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

Problem Set 5

Issued: Tuesday, March 10, 2015 **Due:** Friday, March 20, 2015

Problem 5.1

In class we developed the EM algorithm for maximum likelihood estimation (EM-ML). That is, we gave an iterative procedure to compute

$$\hat{\mathbf{x}}_{\mathrm{ML}}(\mathbf{y}) = \arg\max_{\mathbf{a}} p_{\mathbf{y}}(\mathbf{y};\mathbf{a})$$

and showed that the likelihood was non-decreasing with each iteration.

Develop the EM-MAP algorithm for MAP estimation:

$$\hat{\mathbf{x}}_{\mathrm{MAP}}(\mathbf{y}) = \arg\max_{\mathbf{a}} p_{\mathbf{x}|\mathbf{y}}(\mathbf{a}|\mathbf{y}),$$

where the complete data z is an arbitrary random vector.

Hint: Make use of the decomposition $p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) = p_{\mathbf{z},\mathbf{x}|\mathbf{y}}(\mathbf{z},\mathbf{x}|\mathbf{y})/p_{\mathbf{z}|\mathbf{x},\mathbf{y}}(\mathbf{z}|\mathbf{x},\mathbf{y})$.

Problem 5.2

Suppose w is a Gaussian random variable with mean μ and unit variance. We observe y_i for $i = 1, 2$. Conditioned on $\mathsf{w} = w$, y_i is a Gaussian random variable with mean w and variance σ_i^2 for $i = 1, 2$. Moreover, y_1 and y_2 are conditionally independent given w.

We want to use the EM algorithm to obtain maximum likelihood estimates of the parameters $\mathbf{x} = (\mu, \sigma_1^2, \sigma_2^2)$ based on the observations $\mathsf{y}_1 = y_1$ and $\mathsf{y}_2 = y_2$. For our algorithm, we choose the complete data $\mathsf{z} = (\mathsf{y}_1, \mathsf{y}_2, \mathsf{w})$.

(a) Obtain an explicit expression for the function

$$U(\mathbf{x}, \mathbf{x}') = \mathbb{E}\left[\log p_{\mathbf{z}}(\mathbf{z};\mathbf{x}) \mid \mathsf{y}_1 = y_1, \mathsf{y}_2 = y_2; \mathbf{x}'\right]$$

in terms of the two functions

$$\alpha(\mathbf{x}') = \mathbb{E}\left[\mathsf{w}|\mathsf{y}_1 = y_1, \mathsf{y}_2 = y_2; \mathbf{x}'\right]$$
$$\beta(\mathbf{x}') = \mathrm{var}[\mathsf{w}|\mathsf{y}_1 = y_1, \mathsf{y}_2 = y_2; \mathbf{x}'].$$

Note that, as our notation implies, the expectations and variance are with respect to the distribution $p_{\mathbf{z}|y_1,y_2}(\mathbf{z}|y_1, y_2; \mathbf{x}')$. Note also that you are not being asked to obtain expressions for $\alpha(\mathbf{x}')$ and $\beta(\mathbf{x}')$.

(b) Determine the parameter vector $\mathbf{x}$ that maximizes $U(\mathbf{x}, \mathbf{x}')$, i.e.,

$$\hat{\mathbf{x}} = \arg\max_{\mathbf{x}} U(\mathbf{x}, \mathbf{x}'),$$

and express your answer in terms of the functions $\alpha(\mathbf{x}')$ and $\beta(\mathbf{x}')$.

(c) Starting from an initial guess $\hat{\mathbf{x}}^{(0)}$ for the parameters, give a full description of the EM algorithm for obtaining subsequent estimates $\hat{\mathbf{x}}^{(l)}$ for $l = 1, 2, \ldots$ by characterizing the E-step and M-step in terms of $\alpha(\mathbf{x}')$ and $\beta(\mathbf{x}')$.

Problem 5.3

A study tracks the health of a population of N people over a time period of T years. All members of the population start off healthy, and by the end of the study some have contracted an infectious disease at least once. Let y_i be a Bernoulli random variable taking on the values $\{0, 1\}$ that indicates whether subject i has ever contracted the disease by the end of the study (note that this subject could have been infected more than once). Put $\mathbf{y} = (\mathsf{y}_1, \ldots, \mathsf{y}_N)$. We also know whether or not each subject has been exposed to the disease each year. Let $\mu_{ij} = 1$ if subject i has been exposed to the disease during year j, and $\mu_{ij} = 0$ otherwise.

Our goal is to estimate the infection rate r_j for each year ($j = 1, 2, \ldots, T$) of the study. Put $\mathbf{r} = (r_1, \ldots, r_T)$.

(a) Compute π_i, the probability that subject i is infected during at least one exposure. Assume the probability that an exposed subject is infected during year j is r_j, independent of what happens during any other year of the study. Also note that π_i is a function of the μ_{ij} and r_j.

(b) Note that

$$p_{\mathsf{y}_i}(y_i; \mathbf{r}) = \pi_i^{y_i}(1 - \pi_i)^{(1-y_i)},$$

when confined to $y_i \in \{0, 1\}$, is the probability mass function for y_i. Write down $\log p_{\mathbf{y}}(\mathbf{y}; \mathbf{r})$, and express $\hat{\mathbf{r}}_{\mathrm{ML}}(\mathbf{y})$ as the result of a maximization problem (don't try to solve it). Assume that the subjects under study are distributed throughout a much larger population, so that the chance of one being infected is independent of any of the others being infected. You should find it very difficult to work this maximization by hand as N gets large. We're going to use the EM algorithm to solve it.

(c) Say we had more information. Let z_{ij} be a Bernoulli random variable taking on the values $\{0, 1\}$ that indicates whether subject i was infected at some point

during year j. Put $\mathbf{Z} = [z_{ij}]$. Show that, for a matrix $\mathbf{Z}$ that is valid for all the settings in this problem,

$$\log p_{\mathsf{Z}}(\mathbf{Z}; \mathbf{r}) = \sum_{i=1}^{N} \sum_{j=1}^{T} \left(\mu_{ij} z_{ij} \log r_j + \mu_{ij} (1 - z_{ij}) \log(1 - r_j) \right).$$

(d) Find $U(\mathbf{r}, \mathbf{r}_{\text{old}}) = \mathbb{E}[\log p_{\mathsf{Z}}(\mathbf{Z}; \mathbf{r}) \mid \mathbf{y} = \mathbf{y}; \mathbf{r}_{\text{old}}]$.

(e) Determine $\mathbf{r}_{\text{new}} = \arg\max_{\mathbf{r}} U(\mathbf{r}, \mathbf{r}_{\text{old}})$. Does your expression for this refined estimate of $\mathbf{r}$ make intuitive sense?

(f) Describe how to use your results from parts (d) and (e) to implement the EM algorithm. Are there any cases where you might divide by zero? What should you do instead?

(g) Estimate $\mathbf{r}$ given the data on the course web site. This hypothetical study was done over 10 years with 100 subjects. In what year was the infection rate the worst? Were there any years during which the disease might not have spread at all?

Problem 5.4 (practice)

In the context of the EM-ML algorithm, say that the complete data has a model that is in the canonical exponential family:

$$\log p_{\mathsf{z}}(z; x) = xz - \alpha(x) + \beta(z).$$

In the lecture, we've seen that the algorithm will converge to a fixed point. Let $e^{(n)} = \hat{x}^{(n)} - \hat{x}$ denote the difference between the current estimate at the n^{th} iteration $\hat{x}^{(n)}$ and the estimate at the fixed point $\hat{x}$.

(a) The Taylor expansion of $\mathbb{E}[\mathsf{z}; x]$ about $\hat{x}$ can be represented as

$$\mathbb{E}[\mathsf{z}; x] = \mathbb{E}[\mathsf{z}; \hat{x}] + f(\hat{x})(x - \hat{x}) + o((x - \hat{x})^2).$$

Find $f(\hat{x})$ and express it in terms of the Fisher information $J_{\mathsf{z}}(\cdot)$.

(b) The Taylor expansion of $\mathbb{E}[\mathsf{z} \mid \mathsf{y} = y; x]$ about $\hat{x}$ can be represented as

$$\mathbb{E}[\mathsf{z} \mid \mathsf{y} = y; x] = \mathbb{E}[\mathsf{z} \mid \mathsf{y} = y; \hat{x}] + g(\hat{x})(x - \hat{x}) + o((x - \hat{x})^2).$$

Find $g(\hat{x})$ and express it in terms of the Fisher information $J_{\mathsf{z}|\mathsf{y}=y}(\cdot)$.

(c) Determine $\rho(\hat{x})$ such that $e^{(n)} \approx \rho(\hat{x}) e^{(n-1)}$ for sufficiently small $e^{(0)}$ ($n \in \mathbb{Z}^+$).

You can obtain this result by continuing the style of analysis used in lecture to characterize the fixed point. Does your answer make intuitive sense?

Problem 5.5

(a) Show that the quadratic cost function

$$C(x, q) = A \sum_a (q(a) - \mathbb{1}_x(a))^2 + B(x),$$

where $A > 0$, is proper.

(b) Prove, as our theorem about the uniqueness of local proper cost functions implies, that the linear cost function $C(x, q) = -q(x)$ is not proper.

Problem 5.6

(a) Prove that $I(x; y) = 0$ if and only if x and y are independent random variables. Note that as a special case this implies that $I(x; y \mid z) = 0$ if and only if x and y are independent conditioned on z.

(b) Prove the chain rule for mutual information: $I(x; y, z) = I(x; z) + I(x; y \mid z)$ for any random variables x, y, and z. Note that there are two different such expansions by the symmetry of y and z in $I(x; y, z)$.

(c) Use the results above to prove the data processing inequality: if $x \leftrightarrow y \leftrightarrow t$ forms a Markov chain, then $I(x; t) \leq I(x; y)$.

(d) Say $x \leftrightarrow y \leftrightarrow t$ forms a Markov chain. Prove that the data processing inequality holds with equality if and only if $x \leftrightarrow t \leftrightarrow y$ also forms a Markov chain.

Problem 5.7

In this problem, we develop two alternative versions of the data processing inequality (DPI). Let $\mathcal{P}(\mathcal{X})$ be the set of distributions (probability simplex) over alphabet $\mathcal{X}$, and let x and y be discrete random variables with distributions p_x and p_y respectively. $p_x, p_y \in \mathcal{P}(\mathcal{X})$. Assume that $|\mathcal{X}| < \infty$ and $p_x(a), p_y(a) > 0$ for all $a \in \mathcal{X}$.

Consider processing x and y by the same arbitrary (randomized) function, producing x' and y' with distributions $q_{x'} \in \mathcal{P}(\mathcal{X})$ and $q_{y'} \in \mathcal{P}(\mathcal{X})$, respectively. This random processing can be expressed in the form

$$q_{x'}(b) = \sum_{a \in \mathcal{X}} W(b \mid a) p_x(a), \qquad q_{y'}(b) = \sum_{a \in \mathcal{X}} W(b \mid a) p_y(a), \qquad \text{for all } b \in \mathcal{X},$$

where the (known) conditional distribution $W(\cdot \mid \cdot)$ represents the processing.

(a) Prove the information divergence version of the DPI:

$$D(q_{\mathsf{x}'} \parallel q_{\mathsf{y}'}) \leq D(p_{\mathsf{x}} \parallel p_{\mathsf{y}}) \quad \text{for all } W(\cdot \mid \cdot). \tag{1}$$

Hint: You may use the fact that for two positive finite-length sequences $\{a_i\}, \{b_i\}$,

$$\sum_i a_i \log \frac{a_i}{b_i} \geq \left(\sum_i a_i\right) \log \frac{\sum_i a_i}{\sum_i b_i}, \qquad \text{(log-sum inequality)}$$

with equality if and only if the ratio a_i/b_i is constant for all i.

(b) Let $\mathsf{x}' = g(\mathsf{x})$ and $\mathsf{y}' = g(\mathsf{y})$ for some deterministic function $g(\cdot)$. Assuming $p_{\mathsf{x}} \neq p_{\mathsf{y}}$, specify a *necessary* and *sufficient* condition on $g(\cdot)$ and on the pair of distributions $(p_{\mathsf{x}}, p_{\mathsf{y}})$ such that (1) holds with equality.

In the remainder of the problem, let (y, z) be a pair of discrete random variables with joint distribution $p_{\mathsf{y},\mathsf{z}}(y, z; x) \in \mathcal{P}(\mathcal{Y} \times \mathcal{Z})$, which is parameterized by a scalar x and is positive everywhere. Assume all derivatives of $p_{\mathsf{y},\mathsf{z}}(y, z; x)$ with respect to x exist and the usual regularity conditions hold.

(c) (**Practice**) Prove the chain rule for Fisher information:

$$J_{\mathsf{y},\mathsf{z}}(x) = J_{\mathsf{y}}(x) + J_{\mathsf{z}|\mathsf{y}}(x),$$

where

$$J_{\mathsf{z}|\mathsf{y}}(x) \triangleq \mathbb{E}_{\mathsf{y},\mathsf{z}}\left[\left(\frac{\partial \log p_{\mathsf{z}|\mathsf{y}}(z \mid y; x)}{\partial x}\right)^2\right] = -\,\mathbb{E}_{\mathsf{y},\mathsf{z}}\left[\frac{\partial^2 \log p_{\mathsf{z}|\mathsf{y}}(z \mid y; x)}{\partial x^2}\right].$$

Prove the DPI for Fisher information: if $\mathsf{w} \triangleq \phi(\mathsf{z})$ where $\phi(\cdot)$ is deterministic, then

$$J_{\mathsf{w}}(x) \leq J_{\mathsf{z}}(x). \tag{2}$$

(d) (**Practice**) Determine a *necessary* and *sufficient* condition on $\phi(\cdot)$ such that (2) holds with equality.

Problem 5.8

Show that for a model $p_{\mathsf{y}}(y; x)$,

$$\lim_{\delta \to 0} \frac{D(p_{\mathsf{y}}(\cdot; x) \,\|\, p_{\mathsf{y}}(\cdot; x + \delta))}{\delta^2} = \frac{1}{2} J_{\mathsf{y}}(x),$$

so that for small δ we can make the approximation

$$D(p_{\mathsf{y}}(\cdot; x) \,\|\, p_{\mathsf{y}}(\cdot; x + \delta)) \approx \frac{\delta^2}{2} J_{\mathsf{y}}(x),$$

where $J_{\mathsf{y}}(x)$ is the Fisher information in y about x:

$$J_{\mathsf{y}}(x) = -\mathbb{E}\left[\frac{\partial^2}{\partial x^2}\log p_{\mathsf{y}}(\cdot; x)\right].$$

Assume all the logarithms are base e for convenience.

Problem 5.9 (practice)
Show that for the case of a discrete random vector $\mathbf{y}$ with N iid components each generated according to $p_{\mathsf{y}}(\cdot; x)$, the ML estimate of x can be expressed in the form

$$\hat{x}_{\mathrm{ML}} = \arg\min_x D(\hat{p}_{\mathsf{y}}(\cdot; \mathbf{y}) \,\|\, p_{\mathsf{y}}(\cdot; x)),$$

where $\hat{p}_{\mathsf{y}}(\cdot; \mathbf{y})$ is the empirical distribution of y based on $\mathbf{y}$, i.e. $\hat{p}_{\mathsf{y}}(b; \mathbf{y})$ is the fraction of times the symbol b appears in the sequence $\mathbf{y} = (y_1, \ldots, y_N)$.

Problem 5.10 (practice)
Suppose we are deciding between two hypotheses $\mathsf{H} \in \{H_0, H_1\}$ based on continuous-valued observations y governed by the (bounded) model $p_{\mathsf{y}|\mathsf{H}}$, where $p \triangleq \mathbb{P}(\mathsf{H} = H_1)$ satisfies $0 < p < 1$. We consider the decision rule $\hat{\mathsf{H}}(\mathsf{y})$ that maximizes the mutual information between the decision and the correct hypothesis, i.e.,

$$\hat{\mathsf{H}}(\mathsf{y}) = \arg\max_{f \in \mathcal{F}} I(f(\mathsf{y}); \mathsf{H}), \tag{3}$$

where $\mathcal{F}$ is the set of all possible decision rules, and the mutual information $I(\hat{\mathsf{H}}; \mathsf{H})$ is defined as

$$I(\hat{\mathsf{H}}; \mathsf{H}) = \sum_{i=0}^{1}\sum_{j=0}^{1} p_{\hat{\mathsf{H}},\mathsf{H}}(H_i, H_j)\log\frac{p_{\hat{\mathsf{H}},\mathsf{H}}(H_i, H_j)}{p_{\hat{\mathsf{H}}}(H_i)p_{\mathsf{H}}(H_j)}. \tag{4}$$

(a) Determine (in terms of p) coefficients w_{D} and w_{F} such that the mutual information $I(\hat{\mathsf{H}}; \mathsf{H})$ can be expressed in the form

$$I(\hat{\mathsf{H}}; \mathsf{H}) = H_{\mathrm{B}}(w_{\mathrm{D}}P_{\mathrm{D}} + w_{\mathrm{F}}P_{\mathrm{F}}) - w_{\mathrm{D}}H_{\mathrm{B}}(P_{\mathrm{D}}) - w_{\mathrm{F}}H_{\mathrm{B}}(P_{\mathrm{F}}), \tag{5}$$

where

$$H_{\mathrm{B}}(\epsilon) = -\epsilon\log(\epsilon) - (1-\epsilon)\log(1-\epsilon), \quad \epsilon \in [0, 1],$$

is the binary entropy function, and where

$$P_{\mathrm{D}} = \mathbb{P}\left(\hat{\mathsf{H}} = H_1 \mid \mathsf{H} = H_1\right) \quad \text{and} \quad P_{\mathrm{F}} = \mathbb{P}\left(\hat{\mathsf{H}} = H_1 \mid \mathsf{H} = H_0\right).$$

(b) Show that (5), when viewed as a function of P_{D} with $P_{\mathrm{F}} = P'_{\mathrm{F}}$ fixed, increases monotonically with P_{D} for $P_{\mathrm{D}} > P'_{\mathrm{F}}$.

(c) Argue that the decision rule that optimizes (3) is a likelihood ratio test, i.e.,

$$\frac{p_{\mathsf{y}|\mathsf{H}}(y \mid H_1)}{p_{\mathsf{y}|\mathsf{H}}(y \mid H_0)} \underset{\hat{H}=H_0}{\overset{\hat{H}=H_1}{\gtrless}} \eta. \tag{6}$$

for some threshold η. Note: you are not being asked to determine η.

(d) Determine whether the following statement is true or false: the above "maximum mutual information" decision rule is in general the same as the rule that minimizes the probability of a decision error. Be sure to justify your answer.

Hint: You may find it useful that the optimal threshold η in (6) satisfies

$$\eta = -\frac{(1-p)\log\frac{(1-\alpha)P_{\mathrm{F}}}{\alpha(1-P_{\mathrm{F}})}}{p\log\frac{(1-\alpha)P_{\mathrm{D}}}{\alpha(1-P_{\mathrm{D}})}},$$

where as a reminder P_{D} and P_{F} are implicit functions of η, and where $\alpha \triangleq \mathbb{P}\left(\hat{\mathsf{H}} = H_1\right)$.