Homework 5, due Monday April 27

COMS 4771 Spring 2015

Problem 1 (Expectation-Maximization)**.** Consider the following variant of the MTurk model for m items and n workers.

- Nature picks correct label for item i to be 1 with probability π_i (and 0 otherwise).
- Ask each worker to label each item as 0 or 1.
- If the correct label for item i is 0, then worker j responds with correct label on item i with probability p_j.

  If the correct label for item i is 1, then worker j responds with correct label on item i with probability r_j.
- All choices of Nature are independent, and all the responses of the workers are conditionally independent given the choices of Nature.

The parameters of the model are $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{p}, \boldsymbol{r}) = (\pi_1, \pi_2, \ldots, \pi_m, p_1, p_2, \ldots, p_n, r_1, r_2, \ldots, r_n) \in (0,1)^{m+2n}$, and the random variables involved in the model are distributed as follows.

- (Hidden) Y_i is the correct label for item i;

$$\Pr_{\boldsymbol{\theta}}(Y_i = 0) = 1 - \pi_i, \quad \Pr_{\boldsymbol{\theta}}(Y_i = 1) = \pi_i.$$

- (Observed) $X_{i,j}$ is the response given by worker j for item i;

$$\Pr_{\boldsymbol{\theta}}(X_{i,j} = 0 \,|\, Y_i = 0) = p_j, \qquad \Pr_{\boldsymbol{\theta}}(X_{i,j} = 1 \,|\, Y_i = 0) = 1 - p_j,$$
$$\Pr_{\boldsymbol{\theta}}(X_{i,j} = 0 \,|\, Y_i = 1) = 1 - r_j, \qquad \Pr_{\boldsymbol{\theta}}(X_{i,j} = 1 \,|\, Y_i = 1) = r_j.$$

Derive an E-M algorithm for estimating the parameters given the observed responses. Explain your derivation, and then clearly describe the E and M steps.

Problem 2 (Maximum entropy)**.** You are given a six-sided die and initially believe that it is a fair die (i.e., when rolled, the numbers $\{1, 2, \ldots, 6\}$ are all equally likely to come up).

(a) Upon several independent rolls of the die, you observe the following.

   i. The fraction of times that the die comes up with 4 is 0.2.

   ii. The fraction of times that the die comes up with a number ≤ 3 is 0.2.

   What is the maximum entropy distribution $(p_1, p_2, \ldots, p_6)$ over $\{1, 2, \ldots, 6\}$ that agrees with both of these observations? (Give actual real values for each p_i.) Explain all the steps you take to arrive at your solution.

(b) Repeat (a), but instead using the following observations.

   i. The fraction of times that the die comes up with either 1 or 2 is 0.5.

   ii. The fraction of times that the die comes up with an even number is 0.5.

Problem 3 (Practice problems)**.** The following problems will not be graded.

1. Consider the experiment where you independently toss m balls into two possible bins. The bins have varying widths $w_1, w_2 > 0$, and the probability that a ball lands into a particular bin is proportional to the width of that bin. Without loss of generality, assume that $w_1 + w_2 = 1$. The outcome of this process can be described by a random pair of ball counts for each bin: $\boldsymbol{X} = (X_1, X_2) \in \{0, 1, 2, \ldots, m\}^2$ where $X_1 + X_2 = m$. Let $P_{\boldsymbol{w}}$ be the distribution of $\boldsymbol{X}$, where $\boldsymbol{w} = (w_1, w_2)$. Show that $\mathcal{P} := \{P_{\boldsymbol{w}} : \boldsymbol{w} \in (0,1)^2\}$ is an exponential family. To do this, specify the domain, the base measure, and the feature functions. Also specify the log partition function and the natural parameter space.

2. Suppose you train a classifer using AdaBoost,

$$x \mapsto \operatorname{sign}\left(\sum_{t=1}^{T} \alpha_t \cdot f_t(x)\right).$$

   You would like to predict the probability that a new test point x has label 1. How should you compute this prediction? (Assume you have access to the α_ts and the f_ts.)

3. Consider the *fixed-design setting* for linear regression:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{w}_\star + \boldsymbol{\varepsilon}$$

   where $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is a (non-random) matrix, $\boldsymbol{w}_\star \in \mathbb{R}^p$, and $\boldsymbol{\varepsilon}$ is a zero-mean random vector in $\mathbb{R}^n$ with independent entries and $\operatorname{cov}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$ for some $\sigma^2 > 0$.

   Recall that for $\lambda > 0$, the ridge regression estimator $\hat{\boldsymbol{w}}_\lambda$ satisfies

$$\operatorname{risk}(\hat{\boldsymbol{w}}_\lambda) := \mathbb{E}\left[\frac{1}{n}\|\boldsymbol{X}\boldsymbol{w}_\star - \boldsymbol{X}\hat{\boldsymbol{w}}_\lambda\|_2^2\right] = \sum_{j=1}^{p} \frac{\lambda_j \lambda^2}{(\lambda_j + \lambda)^2} \langle \boldsymbol{v}_j, \boldsymbol{w}_\star \rangle^2 + \frac{\sigma^2}{n} \sum_{j=1}^{p} \left(\frac{\lambda_j}{\lambda_j + \lambda}\right)^2,$$

   and the principal components regression estimator $\hat{\boldsymbol{w}}_{\mathrm{pc}\,\lambda}$ satisfies

$$\operatorname{risk}(\hat{\boldsymbol{w}}_{\mathrm{pc}\,\lambda}) := \mathbb{E}\left[\frac{1}{n}\|\boldsymbol{X}\boldsymbol{w}_\star - \boldsymbol{X}\hat{\boldsymbol{w}}_{\mathrm{pc}\,\lambda}\|_2^2\right] = \sum_{j=1}^{p} \mathbb{1}\{\lambda_j < \lambda\} \lambda_j \langle \boldsymbol{v}_j, \boldsymbol{w}_\star \rangle^2 + \frac{\sigma^2}{n} \sum_{j=1}^{p} \mathbb{1}\{\lambda_j \geq \lambda\}.$$

   Assume $\lambda_1 > \lambda_2 > \cdots > \lambda_p > 0$. Suppose you are told the least value $k \in \mathbb{N}$ such that $\langle \boldsymbol{v}_j, \boldsymbol{w}_\star \rangle = 0$ for all $j > k$. What value (or range of values) of $\lambda > 0$ should be used with principal components regression? And for such a setting of λ, how large can the ratio $\operatorname{risk}(\hat{\boldsymbol{w}}_\lambda) / \operatorname{risk}(\hat{\boldsymbol{w}}_{\mathrm{pc}\,\lambda})$ be? Give an answer in terms of k and p, and specify conditions on the λ_j so that the ratio is as large as possible.

4. In a time-homogeneous HMM, are the marginal distributions of the first observation X_1 and the second observation X_2 the same? Explain why or why not.