Lecture 12, MIT 6.867 (Machine Learning), Fall 2010

Michael Collins

February 22, 2012

Today's Lecture

- Gaussian mixture models, and the EM algorithm
- The general form of the EM algorithm; convergence properties
- The EM algorithm applied to the naive Bayes model

Gaussian Distributions: A Special Case

- If Σ is the identity matrix, then we have a simple case of the Gaussian distribution, where the only parameter is $\underline{\mu}$:

$$N(\underline{x}; \underline{\mu}) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}||\underline{x} - \underline{\mu}||^2\right)$$

- Given data points $\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n$, the maximum-likelihood estimates for $\underline{\mu}$ maximize

$$L(\underline{\theta}) = \sum_{i=1}^{n} \log N(\underline{x}_i; \underline{\mu})$$

Giving (again):

$$\underline{\hat{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \underline{x}_i$$

Gaussian Mixture Models (GMMs)

- Model form for a GMM with k mixture components:

$$p(\underline{x}; \underline{\theta}) = \sum_{z=1}^{k} q(z) N(\underline{x}; \underline{\mu}_z)$$

- The parameter vector $\underline{\theta}$ contains the following parameters:

1. $q(z)$ for $z = 1 \ldots k$. We have $q(z) \geq 0$ for all z, and

$$\sum_{z=1}^{k} q(z) = 1$$

2. $\underline{\mu}_z$ for $z = 1 \ldots k$

Maximum-Likelihood Estimation for GMMs

- The maximum-likelihood estimates for $q(z)$ and $\underline{\mu}_z$ maximize the following function:

$$L(\underline{\theta}) = \sum_{i=1}^{n} \log p(\underline{x}_i; \underline{\theta})$$
$$= \sum_{i=1}^{n} \log \sum_{z=1}^{k} q(z) N(\underline{x}_i; \underline{\mu}_z)$$

- How do we find the ML estimates in this case?

- For an applet demonstrating ML estimation for GMMs, see `http://www.socr.ucla.edu/Applets.dir/MixtureEM.html`

The EM Algorithm for GMMS

Initialization: Set $q^0(z)$ and $\underline{\mu}_z^0$ to some initial values (e.g., random initial values)

Algorithm: For $t = 1 \ldots T$:

1 For $i = 1 \ldots n$, and $z = 1 \ldots k$, calculate

$$\delta(z|i) = p(z|\underline{x}_i; \underline{\theta}^{t-1}) = \frac{q^{t-1}(z) N(\underline{x}_i; \underline{\mu}_z^{t-1})}{\sum_z q^{t-1}(z) N(\underline{x}_i; \underline{\mu}_z^{t-1})}$$

2 Recalculate the parameters:

$$q^t(z) = \frac{n(z)}{n} \quad \text{and} \quad \underline{\mu}_z^t = \frac{\sum_{i=1}^n \delta(z|i) \underline{x}_i}{n(z)}$$

where $n(z) = \sum_{i=1}^n \delta(z|i)$

Today's Lecture

- Gaussian mixture models, and the EM algorithm
- The general form of the EM algorithm; convergence properties
- The EM algorithm applied to the naive Bayes model

Properties of the EM Algorithm

- The algorithm defines a sequence of parameter values $\underline{\theta}^0, \underline{\theta}^1, \ldots, \underline{\theta}^T$
- We'll show that for all t,
$$L(\underline{\theta}^t) \geq L(\underline{\theta}^{t-1})$$
- The algorithm will (usually*) converge to a local maximum of $L(\underline{\theta})$, but it may get stuck in locally optimal solutions
- "usually*": technically, it may also get stuck in a saddle-point of $L(\underline{\theta})$, i.e., a point where the gradient is zero, but which is not a local maximum

Initialization is Important

- EM can get stuck in local maxima: because of this, the initial parameter values are important
- One approach: choose random initial values for the $\underline{\mu}_z$ parameters
- Another approach: choose $\underline{\mu}_z$ parameters to be randomly selected points from the training set
- Typically run the EM algorithm multiple times, pick the best solution

A General Form of the EM Algorithm

- Goal: maximize

$$L(\underline{\theta}) = \sum_{i=1}^{n} \log p(\underline{x}_i; \underline{\theta}) = \sum_{i=1}^{n} \log \sum_{z=1}^{k} p(\underline{x}_i, z; \underline{\theta})$$

- The algorithm: For $t = 1 \ldots T$

$$\underline{\theta}^t = \arg\max_{\underline{\theta}} Q(\underline{\theta}, \underline{\theta}^{t-1})$$

where

$$Q(\underline{\theta}, \underline{\theta}^{t-1}) = \sum_{i=1}^{n} \sum_{z=1}^{k} p(z|\underline{x}_i; \underline{\theta}^{t-1}) \log p(\underline{x}_i, z; \underline{\theta})$$

The Relationship to Estimation with Fully Observed Data

- Maximum-likelihood estimation with fully observed data: training set is $(\underline{x}_i, z_i)$ for $i = 1 \ldots n$, maximize

$$L(\underline{\theta}) = \sum_{i=1}^{n} \sum_{z=1}^{k} \delta(z|i) \log p(\underline{x}_i, z; \underline{\theta})$$

  where $\delta(z|i) = 1$ if $z = z_i$, and 0 otherwise

- Maximum-likelihood estimation with EM: training set is $\underline{x}_i$ for $i = 1 \ldots n$. At each iteration, choose $\underline{\theta}^t$ to maximize

$$Q(\underline{\theta}, \underline{\theta}^{t-1}) = \sum_{i=1}^{n} \sum_{z=1}^{k} \delta(z|i) \log p(\underline{x}_i, z; \underline{\theta})$$

  where $\delta(z|i) = p(z|\underline{x}_i; \underline{\theta}^{t-1})$

Proof of Convergence

- It can be shown (see next slides) that for any $\underline{\theta}'$, $\underline{\theta}$,

$$L(\underline{\theta}') - L(\underline{\theta}) = Q(\underline{\theta}', \underline{\theta}) - Q(\underline{\theta}, \underline{\theta}) + K(\underline{\theta}', \underline{\theta})$$

  where

$$K(\underline{\theta}', \underline{\theta}) = \sum_{i=1}^{n} \sum_{z=1}^{k} p(z|\underline{x}_i; \underline{\theta}) \log \frac{p(z|\underline{x}_i; \underline{\theta})}{p(z|\underline{x}_i; \underline{\theta}')}$$

- In addition, $K(\underline{\theta}', \underline{\theta}) \geq 0$ for all $\underline{\theta}', \underline{\theta}$, hence

$$L(\underline{\theta}^t) - L(\underline{\theta}^{t-1}) \geq Q(\underline{\theta}^t, \underline{\theta}^{t-1}) - Q(\underline{\theta}^{t-1}, \underline{\theta}^{t-1}) \geq 0$$

  (2nd inequality holds because $\underline{\theta}^t = \arg\max_{\underline{\theta}} Q(\underline{\theta}, \underline{\theta}^{t-1})$)

Proof of Convergence (Continued)

- We have

$$L(\underline{\theta}^t) - L(\underline{\theta}^{t-1}) \geq Q(\underline{\theta}^t, \underline{\theta}^{t-1}) - Q(\underline{\theta}^{t-1}, \underline{\theta}^{t-1}) \geq 0$$

  hence the likelihood is non-decreasing at each iteration of EM

- In addition it can be shown that

$$Q(\underline{\theta}', \underline{\theta}) - Q(\underline{\theta}, \underline{\theta}) = 0 \quad \text{iff} \quad \frac{dL(\underline{\theta})}{d\underline{\theta}} = 0$$

  i.e., we're at a stationary point of L. Hence
  $L(\underline{\theta}^t) - L(\underline{\theta}^{t-1}) > 0$ if $\underline{\theta}^{t-1}$ is not a stationary point of L

Proof that $L(\underline{\theta}') - L(\underline{\theta}) = Q(\underline{\theta}', \underline{\theta}) - Q(\underline{\theta}, \underline{\theta}) + K(\underline{\theta}', \underline{\theta})$

(Follows by some simple algebra...)

$$\sum_{i=1}^{n} \sum_{z=1}^{k} p(z|\underline{x}_i; \underline{\theta}) \log p(z|\underline{x}_i; \underline{\theta}') = \sum_{i=1}^{n} \sum_{z=1}^{k} p(z|\underline{x}_i; \underline{\theta}) \log \frac{p(\underline{x}_i, z; \underline{\theta}')}{\sum_{z=1}^{k} p(\underline{x}_i, z; \underline{\theta}')}$$

$$= \sum_{i=1}^{n} \sum_{z=1}^{k} p(z|\underline{x}_i; \underline{\theta}) \log p(\underline{x}_i, z; \underline{\theta}') - \sum_{i=1}^{n} \sum_{z=1}^{k} p(z|\underline{x}_i; \underline{\theta}) \log \sum_{z=1}^{k} p(\underline{x}_i, z; \underline{\theta}')$$

$$= Q(\underline{\theta}', \underline{\theta}) - \sum_{i=1}^{n} \sum_{z=1}^{k} p(z|\underline{x}_i; \underline{\theta}) \log p(\underline{x}_i; \underline{\theta}')$$

$$= Q(\underline{\theta}', \underline{\theta}) - \sum_{i=1}^{n} \log p(\underline{x}_i; \underline{\theta}')$$

$$= Q(\underline{\theta}', \underline{\theta}) - L(\underline{\theta}')$$

Proof that $L(\underline{\theta}') - L(\underline{\theta}) = Q(\underline{\theta}', \underline{\theta}) - Q(\underline{\theta}, \underline{\theta}) + K(\underline{\theta}', \underline{\theta})$

We've shown that

$$\sum_{i=1}^{n} \sum_{z=1}^{k} p(z|\underline{x}_i; \underline{\theta}) \log p(z|\underline{x}_i; \underline{\theta}') = Q(\underline{\theta}', \underline{\theta}) - L(\underline{\theta}') \qquad (1)$$

It follows also that

$$\sum_{i=1}^{n} \sum_{z=1}^{k} p(z|\underline{x}_i; \underline{\theta}) \log p(z|\underline{x}_i; \underline{\theta}) = Q(\underline{\theta}, \underline{\theta}) - L(\underline{\theta}) \qquad (2)$$

If we take (2) - (1) we get the desired result:

$$\sum_{i=1}^{n} \sum_{z=1}^{k} p(z|\underline{x}_i; \underline{\theta}) \log \frac{p(z|\underline{x}_i; \underline{\theta})}{p(z|\underline{x}_i; \underline{\theta}')} = Q(\underline{\theta}, \underline{\theta}) - L(\underline{\theta}) - Q(\underline{\theta}', \underline{\theta}) + L(\underline{\theta}')$$

Today's Lecture

- Gaussian mixture models, and the EM algorithm
- The general form of the EM algorithm; convergence properties
- The EM algorithm applied to the naive Bayes model

Another Example: EM for Naive Bayes

- Assume each $\underline{x} \in \{0,1\}^d$. The model form:

$$p(\underline{x}; \underline{\theta}) = \sum_{z=1}^{k} q(z) \prod_{j=1}^{d} q_j(x_j|z)$$

- Parameters of the model:
  - $q(z)$ for $z = 1 \dots k$
    (constraints: $q(z) \geq 0$, and $\sum_z q(z) = 1$)
  - $q_j(x|z)$ for $j = 1 \dots d$, $x \in \{0,1\}$, and $z = 1 \dots k$
    (constraints: $q_j(x|z) \geq 0$, and $\sum_x q_j(x|z) = 1$)

Guess the optimal parameters...

- I have 5 training examples:

$$\underline{x}_1 = \underline{x}_2 = (1, 1, 0, 0)$$

$$\underline{x}_3 = \underline{x}_4 = \underline{x}_5 = (0, 0, 1, 1)$$

- I choose $k = 2$. What are the maximum-likelihood parameters in this case?

- (An example of how this kind of data might arise: each vector $\underline{x}$ represents a document. $x_1 = 1$ if the document contains "Obama", 0 otherwise. $x_2 = 1$ iff document contains "McCain". $x_3 = 1$ iff a document contains "Philadelphia". $x_4 = 1$ iff a document contains "Tampa".)

A Warm-up: Maximum-Likelihood Estimates for Fully Observed Data

- Training data $(\underline{x}_i, z_i)$ for $i = 1 \ldots n$
- Maximum-likelihood estimates maximize

$$L(\underline{\theta}) = \sum_{i=1}^{n} \log p(\underline{x}_i, z_i; \underline{\theta}) = \sum_{i=1}^{n} \sum_{z=1}^{k} \delta(z|i) \log p(\underline{x}_i, z; \underline{\theta})$$

  where $\delta(z|i) = 1$ if $z = z_i$, 0 otherwise
- Solution:

$$q(z) = \frac{1}{n} \sum_{i=1}^{n} \delta(z|i) \quad q_j(x|z) = \frac{\sum_{i:x_{i,j}=x} \delta(z|i)}{\sum_{i=1}^{n} \delta(z|i)}$$

The EM Algorithm for Naive Bayes

Initialization: Set $q^0(z)$ and $q_j^0(x|z)$ to some initial values (e.g., random initial values)

Algorithm: For $t = 1 \ldots T$:

1 For $i = 1 \ldots n$, and $z = 1 \ldots k$, calculate

$$\delta(z|i) = p(z|\underline{x}_i; \underline{\theta}^{t-1}) = \frac{q^{t-1}(z) \prod_{j=1}^{d} q_j^{t-1}(x_{i,j}|z)}{\sum_z q^{t-1}(z) \prod_{j=1}^{d} q_j^{t-1}(x_{i,j}|z)}$$

2 Recalculate the parameters:

$$q^t(z) = \frac{1}{n} \sum_{i=1}^{n} \delta(z|i) \quad q_j^t(x|z) = \frac{\sum_{i:x_{i,j}=x} \delta(z|i)}{\sum_{i=1}^{n} \delta(z|i)}$$

Clustering

- We've seen models of the form

$$p(\underline{x}; \underline{\theta}) = \sum_z q(z) N(\underline{x}; \underline{\mu}_z)$$

- After training a model using EM, we can assign each point in $\underline{x}_1, \underline{x}_2, \ldots \underline{x}_n$ to a different *cluster*:

$$\begin{aligned} z_i &= \arg\max_z p(z | \underline{x}_i; \underline{\theta}) \\ &= \arg\max_z \frac{q(z) N(\underline{x}_i; \underline{\mu}_z)}{\sum_z q(z) N(\underline{x}_i; \underline{\mu}_z)} \\ &= \arg\max_z q(z) N(\underline{x}_i; \underline{\mu}_z) \end{aligned}$$

K-Means Clustering

- Goal: for a dataset $\underline{x}_1 \ldots \underline{x}_n$, try to find:

  1. *cluster labels* $z_1 \ldots z_n$, where each $z_i \in \{1, 2, \ldots k\}$

  2. *cluster centers* $\underline{\mu}_1 \ldots \underline{\mu}_k$

- We will always have:

$$z_i = \arg\min_{z} ||\underline{x}_i - \underline{\mu}_z||^2$$

  i.e., each point gets assigned to the cluster with the closest center

- The quality of a clustering is measured as

$$J(z_1, z_2, \ldots, z_n, \underline{\mu}_1 \ldots \underline{\mu}_k) = \sum_{i=1}^{n} ||\underline{x}_i - \underline{\mu}_{z_i}||^2$$

The K-means Clustering Algorithm

Initialization: Set $\underline{\mu}_z^0$ for $z = 1 \ldots k$ to some initial values (e.g., random initial values)

Algorithm: For $t = 1 \ldots T$:

1 For $i = 1 \ldots n$, calculate $z_i^{t-1} = \arg\min_z ||\underline{x}_i - \underline{\mu}_z^{t-1}||^2$

2 Recalculate the cluster centers:

$$\underline{\mu}_z^t = \frac{\sum_{i=1}^n \delta(z|i)\underline{x}_i}{\sum_{i=1}^n \delta(z|i)}$$

where $\delta(z|i) = 1$ if $z^{t-1} = z_i$, 0 otherwise

Output: cluster centers $\underline{\mu}_z^T$ for $z = 1 \ldots k$,
cluster labels $z_i^T = \arg\min_z ||\underline{x}_i - \underline{\mu}_z^T||^2$ for $i = 1 \ldots n$

Convergence Properties of K-means

- Consider again our objective function (which we're aiming to minimize):

$$J(z_1, z_2, \ldots, z_n, \underline{\mu}_1 \ldots \underline{\mu}_k) = \sum_{i=1}^{n} ||\underline{x}_i - \underline{\mu}_{z_i}||^2$$

- Step 1 of k-means: minimizes J with respect to the z_i variables (keeping the $\underline{\mu}_z$ variables fixed)

- Step 2: minimizes J with respect to the $\underline{\mu}_z$ variables (keeping the z_i variables fixed)

- K-means will converge to a local minimum of J