ENGG 5202: Homework #1

Due on Thursday, March 3, 2016, 5:30pm

Xiaogang Wang

Problem 1

[**20 points**]

Consider a one-dimensional two-category classification problem with equal priors, $P(\omega_1) = P(\omega_2) = 1/2$, where the densities have the form

$$p(x|\omega_i) = \begin{cases} 0 & x < 0 \\ \frac{2}{\theta_i}\left(1 - \frac{x}{\theta_i}\right) & 0 \le x \le \theta_i \\ 0 & otherwise, \end{cases}$$

The following data were collected: $\mathcal{D}_1 = \{2, 5\}$ and $\mathcal{D}_2 = \{3, 9\}$ for ω_1 and ω_2 respectively. Find the maximum-likelihood values of $\hat{\theta}_1$ and $\hat{\theta}_2$.

Problem 2

[**20 points**]

Let x have an exponential density,

$$p(x|\theta) = \begin{cases} \theta e^{-\theta x}, & x \ge 0 \\ 0, & otherwise. \end{cases}$$

1. Plot $p(x|\theta)$ versus x for $\theta = 1$. Plot $p(x|\theta)$ versus θ, $(0 \le \theta \le 5)$, for $x = 2$.
2. Suppose that n samples $x_1, \ldots, x_n$ are drawn independently according to $p(x|\theta)$. Show that the maximum-likelihood estimate for θ is given by

$$\hat{\theta} = \frac{n}{\sum_{k=1}^{n} x_k}$$

3. If the samples are generated from $p(x|\theta)$ with $\theta = 1$, what's the maximum-likelihood estimate $\hat{\theta}$ for large n? Does $\hat{\theta}$ approach to the true θ when n is very large?

Problem 3

[**40 points**]

In this problem, we will study how to use the EM algorithm to estimate the parameters of a mixture model of Gaussian components. A mixture of m Gaussian components, for example, is a simple hidden or latent variable model,

$$p(\mathbf{x};\boldsymbol{\theta}) = \sum_{j=1}^{m} p_j N(\mathbf{x};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)$$

where $\sum_{j=1}^{m} p_j = 1$. The distinction between hidden and observed variables depends on the data we expect to see in estimating these models. In a typical setting, we observe only $\mathbf{x}$ samples and the corresponding choices of the mixture components remain hidden. Thus $\mathbf{x}$ is an observed (vector valued) variable, while j is hidden.

A direct optimization of latent variable models is often difficult since the likelihood of the data involves summations over the values of the hidden variables. Specifically, in the log-likelihood of the data for a mixture of m Gaussians

$$l(\mathbf{x}_1,\ldots,\mathbf{x}_n;\boldsymbol{\theta}) = \log \prod_{k=1}^{n} p(\mathbf{x}_k;\boldsymbol{\theta}) = \sum_{k=1}^{n} \log \sum_{j=1}^{m} p_j N(\mathbf{x}_k;\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)$$

the summation over the mixture components appear inside the logarithm, coupling the means and covariances of the Gaussian components.

Suppose for a moment that we also observed the selections of the mixture components, $z_1,\ldots,z_n$, in addition to $\mathbf{x}_1,\ldots,\mathbf{x}_n$. In this case we could write down the complete log-likelihood of the data

$$l_c(\mathbf{x}_1,\ldots,\mathbf{x}_n,z_1,\ldots,z_n;\theta) = \sum_{k=1}^{n} \log(p_{z_k} N(\mathbf{x}_k;\boldsymbol{\mu}_{z_k},\boldsymbol{\Sigma}_{z_k}))$$

where each sample $(\mathbf{x}_k, z_k)$ contains a value assignment to all the variables in the model. The means and covariances of the Gaussians, along with the mixing proportions, could now be estimated independently of each other.

Given only $\mathbf{x}_1,\ldots,\mathbf{x}_n$, we can nevertheless use the current setting of the model parameters, $\boldsymbol{\theta}^{(t)}$, to infer what the mixture selections would have been. In other words, we can evaluate the expected complete log-likelihood:

$$\begin{aligned} Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)}) &= \mathcal{E}\left\{ l_c(\mathbf{x}_1,\ldots,\mathbf{x}_n,z_1,\ldots,z_n;\boldsymbol{\theta}) | \mathbf{x}_1,\ldots,\mathbf{x}_n,\boldsymbol{\theta}^{(t)} \right\} \\ &= \sum_{k=1}^{n} \mathcal{E}\left\{ \log\left(p_{z_k} N(\mathbf{x}_k;\boldsymbol{\mu}_{z_k},\boldsymbol{\Sigma}_{z_k})\right) | \mathbf{x}_k,\boldsymbol{\theta}^{(t)} \right\} \\ &= \sum_{k=1}^{n}\sum_{j=1}^{m} P(z_k = j|\mathbf{x}_k,\boldsymbol{\theta}^{(t)}) \log\left(p_j N\left(\mathbf{x}_k;\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j\right)\right) \end{aligned}$$

where the expectations are over $z_1,\ldots,z_n$ given $\boldsymbol{\theta}^{(t)}$ and $\mathbf{x}_1,\ldots,\mathbf{x}_n$. $Q(\boldsymbol{\theta};\boldsymbol{\theta}^{(t)})$ should be viewed as a function of possible new settings of parameters

Problem 3 continued on next page...

$$\boldsymbol{\theta} = \{p_1, \ldots, p_m, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_m\}$$

for a fixed $\boldsymbol{\theta}^{(t)}$ and $\mathbf{x}_1, \ldots, \mathbf{x}_n$ (which we have suppressed in the notation). Evaluating the expected complete log-likelihood for a hidden variable model as a function of possible new parameters $\boldsymbol{\theta}$ defines the E-step of the EM algorithm.

The M-step of the EM algorithm corresponds simply to finding $\boldsymbol{\theta}^{(t+1)}$ that maximize the expected complete log-likelihood:

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}; \boldsymbol{\theta}^{(t)}).$$

This estimation step is again easy, as if we had complete (but weighted) data.

More generally, we do not actually need to solve for the parameters $\boldsymbol{\theta}^{(t+1)}$ that exactly maximize the expected complete log-likelihood in the M-step. To guarantee that the log-likelihood of the observed data increases after every M-step, it suffices to find any $\boldsymbol{\theta}^{(t+1)}$ for which

$$Q(\boldsymbol{\theta}^{(t+1)}; \boldsymbol{\theta}^{(t)}) > Q(\boldsymbol{\theta}^{(t)}; \boldsymbol{\theta}^{(t)}).$$

When we settle for a suboptimal $\boldsymbol{\theta}^{(t+1)}$, with the above constraint, the M-step is known as a generalized M-step.

Let us now turn to a slightly more complicated hidden variable density model. We consider the case of a single variable x. Specifically, we define a one dimensional density $p(x; \boldsymbol{\theta})$, $x \in \mathcal{R}$, in terms of two hidden variables

$$p(x; \boldsymbol{\theta}) = \sum_{j=1}^{m} \sum_{i=1}^{l} p_j q_i N(x; \mu_j, \sigma_i^2) \tag{1}$$

where $\sum_{j=1}^{m} p_j = 1$, $\sum_{i=1}^{l} q_i = 1$, and $\boldsymbol{\theta} = \{p_1, \ldots, p_m, \mu_1, \ldots, \mu_m, q_1, \ldots, q_l, \sigma_1^2, \ldots, \sigma_l^2\}$. We could view this as a simple mixture model with ml Gaussian components indexed by (j, i). However, unlike before, the parameters of the ml components cannot be set independently. For example, there are only m possible means, not ml. Alternatively, we could view this as a mixture of m non-Gaussian components, where each component distribution is a scale mixture, $p(x|j; \theta) = \sum_{i=1}^{l} q_i N(x; \mu_j, \sigma_i^2)$, combining Gaussians with different variances (scales). These m components are again not parameterized independently of each other.

We introduce hidden variables $z_1, \ldots, z_n$ and $y_1, \ldots, y_n$ to indicate the selection of the mixture components. The complete log-likelihood of the data is

$$l_c(x_1, \ldots, x_n, z_1, \ldots, z_n, y_1, \ldots, y_n; \boldsymbol{\theta}) = \sum_{k=1}^{n} \log\left(p_{z_k} q_{y_k} N\left(x; \mu_{z_k}, \sigma_{y_k}^2\right)\right).$$

As a hidden variable model, $p(x; \boldsymbol{\theta})$ in Eq.(1) can be estimated from samples $x_1, ..., x_n$ via the EM algorithm. For this purpose, we will have to evaluate the expected complete log-likelihood (E-step) and solve for the parameters in the (generalized) M-step.

Answer the following questions regarding to the model introduced in Eq. (1) with two hidden variables.

1. Provide the expression for the posterior probability $p(z_k, y_k|x_k, \boldsymbol{\theta}^{(t)})$ over the hidden variables given x_k and $\theta^{(t)}$ (current setting of the model parameters).

2. Write down the portion of the expected complete log-likelihood that pertains to the new mean parameters $\mu_1, \ldots, \mu_m$ we intend to find in the M-step. The dependence on these means should be made explicit.

3. Solving the M-step for the maximizing parameters $\theta^{(t+1)}$ would require us to jointly optimize the means $\mu_1, \ldots, \mu_m$ and the variances $\sigma_1^2, \ldots, \sigma_l^2$ in this case. We can, however, easily specify a simpler

Problem 3 continued on next page...

generalized M-step: solve for μ_j's given fixed σ_i^2's, and subsequently solve for σ_i^2's given the new values of μ_j's. Provide an expression for the maximizing μ_j's given fixed $\sigma_1^2, \ldots, \sigma_l^2$.

4. We have provided you with MATLAB code to test the new latent variable model. Load `p2data.mat` into MATLAB to get $n = 300$ samples x drawn from distribution `trueparam` (also provided). Estimate the parameters $\boldsymbol{\theta}$ of the latent variable model via the EM-algorithm, `param = em(x,m,l)`, using $m = 1, \ldots, 4$, and $l = 1, 2, 3$. Plot the samples, estimated density, and the data generating distribution using

   ```
   plotdensity(x,param,'b'); hold on;
   plotdensity(x,trueparam,'r'); hold off;
   ```

   You may have to run the EM-algorithm for a few times to get the best results for each combination (m, l) as the initial parameters are selected at random. Provide a brief explanation why there might be a difference in the approximation quality between setting $l = 1$ (single variance parameter) and $l > 1$, even if m can take different values.

Problem 4

[**20 points**]

HMM has two hidden states $z_t \in \{1, 2\}$ and three possible observation values $x_t \in \{1, 2, 3\}$. The initial state probabilities are $\{\pi_1 = 0.6, \pi_2 = 0.4\}$. The transition probabilities are $\{a_{11} = 0.7, a_{12} = 0.3, a_{21} = 0.4, a_{22} = 0.6\}$. The emission probabilities are $\{b_{11} = 0.1, b_{12} = 0.4, b_{13} = 0.5, b_{21} = 0.6, b_{22} = 0.3, b_{23} = 0.1\}$.

1. Write a program (attach the matlab code) to sample data sequences of length 10 from the HMM described above. Write down one observation sequence $(x_1, \ldots, x_{10})$ generated by your program.

2. Implement the Viterbi algorithm (attach the matlab code) to decode the observation sequence you get in part 1 and write down the most likely sequence of hidden states.