COMS 4771 Machine Learning (Spring 2015)
Problem Set #5

Solutions - `coms4771@machinelearning.nyc`
Discussants: `None`

April 28, 2015

Problem 1

Let $\boldsymbol{x} \in \{0,1\}^{m \times n}$ denote the matrix of labels provided by the workers.

Let $\hat{\boldsymbol{\theta}}$ be the current parameters. Define $q_i := \Pr_{\hat{\boldsymbol{\theta}}}(Y_i = 1 \mid \boldsymbol{X} = \boldsymbol{x})$ for each item $i \in [m]$. First, observe that independence assumptions imply that $q_i = \Pr_{\hat{\boldsymbol{\theta}}}(Y_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}_i)$, where $\boldsymbol{X}_i = (X_{i,1}, X_{i,2}, \dots, X_{i,n})$ and $\boldsymbol{x}_i = (x_{i,1}, x_{i,2}, \dots, x_{i,n})$. By Bayes' rule,

$$q_i = \frac{\Pr_{\hat{\boldsymbol{\theta}}}(\boldsymbol{X}_i = \boldsymbol{x}_i \mid Y_i = 1)\Pr_{\hat{\boldsymbol{\theta}}}(Y_i = 1)}{\Pr_{\hat{\boldsymbol{\theta}}}(\boldsymbol{X}_i = \boldsymbol{x}_i \mid Y_i = 0)\Pr_{\hat{\boldsymbol{\theta}}}(Y_i = 0) + \Pr_{\hat{\boldsymbol{\theta}}}(\boldsymbol{X}_i = \boldsymbol{x}_i \mid Y_i = 1)\Pr_{\hat{\boldsymbol{\theta}}}(Y_i = 1)}.$$

The numerator is

$$\Pr_{\hat{\boldsymbol{\theta}}}(\boldsymbol{X}_i = \boldsymbol{x}_i \mid Y_i = 1)\Pr_{\hat{\boldsymbol{\theta}}}(Y_i = 1) = \hat{\pi}_i \prod_{j=1}^{m} \hat{r}_j^{x_{i,j}} (1 - \hat{r}_j)^{1 - x_{i,j}}.$$

The term in the denominator that isn't the same as the numerator is

$$\Pr_{\hat{\boldsymbol{\theta}}}(\boldsymbol{X}_i = \boldsymbol{x}_i \mid Y_i = 0)\Pr_{\hat{\boldsymbol{\theta}}}(Y_i = 0) = \hat{\pi}_i \prod_{j=1}^{m} \hat{p}_j^{1 - x_{i,j}} (1 - \hat{p}_j)^{x_{i,j}}.$$

Therefore

$$q_i = \frac{\hat{\pi}_i \prod_{j=1}^{m} \hat{r}_j^{x_{i,j}} (1 - \hat{r}_j)^{1 - x_{i,j}}}{\hat{\pi}_i \prod_{j=1}^{m} \hat{r}_j^{x_{i,j}} (1 - \hat{r}_j)^{1 - x_{i,j}} + \hat{\pi}_i \prod_{j=1}^{m} \hat{p}_j^{1 - x_{i,j}} (1 - \hat{p}_j)^{x_{i,j}}}.$$

The complete log-likelihood of $\boldsymbol{\theta}$ given $\boldsymbol{x}$ and true labels $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_m)$ is

$$\sum_{i=1}^{m} \Biggl\{ Y_i \ln \pi_i + (1 - Y_i) \ln(1 - \pi_i) + \sum_{j=1}^{n} Y_i [x_{i,j} \ln r_j + (1 - x_{i,j}) \ln(1 - r_j)] \\ + (1 - Y_i)[(1 - x_{i,j}) \ln p_j + x_{i,j} \ln(1 - p_j)] \Biggr\}.$$

So the expected complete log-likelihood is

$$\sum_{i=1}^{m} \Bigg\{ q_i \ln \pi_i + (1 - Y_i) \ln(1 - \pi_i) + \sum_{j=1}^{n} q_i [x_{i,j} \ln r_j + (1 - x_{i,j}) \ln(1 - r_j)] \\ + (1 - q_i)[(1 - x_{i,j}) \ln p_j + x_{i,j} \ln(1 - p_j)] \Bigg\}.$$

The maximizing parameters are

$$\begin{aligned} \hat{\pi}_i &= q_i, \\ \hat{p}_j &= \frac{\sum_{i=1}^{m} (1 - q_i)(1 - x_{i,j})}{\sum_{i=1}^{m} (1 - q_i)}, \\ \hat{r}_j &= \frac{\sum_{i=1}^{m} q_i x_{i,j}}{\sum_{i=1}^{m} q_i}. \end{aligned}$$

So here are the E and M steps.

- E-step:

$$q_i := \frac{\hat{\pi}_i \prod_{j=1}^{m} \hat{r}_j^{x_{i,j}} (1 - \hat{r}_j)^{1 - x_{i,j}}}{\hat{\pi}_i \prod_{j=1}^{m} \hat{r}_j^{x_{i,j}} (1 - \hat{r}_j)^{1 - x_{i,j}} + \hat{\pi}_i \prod_{j=1}^{m} \hat{p}_j^{1 - x_{i,j}} (1 - \hat{p}_j)^{x_{i,j}}} \quad \forall i \in [m].$$

- M-step:

$$\begin{aligned} \hat{\pi}_i &= q_i \quad \forall i \in [m], \\ \hat{p}_j &= \frac{\sum_{i=1}^{m} (1 - q_i)(1 - x_{i,j})}{\sum_{i=1}^{m} (1 - q_i)} \quad \forall j \in [n], \\ \hat{r}_j &= \frac{\sum_{i=1}^{m} q_i x_{i,j}}{\sum_{i=1}^{m} q_i} \quad \forall j \in [n]. \end{aligned}$$

Problem 2

This problem is just a matter of writing down the log partition function $G(\boldsymbol{\eta})$, taking its derivatives, and then solving some equations.

(a) Our domain is $\mathcal{X} = \{1, 2, \ldots, 6\}$. Let the first feature function be $T_1(x) = \mathbb{1}\{x = 4\}$, and let the second feature function be $T_2(x) = \mathbb{1}\{x \leq 3\}$. I'm going to use $\pi(x) = 1$ as the base distribution. (Using $\pi(x) = 1/6$ will give the same result.)

Then the log partition function $G(\boldsymbol{\eta})$ is

$$G(\boldsymbol{\eta}) = \ln\big(e^{\eta_2} + e^{\eta_2} + e^{\eta_2} + e^{\eta_1} + e^0 + e^0\big) = \ln(e^{\eta_1} + 3e^{\eta_2} + 2).$$

Now take derivatives with respect to η_1 and η_2:

$$\frac{\partial G(\boldsymbol{\eta})}{\partial \eta_1} = \frac{e^{\eta_1}}{e^{\eta_1} + 3e^{\eta_2} + 2}, \quad \frac{\partial G(\boldsymbol{\eta})}{\partial \eta_2} = \frac{3e^{\eta_2}}{e^{\eta_1} + 3e^{\eta_2} + 2}.$$

We now just solve the system of equations

$$\frac{\partial G(\boldsymbol{\eta})}{\partial \eta_1} = \frac{e^{\eta_1}}{e^{\eta_1} + 3e^{\eta_2} + 2} = 0.2, \quad \frac{\partial G(\boldsymbol{\eta})}{\partial \eta_2} = \frac{3e^{\eta_2}}{e^{\eta_1} + 3e^{\eta_2} + 2} = 0.2$$

for η_1 and η_2. (I find it easier to solve for e^{η_1} and e^{η_2}.) Eventually you get $\eta_1 = -\ln(1.5)$ and $\eta_2 = -\ln(4.5)$. Now get $p_1, p_2, \ldots, p_6$ by plugging-in: you should get $\boldsymbol{p} = (1/15, 1/15, 1/15, 1/5, 3/10, 3/10)$.

(b) This is similar. You should get $\boldsymbol{p} = (1/4, 1/4, 1/8, 1/8, 1/8, 1/8)$.

Problem 3

1. This is the binomial distribution $\mathrm{Bin}(m, p)$ where $p = 1/(1 + e^{-\eta})$. The domain is $\mathcal{X} = \{(x_1, x_2) \in \mathbb{Z}_+^2 : x_1 + x_2 = m\}$. The base measure is $\pi(x_1, x_2) = \binom{m}{x_1}$, and the sole feature function is $T_1(x_1, x_2) = x_1$. The natural parameter space is $\mathcal{N} = \mathbb{R}$ and the log partition function is $G(\eta) = m \ln(1 + e^\eta)$.

2. Recall that AdaBoost can be interpreted as a descent algorithm for minimizing the exponential loss. Also, recall that for all x, the minimizer of $\hat{y} \mapsto \mathbb{E}[\ell_{\exp}(Y\hat{y})|X = x]$ is

$$\hat{y} = \frac{1}{2} \ln \frac{\eta(x)}{1 - \eta(x)}$$

   where $\eta(x) := \Pr[Y = +1|X = x]$.

   Therefore, for a given x, the prediction of $\eta(x)$ is

$$\frac{\exp(2g(x))}{1 + \exp(2g(x))}$$

   where $g(x) := \sum_{t=1}^{T} \alpha_t f_t(x)$.

3. We should pick a value of λ such that $\lambda_k \geq \lambda$, as this guarantees

$$\mathrm{risk}(\hat{\boldsymbol{w}}_{\mathrm{pc}\,\lambda}) = \frac{\sigma^2 k}{n},$$

   which goes to zero as $n \to \infty$. For such values of λ, the ratio $\mathrm{risk}(\hat{\boldsymbol{w}}_\lambda)/\,\mathrm{risk}(\hat{\boldsymbol{w}}_{\mathrm{pc}\,\lambda})$ is at least

$$\frac{1}{k} \sum_{j=1}^{p} \left(\frac{\lambda_j}{\lambda_j + \lambda} \right)^2.$$

   If all λ_j are at least λ, then the ratio is at least

$$\frac{1}{k} \sum_{j=k+1}^{p} \left(\frac{\lambda_j}{\lambda_j + \lambda} \right)^2 \geq \frac{p-k}{4k} = \Omega\Big(\frac{p}{k}\Big).$$

4. No. If $\boldsymbol{\pi}$ is the marginal distribution of Y_1, and $\boldsymbol{A}$ is the transition matrix, then $\boldsymbol{\nu} := \boldsymbol{A}^\top \boldsymbol{\pi}$ is the marginal distribution of Y_2. These need not be the same (unless $\boldsymbol{\pi}$ is a stationary distribution for the hidden state Markov chain), and hence the marginal distributions for X_1 and X_2 need not be the same. (If the conditional distribution of X_t given $Y_t = i$ is P_i, then the marginal distribution of X_1 is the mixture distribution $\pi_1 P_1 + \pi_2 P_2 + \cdots + \pi_k P_k$, while the marginal distribution of X_2 is $\nu_1 P_1 + \nu_2 P_2 + \cdots \nu_k P_k$.)

5. Only (d) is uniquely defined.

   (a) Let W_{pca} be the rank 10 PCA subspace for $\boldsymbol{X}$. Note that W_{pca} is a ten-dimensional subspace of $\mathbb{R}^d$. Each non-zero $\boldsymbol{v} \in W_{\mathrm{pca}}$ is an eigenvalue of $\boldsymbol{X}^\top \boldsymbol{X}$ with eigenvalue one—hence, each is a "top eigenvector" of $\boldsymbol{X}^\top \boldsymbol{X}$.

(b) Each non-zero $\boldsymbol{v} \in W_{\text{pca}}$ also determines a one-dimensional subspace which has the minimum squared reconstruction error among all one-dimensional subspaces.

(c) Any orthonormal basis for W_{pca} is a set of ten unit-length eigenvectors for $\boldsymbol{X}^\top \boldsymbol{X}$ which are mutually orthogonal and each has corresponding eigenvalue one—hence, it is a set of top 10 unit-length eigenvectors for $\boldsymbol{X}^\top \boldsymbol{X}$.

(d) Any other subspace of dimension ten (besides W_{pca}) must contain a unit-vector $\boldsymbol{u} \notin W_{\text{pca}}$ for which the empirical variance of the data points in direction $\boldsymbol{u}$ is less than one: indeed, the residual vector $\boldsymbol{r} := (\boldsymbol{I} - \boldsymbol{\Pi}_{W_{\text{pca}}})\boldsymbol{u} \neq \boldsymbol{0}$ is orthogonal to W_{pca}, and hence the empirical variance of the data points in direction $\boldsymbol{u}$ is $1 - \|\boldsymbol{r}\|_2^2 < 1$. Hence, such a subspace cannot minimize the squared reconstruction error among all ten-dimensional subspaces.

6. The ordinary least squares optimization problem is

$$\min_{\boldsymbol{w} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} \big(\langle \boldsymbol{w}, \boldsymbol{x}^{(i)} \rangle - y^{(i)} \big)^2 .$$

We start with some initial vector $\boldsymbol{w}^{(1)} \in \mathbb{R}^p$. Then for $t = 1, 2, \ldots, n$:

(a) Compute $\boldsymbol{\lambda}^{(t)} := 2\big(\langle \boldsymbol{w}^{(t)}, \boldsymbol{x}^{(\pi(i))} \rangle - y^{(\pi(i))} \big) \boldsymbol{x}^{(\pi(i))}$.

(b) Update $\boldsymbol{w}^{(t+1)} := \boldsymbol{w}^{(t)} - \eta_t \boldsymbol{\lambda}^{(t)}$.

Return $\boldsymbol{w}^{(n+1)}$.