Disclaimer

- The material provided in this document is not my original work and is a summary of some one else's work(s).
- A simple Google search of the title of the document will direct you to the original source of the material.
- I do not guarantee the accuracy, completeness, timeliness, validity, non-omission, merchantability or fitness of the contents of this document for any particular purpose.
- Downloaded from najeebkhan.github.io

THE EM ALGORITHM

Presented by

Najeeb

July 14, 2014

Introduction

- Given the true densities for all classes, one may determine the optimum Bayes classifier

$$\arg\max_k P(w_k \mid x) = \arg\max_k P(x \mid w_k) P(w_k)$$

- Given parameters trained to maximize the likelihood for the data in each class, we can approximate the Bayes classifier
- However, in many problems, there is no analytical solution for the parameters of the statistical classifier given the training data

Introduction

- An example of such a problem is finding the means, variances, and mixture weights for a sum of Gaussian densities, given a sample of data points
- We will postulate a model in which each individual Gaussian generates some of the observed data points
- Given that there is no way to directly compute estimates for these quantities, we need an optimization technique to find the parameters for each Gaussian that will maximize the likelihood of all of the observed data

Introduction

- The dominant approach to such problems is called the Expectation Maximization (EM) algorithm
- The parameter estimation problem is structured to incorporate variables representing information that is not directly observed, but that is assumed to be part of the model that generated the data
- Such a variable is often called hidden or missing
- In the Gaussian mixture case, a hidden variable could be the index of the Gaussian that generated a data point

Introduction

- The key idea of EM is to
  - Estimate the densities by taking an expectation of the logarithm of the joint density between the known and unknown components
  - Maximize this function by updating the parameters that are used in the probability estimation

Expectation

- The log likelihood can be expressed in terms of hidden variables as

$$\ell(X \mid \Theta) = \log P(X \mid \Theta)$$

$$P(k, X \mid \Theta) = \frac{P(k, X, \Theta)}{P(\Theta)} = \frac{P(k \mid X, \Theta) P(X, \Theta)}{P(\Theta)}$$

$$P(k, X \mid \Theta) = P(k \mid X, \Theta) P(X \mid \Theta)$$

$$P(X \mid \Theta) = \frac{P(k, X \mid \Theta)}{P(k \mid X, \Theta)}$$

$$\log P(X \mid \Theta) = \log P(k, X \mid \Theta) - \log P(k \mid X, \Theta)$$

Expectation

$$\log P(X \mid \Theta) = \log P(k, X \mid \Theta) - \log P(k \mid X, \Theta)$$

$$\sum_{k=1}^{K} P(k \mid X, \Theta^i) \log P(X \mid \Theta) =$$

$$\sum_{k=1}^{K} P(k \mid X, \Theta^i) \log P(k, X \mid \Theta) - P(k \mid X, \Theta^i) \log P(k \mid X, \Theta)$$

$$\log P(X \mid \Theta) \sum_{k=1}^{K} P(k \mid X, \Theta^i) = Q(\Theta, \Theta^i) - H(\Theta, \Theta^i)$$

$$Q(\Theta, \Theta^i) = \sum_{k=1}^{K} P(k \mid X, \Theta^i) \log P(k, X \mid \Theta)$$

$$H(\Theta, \Theta^i) = \sum_{k=1}^{K} P(k \mid X, \Theta^i) \log P(k \mid X, \Theta)$$

Q function

- If the Q function is maximized the likelihood is also maximized

$$
\begin{aligned}
Q(\Theta,\Theta^i) &= \sum_{k=1}^{K} P(k \mid X,\Theta^i)\log P(k, X \mid \Theta) \\
&= \sum_{k=1}^{K} P(k \mid X,\Theta^i)\log\left[P(k \mid X,\Theta)P(X \mid \Theta)\right] \\
&= \sum_{k=1}^{K} P(k \mid X,\Theta^i)\log P(k \mid X,\Theta) + \log P(X \mid \Theta)\sum_{k=1}^{K} P(k \mid X,\Theta^i) \\
&= \sum_{k=1}^{K} P(k \mid X,\Theta^i)\log P(k \mid X,\Theta) + \log P(X \mid \Theta)
\end{aligned}
$$

Q function

$$Q(\Theta,\Theta^i)=\sum_{k=1}^{K} P(k \mid X,\Theta^i)\log P(k \mid X,\Theta)+\log P(X \mid \Theta) \qquad (1)$$

$$Q(\Theta^i,\Theta^i)=\sum_{k=1}^{K} P(k \mid X,\Theta^i)\log P(k \mid X,\Theta^i)+\log P(X \mid \Theta^i) \qquad (2)$$

$$\log P(X \mid \Theta)-\log P(X \mid \Theta^i)=Q(\Theta,\Theta^i)-Q(\Theta^i,\Theta^i)+NNC$$

Q function

$$Q(\Theta,\Theta^i)=\sum_{k=1}^{K} P(k \mid X,\Theta^i)\log P(k \mid X,\Theta)+\log P(X \mid \Theta) \quad (1)$$

$$Q(\Theta^i,\Theta^i)=\sum_{k=1}^{K} P(k \mid X,\Theta^i)\log P(k \mid X,\Theta^i)+\log P(X \mid \Theta^i) \quad (2)$$

$$\log P(X \mid \Theta)-\log P(X \mid \Theta^i)=Q(\Theta,\Theta^i)-Q(\Theta^i,\Theta^i)+NNC$$

- If a change to Θ increases Q, log P(x | Θ) increases

Q function

- In principle, we could maximize the expectation for each value of Θ and then re-estimate Θ
- In many cases, it is possible to structure the problem so that we can analytically determine the choice for the parameters that will maximize the expectation in each iteration
- Then a new expression for the expectation can be determined, followed by a new parameter estimation, and so on

EM For Mixture Gaussian Density Estimation

- We can always decompose an unknown probability density P(X | Θ) as

$$P(X \mid \Theta) = \sum_{k=1}^{K} P(X, k \mid \Theta) = \sum_{k=1}^{K} P(k \mid \Theta) P(X \mid k, \Theta)$$

- We can write an expression for the log joint density for observed and hidden variables

$$Z = \log P(x, k \mid \Theta) = \log[P(k \mid \Theta) P(x \mid k, \Theta)]$$

- With an expected value over N samples and K mixture components of

$$Q = \sum_{k=1}^{K} \sum_{n=1}^{N} P(k \mid x_n, \Theta^i) \log[P(k \mid \Theta) P(x_n \mid k, \Theta)]$$

- Θ^i is the old parameter

EM For Mixture Gaussian Density Estimation

- Using properties of the log

$$Q = \sum_{k=1}^{K}\sum_{n=1}^{N} P(k \mid x_n, \Theta^i) \log P(k \mid \Theta) + \sum_{k=1}^{K}\sum_{n=1}^{N} P(k \mid x_n, \Theta^i) \log P(x_n \mid k, \Theta)$$

- We will often assume models for which the Θ parameters of the two terms are disjointed and can be optimized separately

EM For Mixture Gaussian Density Estimation

- Assuming that each component P(x|k) is Gaussian

$$Q = \sum_{k=1}^{K}\sum_{n=1}^{N} P(k \mid x_n, \Theta^i) \log P(k \mid \Theta) +$$

$$\sum_{k=1}^{K}\sum_{n=1}^{N} P(k \mid x_n, \Theta^i)[-\log \sigma_k - \frac{(x_n - \mu_k)^2}{2\sigma_k^2} + C]$$

- Given this choice of parametric form for the mixture density, we can use standard optimization methods to find the best value for the parameters

EM For Mixture Gaussian Density Estimation

- Solving for the means

$$\frac{\partial Q}{\partial \mu_j} = 0$$

$$\sum_{n=1}^{N} P(j \mid x_n, \Theta^i)\left(\frac{x_n}{\sigma_j^2} - \frac{\mu_j}{\sigma_j^2}\right) = 0$$

$$\sum_{n=1}^{N} P(j \mid x_n, \Theta^i) x_n = \sum_{n=1}^{N} P(j \mid x_n, \Theta^i)\mu_j$$

$$\mu_j = \frac{\sum_{n=1}^{N} P(j \mid x_n, \Theta^i) x_n}{\sum_{n=1}^{N} P(j \mid x_n, \Theta^i)}$$

EM For Mixture Gaussian Density Estimation

- Solving for the means

$$\frac{\partial Q}{\partial \mu_j} = 0$$

$$\sum_{n=1}^{N} P(j \mid x_n, \Theta^i)\left(\frac{x_n}{\sigma_j^2} - \frac{\mu_j}{\sigma_j^2}\right) = 0$$

$$\sum_{n=1}^{N} P(j \mid x_n, \Theta^i) x_n = \sum_{n=1}^{N} P(j \mid x_n, \Theta^i)\mu_j$$

$$\mu_j = \frac{\sum_{n=1}^{N} P(j \mid x_n, \Theta^i) x_n}{\sum_{n=1}^{N} P(j \mid x_n, \Theta^i)}$$

EM For Mixture Gaussian Density Estimation

- Solving for the variances

$$\frac{\partial Q}{\partial \sigma_j} = 0$$

$$\sum_{n=1}^{N} P(j \mid x_n, \Theta^i)\left(-\frac{1}{\sigma_j} + \frac{(x_n - \mu_j)^2}{\sigma_j^3} \right) = 0$$

$$\frac{1}{\sigma_j}\sum_{n=1}^{N} P(j \mid x_n, \Theta^i) = \frac{1}{\sigma_j^3}\sum_{n=1}^{N} P(j \mid x_n, \Theta^i)(x_n - \mu_j)^2$$

$$\sigma_j^2 = \frac{\sum_{n=1}^{N} P(j \mid x_n, \Theta^i)(x_n - \mu_j)^2}{\sum_{n=1}^{N} P(j \mid x_n, \Theta^i)}$$

EM For Mixture Gaussian Density Estimation

- Solving for the Mixture Weights

$$Q^* = \sum_{k=1}^{K}\sum_{n=1}^{N} P(k \mid x_n, \Theta^i) \log P(k \mid \Theta) +$$

$$\sum_{k=1}^{K}\sum_{n=1}^{N} P(k \mid x_n, \Theta^i)[-\log \sigma_k - \frac{(x_n - \mu_k)^2}{2\sigma_k^2} + C] +$$

$$\lambda\left[\sum_{k=1}^{K} P(k \mid \Theta) - 1\right]$$

EM For Mixture Gaussian Density Estimation

- Solving for the Mixture Weights

$$\frac{\partial Q^*}{\partial P(j \mid \Theta)} = 0$$

$$\frac{1}{P(j \mid \Theta)} \sum_{n=1}^{N} P(j \mid x_n, \Theta^i) + \lambda = 0$$

$$P(j \mid \Theta) = \frac{1}{\lambda} \sum_{n=1}^{N} P(j \mid x_n, \Theta^i)$$

$$P(j \mid \Theta) = \frac{1}{N} \sum_{n=1}^{N} P(j \mid x_n, \Theta^i)$$

EM For Mixture Gaussian Density Estimation

- Solving for the Mixture Weights

$$P(j \mid x_n, \Theta^i) = \frac{P(x_n \mid j, \Theta^i) P(j \mid \Theta^i)}{P(x_n \mid \Theta^i)}$$

$$P(j \mid x_n, \Theta^i) = \frac{P(x_n \mid j, \Theta^i) P(j \mid \Theta^i)}{\sum_{k=1}^{K} P(x_n \mid k, \Theta^i) P(k \mid \Theta^i)}$$

HMM Training

$$Q = \sum_{k=1}^{K} \sum_{n=1}^{N} P(k \mid x_n, \Theta^i) \log[P(k \mid \Theta) P(x_n \mid k, \Theta)]$$

- For the case of HMM we use the state sequence as the hidden variable

$$Q = \sum_{Q} P(Q \mid X_1^N, \Theta_{old}, M) \log\Big[P(X_1^N \mid Q, \Theta, M) P(Q \mid \Theta, M) \Big]$$

$$\begin{aligned} Q = & \sum_{n=1}^{N} \sum_{k=1}^{L(M)} P(q_k^n \mid X_1^N, \Theta_{old}, M) \log P(x_n \mid q_k^n, \Theta, M) \\ & + \sum_{k=1}^{L(M)} P(q_k^1 \mid X_1^N, \Theta_{old}, M) \log P(x_n \mid q_k^1, \Theta, M) \\ & + \sum_{n=2}^{N} \sum_{k=1}^{L(M)} \sum_{l=1}^{L(M)} P(q_l^n, q_k^{n-1} \mid X_1^N, \Theta_{old}, M) \log P(q_l^n \mid q_k^{n-1}, \Theta, M) \end{aligned}$$

HMM Training

- Assuming single univariate Gaussian is associated with each hmm state
-

$$\mu_j = \frac{\sum_{n=1}^{N} P(q_j^n \mid X_1^N, \Theta_{old}, M) x_n}{\sum_{n=1}^{N} P(q_j^n \mid X_1^N, \Theta^i)}$$

$$\sigma_j^2 = \frac{\sum_{n=1}^{N} P(q_j^n \mid X_1^N, \Theta_{old}, M)(x_n - \mu_j)^2}{\sum_{n=1}^{N} P(q_j^n \mid X_1^N, \Theta^i)}$$