histograms are a useful way to model such distributions. We use H_m to denote the class of histogram models with m bins, i.e., under H_m the models take the form

$$p^m(y; \mathbf{x}^m) = x_k^m, \quad \text{for } y \in \left(\frac{k-1}{m}, \frac{k}{m}\right], \qquad \text{for } k = 1, 2, \ldots, m,$$

where $\mathbf{x}^m = (x_1^m, \ldots, x_m^m)$. It can easily be shown that the maximum likelihood (ML) estimate $\hat{x}_k^m(\mathbf{y})$ of x_k^m takes the form

$$\hat{x}_k^m(\mathbf{y}) = \frac{m n_k^m(\mathbf{y})}{N}, \qquad k = 1, 2, \ldots, m,$$

where $n_k^m(\mathbf{y})$ is the number of the y_i's that are in $((k-1)/m, k/m]$. So $\hat{\mathsf{x}}^m = \hat{\mathsf{x}}^m(\mathsf{y}) = (\hat{x}_1^m(\mathsf{y}), \ldots, \hat{x}_m^m(\mathsf{y}))$ denotes the vector of ML estimates.

In this problem we analyze some aspects of the asymptotics of histogram estimation and attempt to find a good choice of model H_m, for data of length N.

(a) Show that for any $z \in (0, 1]$, the bias in the estimate of $q(z)$ is of the form

$$\mathbb{E}\left[p^m(z; \hat{\mathsf{x}}^m)\right] - q(z) = O(\frac{1}{m})q'(z) + o\left(\frac{1}{m}\right), \qquad m \to \infty, \tag{1}$$

for some a (that does not depend on m or z), where $O(1/m)$ denotes terms that decay no slower than $1/m$, and $o(1/m)$ denotes terms that decay faster than $1/m$.

Hints: You may find one or both of the following facts useful:

- A continuous function $f(x)$ can be expanded as:

$$f(x) = f(a) + (x-a)f'(a) + \frac{(x-a)^2}{2}f''(a) + o((x-a)^2),$$

when all the required derivatives exist.

- As $m \to \infty$, the bin width approaches 0.

(b) Determine the constant b (that does not depend on m or N) such that when m is large enough that the bin width is effectively 0, the variance of the histogram estimate at z is given by

$$\mathrm{var}\left(p^m(z; \hat{\mathsf{x}}^m)\right) = b\frac{m}{N}q(z) + \frac{1}{N}O_m(1), \qquad m \to \infty, \tag{2}$$

where $O_m(1)$ denotes terms that do not grow with m and that do not depend on N.

(c) Let m be large enough that the $o(1/m)$ term of (1) is negligible, and N be large enough that the $O_m(1)/N$ term of (2) is neglible (we cannot neglect the $q(z)\, b\, m/N$ term without prescribing how m and N scale with respect to one another).

In this case, the *total* mean-square modeling error, given by

$$\mathrm{MSE} = \int_0^1 \mathbb{E}\left[(p^m(z; \hat{\mathbf{x}}^m) - q(z))^2 \right] \,\mathrm{d}z, \tag{3}$$

is minimized for large N when m scales with N according to $m = \alpha N^{\beta}$, for some $\alpha > 0$ and $0 < \beta < 1$ that do not depend on N. Determine β.

(d) Determine functions $g_{\mathrm{BIC}}(\cdot)$ and $g_{\mathrm{AIC}}(\cdot)$ so that:

(i) when $m = \gamma_{\mathrm{BIC}}\, g_{\mathrm{BIC}}(N)$, H_m optimizes the Bayes Information Criterion.

(ii) when $m = \gamma_{\mathrm{AIC}}\, g_{\mathrm{AIC}}(N)$, H_m optimizes the Akaike Information Criterion.

Here γ_{BIC} and γ_{AIC} are constants independent of N. You may directly use that, for large N, these respective criteria choose the model order so as to maximize the score functions

$$\begin{aligned} L_{\mathrm{BIC}}(H_m) &= \frac{1}{N} L_{\mathbf{y}}^N(\hat{\mathbf{x}}^m, H_m) - \frac{K_m}{2} \frac{\log N}{N} \\ L_{\mathrm{AIC}}(H_m) &= \frac{1}{N} L_{\mathbf{y}}^N(\hat{\mathbf{x}}^m, H_m) - \frac{K_m}{N} \end{aligned}$$

where K_m is the number of parameters that has to be estimated for model H_m, and the $L_{\mathbf{y}}^N(\hat{\mathbf{x}}^m, H_m)$ are log likelihoods of model H_m when the ML parameters are used in the model.

Note: Part (e) may be attempted independently of part (c), assuming a value of $\beta \in (0, 1)$.

(e) Consider the following candidate scalings for m with N:

$$\text{(I) } m \sim N^{\beta}, \qquad \text{(II) } m \sim g_{\mathrm{BIC}}(N), \qquad \text{(III) } m \sim g_{\mathrm{AIC}}(N)$$

where β is defined in (c) and $g_{\mathrm{BIC}}(\cdot)$ and $g_{\mathrm{AIC}}(\cdot)$ in (d).

(i) For each of the models, state whether or not the mean-square modeling error (3) decays to 0 as $N \to \infty$.

(ii) Order the models in the decreasing order of the rate at which each of bias and variance decay as $N \to \infty$ (give a separate order for each of bias and variance).