EM for Naive Bayes and Gaussian Mixture Models, k-Means Clustering

Karl Stratos

June 27, 2018

EM Template

Input: model $P_\Phi(\boldsymbol{x}, z)$, unlabeled data $U = \{\boldsymbol{x}^{(i)}\}_{i=1}^n$, T

Output: local maximizer of $L_U(\Phi) := \sum_{i=1}^n \log P_\Phi(\boldsymbol{x}^{(i)})$

1. Initialize parameters $\Phi^{(0)}$.
2. For $t = 0 \dots T-1$,

$$\Phi^{(t+1)} \leftarrow \arg\max_\Phi \sum_{i=1}^n \sum_{z=1}^m P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) \times \log P_\Phi(\boldsymbol{x}^{(i)}, z)$$

3. Return $\Phi^{(T)}$.

EM Template

Input: model $P_\Phi(\boldsymbol{x}, z)$, unlabeled data $U = \{\boldsymbol{x}^{(i)}\}_{i=1}^n$, T

Output: local maximizer of $L_U(\Phi) := \sum_{i=1}^n \log P_\Phi(\boldsymbol{x}^{(i)})$

1. Initialize parameters $\Phi^{(0)}$.
2. For $t = 0 \dots T-1$,

$$\Phi^{(t+1)} \leftarrow \arg\max_\Phi \sum_{i=1}^n \sum_{z=1}^m P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) \times \log P_\Phi(\boldsymbol{x}^{(i)}, z)$$

3. Return $\Phi^{(T)}$.

See yesterday's lecture for how this is derived by alternating maximization of the ELBO$(\Phi, \Psi) \leq L_U(\Phi)$ where Ψ defines an auxiliary posterior $P_\Psi(y|\boldsymbol{x})$.

Overview

- EM for Naive Bayes
  - Maximum Likelihood Estimation with Labeled Data
  - Maximum Likelihood Estimation with Unlabeled Data
- EM for Gaussian Mixture Models
  - Maximum Likelihood Estimation with Labeled Data
  - Maximum Likelihood Estimation with Unlabeled Data
- k-Means Clustering

Naive Bayes: Definition

A **naive Bayes (NB)** model with m labels and d binary-valued feature types has $m + 2dm$ parameters, denoted by Φ:

- $q(z) \geq 0$ for each $z \in \{1 \dots m\}$ such that

$$\sum_z q(z) = 1$$

- $q(0|z, j) \geq 0$ and $q(1|z, j) \geq 0$ such that

$$q(0|z, j) + q(1|z, j) = 1$$

for each $j \in \{1 \dots d\}$ and $z \in \{1 \dots m\}$

Naive Bayes: Definition

A **naive Bayes (NB)** model with m labels and d binary-valued feature types has $m + 2dm$ parameters, denoted by Φ:

- $q(z) \geq 0$ for each $z \in \{1 \dots m\}$ such that

$$\sum_z q(z) = 1$$

- $q(0|z, j) \geq 0$ and $q(1|z, j) \geq 0$ such that

$$q(0|z, j) + q(1|z, j) = 1$$

  for each $j \in \{1 \dots d\}$ and $z \in \{1 \dots m\}$

Φ defines a joint distribution over $\boldsymbol{x} = (x_1 \dots x_d) \in \{0, 1\}^d$ and $z \in \{1 \dots m\}$ by

$$P_\Phi(\boldsymbol{x}, z) := q(z) \prod_{j=1}^{d} q(x_j|z, j)$$

Log Likelihood of Labeled Data

If $S = \left\{(\boldsymbol{x}^{(i)}, z^{(i)})\right\}_{i=1}^{n}$ is a set of n iid labeled samples, the log likelihood of S under Φ is

$$L_S(\Phi) = \sum_{i=1}^{n} \log P_\Phi(\boldsymbol{x}^{(i)}, z^{(i)})$$

Log Likelihood of Labeled Data

If $S = \left\{(\boldsymbol{x}^{(i)}, z^{(i)})\right\}_{i=1}^{n}$ is a set of n iid labeled samples, the log likelihood of S under Φ is

$$L_S(\Phi) = \sum_{i=1}^{n} \log P_\Phi(\boldsymbol{x}^{(i)}, z^{(i)})$$

$$= \sum_{i=1}^{n} \log q(z^{(i)}) + \sum_{j=1}^{m} \log q(x_j^{(i)} | z, j)$$

Log Likelihood of Labeled Data

If $S = \left\{(\boldsymbol{x}^{(i)}, z^{(i)})\right\}_{i=1}^{n}$ is a set of n iid labeled samples, the log likelihood of S under Φ is

$$
\begin{aligned}
L_S(\Phi) &= \sum_{i=1}^{n} \log P_\Phi(\boldsymbol{x}^{(i)}, z^{(i)}) \\
&= \sum_{i=1}^{n} \log q(z^{(i)}) + \sum_{j=1}^{m} \log q(x_j^{(i)} | z, j) \\
&= \sum_{z=1}^{m} \textbf{count}(z) \log q(z) + \sum_{z=1}^{m} \sum_{j=1}^{m} \sum_{x \in \{0,1\}} \textbf{count}(z, j, x) \log q(x | z, j)
\end{aligned}
$$

where

$$
\textbf{count}(z) := \sum_{\substack{i=1: \\ z^{(i)}=z}}^{n} 1 \qquad \textbf{count}(z, j, x) := \sum_{\substack{i=1: \\ z^{(i)}=z \\ x_j^{(i)}=x}}^{n} 1
$$

MLE with Labeled Data

What are the parameter values $q(z)$ and $q(x|z,j)$ that maximize

$$\sum_{z=1}^{m} \textbf{count}(z) \log q(z) + \sum_{z=1}^{m} \sum_{j=1}^{m} \sum_{x \in \{0,1\}} \textbf{count}(z,j,x) \log q(x|z,j)$$

under the constraints that they are nonnegative, $\sum_z q(z) = 1$, and $\sum_x q(x|z,j) = 1$?

MLE with Labeled Data

What are the parameter values $q(z)$ and $q(x|z,j)$ that maximize

$$\sum_{z=1}^{m} \textbf{count}(z) \log q(z) + \sum_{z=1}^{m} \sum_{j=1}^{m} \sum_{x \in \{0,1\}} \textbf{count}(z,j,x) \log q(x|z,j)$$

under the constraints that they are nonnegative, $\sum_z q(z) = 1$, and $\sum_x q(x|z,j) = 1$?

Answer: See the lemma in yesterday's lecture for why.

$$q(z) = \frac{\textbf{count}(z)}{n}$$

$$q(x|z,j) = \frac{\textbf{count}(z,j,x)}{\textbf{count}(z,j,0) + \textbf{count}(z,j,1)}$$

Overview

- EM for Naive Bayes
  - Maximum Likelihood Estimation with Labeled Data
  - Maximum Likelihood Estimation with Unlabeled Data
- EM for Gaussian Mixture Models
  - Maximum Likelihood Estimation with Labeled Data
  - Maximum Likelihood Estimation with Unlabeled Data
- k-Means Clustering

Log Likelihood of Unlabeled Data

If $U = \left\{\boldsymbol{x}^{(i)}\right\}_{i=1}^{n}$ is a set of n iid unlabeled samples, the log likelihood of U under Φ is

$$L_U(\Phi) = \sum_{i=1}^{n} \log P_\Phi(\boldsymbol{x}^{(i)})$$

Log Likelihood of Unlabeled Data

If $U = \left\{\boldsymbol{x}^{(i)}\right\}_{i=1}^{n}$ is a set of n iid unlabeled samples, the log likelihood of U under Φ is

$$\begin{aligned} L_U(\Phi) &= \sum_{i=1}^{n} \log P_\Phi(\boldsymbol{x}^{(i)}) \\ &= \sum_{i=1}^{n} \log \left(\sum_{z=1}^{m} P_\Phi(\boldsymbol{x}^{(i)}, z) \right) \end{aligned}$$

Log Likelihood of Unlabeled Data

If $U = \{\boldsymbol{x}^{(i)}\}_{i=1}^{n}$ is a set of n iid unlabeled samples, the log likelihood of U under Φ is

$$
\begin{aligned}
L_U(\Phi) &= \sum_{i=1}^{n} \log P_\Phi(\boldsymbol{x}^{(i)}) \\
&= \sum_{i=1}^{n} \log \left(\sum_{z=1}^{m} P_\Phi(\boldsymbol{x}^{(i)}, z) \right) \\
&= \sum_{i=1}^{n} \log \left(\sum_{z=1}^{m} \log q(z) + \sum_{j=1}^{m} \log q(x_j^{(i)} | z, j) \right)
\end{aligned}
$$

Unfortunately, finding valid parameter values $q(z)$ and $q(x|z,j)$ that maximize this **marginalized** log likelihood is not as trivial (e.g., there is no closed-form solution).

Explanation of EM for This Problem

- EM is a **local search** algorithm to iteratively optimize

$$L_U(\Phi) = \sum_{i=1}^{n} \log \left(\sum_{z=1}^{m} P_\Phi(\boldsymbol{x}^{(i)}, z) \right)$$

That is, it calculates $\Phi^{(1)} \dots \Phi^{(T)}$ such that

$$L_U(\Phi^{(1)}) \leq L_U(\Phi^{(2)}) \leq \cdots \leq L_U(\Phi^{(T-1)}) \leq L_U(\Phi^{(T)})$$

Explanation of EM for This Problem

- EM is a **local search** algorithm to iteratively optimize

$$L_U(\Phi) = \sum_{i=1}^{n} \log \left(\sum_{z=1}^{m} P_\Phi(\boldsymbol{x}^{(i)}, z) \right)$$

  That is, it calculates $\Phi^{(1)} \dots \Phi^{(T)}$ such that

$$L_U(\Phi^{(1)}) \leq L_U(\Phi^{(2)}) \leq \cdots \leq L_U(\Phi^{(T-1)}) \leq L_U(\Phi^{(T)})$$

- Importantly, each EM update *is trivial*: it has a closed-form solution.

Explanation of EM for This Problem

- EM is a **local search** algorithm to iteratively optimize

$$L_U(\Phi) = \sum_{i=1}^{n} \log \left(\sum_{z=1}^{m} P_\Phi(\boldsymbol{x}^{(i)}, z) \right)$$

  That is, it calculates $\Phi^{(1)} \dots \Phi^{(T)}$ such that

$$L_U(\Phi^{(1)}) \leq L_U(\Phi^{(2)}) \leq \cdots \leq L_U(\Phi^{(T-1)}) \leq L_U(\Phi^{(T)})$$

- Importantly, each EM update *is trivial*: it has a closed-form solution.

- As usual with local search algorithms, it only finds a local optimum and is not guaranteed to find a global optimum.

Posterior Probabilities

At each iteration t, we use the current parameter estimates

$$\Phi^{(t)} = \left\{ q^{(t)}(z),\ q^{(t)}(x|z,j) \right\}$$

to calculate the **posterior probabilities** on *individual* samples $\boldsymbol{x}^{(i)}$. This can be easily precomputed by Bayes rule: for every $i \in \{1 \dots n\}$ and $z \in \{1 \dots m\}$, calculate

$$P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) = \frac{P_{\Phi^{(t)}}(\boldsymbol{x}^{(i)}, z)}{P_{\Phi^{(t)}}(\boldsymbol{x}^{(i)})} = \frac{q^{(t)}(z) \prod_{j=1}^{d} q^{(t)}(x_j^{(i)}|z,j)}{\sum_{z=1}^{m} q^{(t)}(z) \prod_{j=1}^{d} q^{(t)}(x_j^{(i)}|z,j)}$$

Expected Log Likelihood of Labeled Data Under $\Phi^{(t)}$

$$\sum_{i=1}^{n}\sum_{z=1}^{m} P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) \times \log P_{\Phi}(\boldsymbol{x}^{(i)}, z)$$

Expected Log Likelihood of Labeled Data Under $\Phi^{(t)}$

$$\sum_{i=1}^{n}\sum_{z=1}^{m} P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) \times \log P_{\Phi}(\boldsymbol{x}^{(i)}, z)$$

$$= \sum_{i=1}^{n}\sum_{z=1}^{m} P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) \times \left(\log q(z) + \sum_{j=1}^{m} \log q(x_j^{(i)}|z, j) \right)$$

Expected Log Likelihood of Labeled Data Under $\Phi^{(t)}$

$$\sum_{i=1}^{n}\sum_{z=1}^{m} P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) \times \log P_{\Phi}(\boldsymbol{x}^{(i)}, z)$$

$$= \sum_{i=1}^{n}\sum_{z=1}^{m} P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) \times \left(\log q(z) + \sum_{j=1}^{m} \log q(x_j^{(i)}|z, j) \right)$$

$$= \sum_{z=1}^{m}\sum_{i=1}^{n} P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) \log q(z) + \sum_{z=1}^{m}\sum_{j=1}^{m}\sum_{i=1}^{n} P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) \log q(x_j^{(i)}|z, j)$$

Expected Log Likelihood of Labeled Data Under $\Phi^{(t)}$

$$\sum_{i=1}^{n}\sum_{z=1}^{m} P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) \times \log P_{\Phi}(\boldsymbol{x}^{(i)}, z)$$

$$= \sum_{i=1}^{n}\sum_{z=1}^{m} P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) \times \left(\log q(z) + \sum_{j=1}^{m} \log q(x_j^{(i)}|z,j) \right)$$

$$= \sum_{z=1}^{m}\sum_{i=1}^{n} P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) \log q(z) + \sum_{z=1}^{m}\sum_{j=1}^{m}\sum_{i=1}^{n} P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) \log q(x_j^{(i)}|z,j)$$

$$= \sum_{z=1}^{m} \widehat{\textbf{count}}_t(z) \log q(z) + \sum_{z=1}^{m}\sum_{j=1}^{m}\sum_{x\in\{0,1\}} \widehat{\textbf{count}}_t(z,j,x) \log q(x|z,j)$$

where

$$\widehat{\textbf{count}}_t(z) := \sum_{i=1}^{n} P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) \qquad \widehat{\textbf{count}}_t(z,j,x) := \sum_{i=1:\, x_j^{(i)}=x}^{n} P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)})$$

MLE in the t-th Iteration of EM

What are the parameter values $q(z)$ and $q(x|z,j)$ that maximize

$$\sum_{z=1}^{m} \widehat{\textbf{count}}_t(z) \log q(z) + \sum_{z=1}^{m}\sum_{j=1}^{m}\sum_{x\in\{0,1\}} \widehat{\textbf{count}}_t(z,j,x) \log q(x|z,j)$$

under the constraints that they are nonnegative, $\sum_z q(z) = 1$, and $\sum_x q(x|z,j) = 1$?

MLE in the t-th Iteration of EM

What are the parameter values $q(z)$ and $q(x|z,j)$ that maximize

$$\sum_{z=1}^{m} \widehat{\textbf{count}}_t(z) \log q(z) + \sum_{z=1}^{m}\sum_{j=1}^{m}\sum_{x\in\{0,1\}} \widehat{\textbf{count}}_t(z,j,x) \log q(x|z,j)$$

under the constraints that they are nonnegative, $\sum_z q(z) = 1$, and $\sum_x q(x|z,j) = 1$?

Answer:

$$q(z) = \frac{\widehat{\textbf{count}}_t(z)}{n} \qquad q(x|z,j) = \frac{\widehat{\textbf{count}}_t(z,j,x)}{\sum_{x\in\{0,1\}} \widehat{\textbf{count}}_t(z,j,x)}$$

EM for NB

1. Initialize NB parameters $\Phi^{(0)}$.
2. For $t = 0 \dots T-1$,
   2.1 For $i = 1 \dots n$ and $z = 1 \dots m$, calculate current posterior

$$P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) \leftarrow \frac{q^{(t)}(z) \prod_{j=1}^{d} q^{(t)}(x_j^{(i)}|z,j)}{\sum_{z=1}^{m} q^{(t)}(z) \prod_{j=1}^{d} q^{(t)}(x_j^{(i)}|z,j)}$$

   2.2 "Count" $\widehat{\textbf{count}}_t(z) \leftarrow \sum_{i=1}^{n} P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)})$ and $\widehat{\textbf{count}}_t(z,j,x) \leftarrow \sum_{i=1:\, x_j^{(i)}=x}^{n} P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)})$ and set $\Phi^{(t+1)} = \left\{q^{(t+1)}(z),\ q^{(t+1)}(x|z,j)\right\}$ by

$$q^{(t+1)}(z) \leftarrow \frac{\widehat{\textbf{count}}_t(z)}{n} \qquad q^{(t+1)}(x|z,j) \leftarrow \frac{\widehat{\textbf{count}}_t(z,j,x)}{\sum_{x \in 0,1} \widehat{\textbf{count}}_t(z,j,x)}$$

3. Return $\Phi^{(T)}$.

Overview

- EM for Naive Bayes
  - Maximum Likelihood Estimation with Labeled Data
  - Maximum Likelihood Estimation with Unlabeled Data
- EM for Gaussian Mixture Models
  - Maximum Likelihood Estimation with Labeled Data
  - Maximum Likelihood Estimation with Unlabeled Data
- k-Means Clustering

Gaussian Mixture Model: Definition

A **Gaussian mixture model (GMM)** with m clusters with identity covariance matrix in $\mathbb{R}^d$ has $m + dm$ parameters, denoted by Φ:

- $\pi(z) \geq 0$ for each $z \in \{1 \dots m\}$ such that

$$\sum_z \pi(z) = 1$$

- $\mu_z \in \mathbb{R}^d$ for each $z \in \{1 \dots m\}$.

Gaussian Mixture Model: Definition

A **Gaussian mixture model (GMM)** with m clusters with identity covariance matrix in $\mathbb{R}^d$ has $m + dm$ parameters, denoted by Φ:

- $\pi(z) \geq 0$ for each $z \in \{1 \dots m\}$ such that

$$\sum_z \pi(z) = 1$$

- $\mu_z \in \mathbb{R}^d$ for each $z \in \{1 \dots m\}$.

Φ defines a joint distribution over $\boldsymbol{x} \in \mathbb{R}^d$ and $z \in \{1 \dots m\}$ by

$$P_\Phi(\boldsymbol{x}, z) := \pi(z) \times \underbrace{\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} ||\boldsymbol{x} - \mu_z||_2^2\right)}_{\mathcal{N}(\boldsymbol{x}|\mu_z, I_d)}$$

Log Likelihood of Labeled Data and MLE

If $S = \left\{(\boldsymbol{x}^{(i)}, z^{(i)})\right\}_{i=1}^{n}$ is a set of n iid labeled samples, the log likelihood of S under Φ is

$$L_S(\Phi) = \sum_{i=1}^{n} \log P_\Phi(\boldsymbol{x}^{(i)}, y^{(i)})$$

Log Likelihood of Labeled Data and MLE

If $S = \left\{(\boldsymbol{x}^{(i)}, z^{(i)})\right\}_{i=1}^{n}$ is a set of n iid labeled samples, the log likelihood of S under Φ is

$$\begin{aligned} L_S(\Phi) &= \sum_{i=1}^{n} \log P_\Phi(\boldsymbol{x}^{(i)}, y^{(i)}) \\ &= \sum_{i=1}^{n} \log \pi(z^{(i)}) - \frac{1}{2} \left|\left|\boldsymbol{x}^{(i)} - \mu_{z^{(i)}}\right|\right|_2^2 - \log \sqrt{2\pi} \end{aligned}$$

Log Likelihood of Labeled Data and MLE

If $S = \left\{(\boldsymbol{x}^{(i)}, z^{(i)})\right\}_{i=1}^{n}$ is a set of n iid labeled samples, the log likelihood of S under Φ is

$$
\begin{aligned}
L_S(\Phi) &= \sum_{i=1}^{n} \log P_\Phi(\boldsymbol{x}^{(i)}, y^{(i)}) \\
&= \sum_{i=1}^{n} \log \pi(z^{(i)}) - \frac{1}{2} \left|\left| \boldsymbol{x}^{(i)} - \mu_{z^{(i)}} \right|\right|_2^2 - \log \sqrt{2\pi} \\
&= \left(\sum_{z=1}^{m} \textbf{count}(z) \log \pi(z) \right) + \left(-\frac{1}{2} \sum_{i=1}^{n} \left|\left| \boldsymbol{x}^{(i)} - \mu_{z^{(i)}} \right|\right|_2^2 \right) + C
\end{aligned}
$$

Log Likelihood of Labeled Data and MLE

If $S = \left\{(\boldsymbol{x}^{(i)}, z^{(i)})\right\}_{i=1}^{n}$ is a set of n iid labeled samples, the log likelihood of S under Φ is

$$
\begin{aligned}
L_S(\Phi) &= \sum_{i=1}^{n} \log P_\Phi(\boldsymbol{x}^{(i)}, y^{(i)}) \\
&= \sum_{i=1}^{n} \log \pi(z^{(i)}) - \frac{1}{2} \left|\left| \boldsymbol{x}^{(i)} - \mu_{z^{(i)}} \right|\right|_2^2 - \log \sqrt{2\pi} \\
&= \left(\sum_{z=1}^{m} \textbf{count}(z) \log \pi(z) \right) + \left(-\frac{1}{2} \sum_{i=1}^{n} \left|\left| \boldsymbol{x}^{(i)} - \mu_{z^{(i)}} \right|\right|_2^2 \right) + C
\end{aligned}
$$

Parameter values $pi(z)$ (with probability constraints) and μ_z (with no constrainst) that maximize $L_S(\Phi)$ are thus

$$
\pi(z) = \frac{\textbf{count}(z)}{n} \qquad \mu_z = \frac{1}{\textbf{count}(z)} \sum_{i=1:\ z^{(i)}=z}^{n} \boldsymbol{x}^{(i)}
$$

Overview

- EM for Naive Bayes
  - Maximum Likelihood Estimation with Labeled Data
  - Maximum Likelihood Estimation with Unlabeled Data
- EM for Gaussian Mixture Models
  - Maximum Likelihood Estimation with Labeled Data
  - Maximum Likelihood Estimation with Unlabeled Data
- k-Means Clustering

Log Likelihood of Unlabeled Data

If $U = \left\{\boldsymbol{x}^{(i)}\right\}_{i=1}^{n}$ is a set of n iid unlabeled samples, the log likelihood of U under Φ is

$$L_U(\Phi) = \sum_{i=1}^{n} \log P_\Phi(\boldsymbol{x}^{(i)})$$

Log Likelihood of Unlabeled Data

If $U = \left\{\boldsymbol{x}^{(i)}\right\}_{i=1}^{n}$ is a set of n iid unlabeled samples, the log likelihood of U under Φ is

$$\begin{aligned} L_U(\Phi) &= \sum_{i=1}^{n} \log P_\Phi(\boldsymbol{x}^{(i)}) \\ &= \sum_{i=1}^{n} \log \left(\sum_{z=1}^{m} P_\Phi(\boldsymbol{x}^{(i)}, z) \right) \end{aligned}$$

Log Likelihood of Unlabeled Data

If $U = \left\{\boldsymbol{x}^{(i)}\right\}_{i=1}^{n}$ is a set of n iid unlabeled samples, the log likelihood of U under Φ is

$$
\begin{aligned}
L_U(\Phi) &= \sum_{i=1}^{n} \log P_\Phi(\boldsymbol{x}^{(i)}) \\
&= \sum_{i=1}^{n} \log \left(\sum_{z=1}^{m} P_\Phi(\boldsymbol{x}^{(i)}, z) \right) \\
&= \sum_{i=1}^{n} \log \left(\sum_{z=1}^{m} \pi(z) \times \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left|\left| \boldsymbol{x}^{(i)} - \mu_z \right|\right|_2^2 \right) \right)
\end{aligned}
$$

Log Likelihood of Unlabeled Data

If $U = \left\{\boldsymbol{x}^{(i)}\right\}_{i=1}^{n}$ is a set of n iid unlabeled samples, the log likelihood of U under Φ is

$$
\begin{aligned}
L_U(\Phi) &= \sum_{i=1}^{n} \log P_\Phi(\boldsymbol{x}^{(i)}) \\
&= \sum_{i=1}^{n} \log \left(\sum_{z=1}^{m} P_\Phi(\boldsymbol{x}^{(i)}, z) \right) \\
&= \sum_{i=1}^{n} \log \left(\sum_{z=1}^{m} \pi(z) \times \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left|\left| \boldsymbol{x}^{(i)} - \mu_z \right|\right|_2^2 \right) \right)
\end{aligned}
$$

Again, finding valid parameter values $\pi(z)$ and μ_z that maximize this **marginalized** log likelihood is not as trivial (there is no closed-form solution).

Log Likelihood of Unlabeled Data

If $U = \{\boldsymbol{x}^{(i)}\}_{i=1}^{n}$ is a set of n iid unlabeled samples, the log likelihood of U under Φ is

$$\begin{aligned} L_U(\Phi) &= \sum_{i=1}^{n} \log P_\Phi(\boldsymbol{x}^{(i)}) \\ &= \sum_{i=1}^{n} \log \left(\sum_{z=1}^{m} P_\Phi(\boldsymbol{x}^{(i)}, z) \right) \\ &= \sum_{i=1}^{n} \log \left(\sum_{z=1}^{m} \pi(z) \times \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \left|\left| \boldsymbol{x}^{(i)} - \mu_z \right|\right|_2^2 \right) \right) \end{aligned}$$

Again, finding valid parameter values $\pi(z)$ and μ_z that maximize this **marginalized** log likelihood is not as trivial (there is no closed-form solution).

EM is useful here again because each iteration *does* have a trivial solution.

Posterior Probabilities

At each iteration t, we use the current parameter estimates

$$\Phi^{(t)} = \left\{ \pi^{(t)}(z),\ \mu_z^{(t)} \right\}$$

to calculate the **posterior probabilities** on *individual* samples $\boldsymbol{x}^{(i)}$. This can again be easily precomputed by Bayes rule: for every $i \in \{1 \dots n\}$ and $z \in \{1 \dots m\}$, calculate

$$P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) = \frac{P_{\Phi^{(t)}}(\boldsymbol{x}^{(i)}, z)}{P_{\Phi^{(t)}}(\boldsymbol{x}^{(i)})} = \frac{\pi^{(t)}(z) \times \mathcal{N}(\boldsymbol{x}|\mu_z^{(t)}, I_d)}{\sum_{z=1}^{m} \pi^{(t)}(z) \times \mathcal{N}(\boldsymbol{x}|\mu_z^{(t)}, I_d)}$$

Expected Log Likelihood of Labeled Data Under $\Phi^{(t)}$

$$\sum_{i=1}^{n}\sum_{z=1}^{m} P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) \times \log P_{\Phi}(\boldsymbol{x}^{(i)}, z)$$

Expected Log Likelihood of Labeled Data Under $\Phi^{(t)}$

$$\sum_{i=1}^{n}\sum_{z=1}^{m} P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) \times \log P_{\Phi}(\boldsymbol{x}^{(i)}, z)$$

$$= \sum_{i=1}^{n}\sum_{z=1}^{m} P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) \times \left(\log \pi(z) - \frac{1}{2}\left|\left|\boldsymbol{x}^{(i)} - \mu_z\right|\right|_2^2 - \log\sqrt{2\pi}\right)$$

Expected Log Likelihood of Labeled Data Under $\Phi^{(t)}$

$$\sum_{i=1}^{n}\sum_{z=1}^{m} P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) \times \log P_{\Phi}(\boldsymbol{x}^{(i)}, z)$$

$$= \sum_{i=1}^{n}\sum_{z=1}^{m} P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) \times \left(\log \pi(z) - \frac{1}{2}\left|\left|\boldsymbol{x}^{(i)} - \mu_z\right|\right|_2^2 - \log\sqrt{2\pi}\right)$$

$$= \sum_{z=1}^{m} \underbrace{\sum_{i=1}^{n} P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)})}_{\widehat{\textbf{count}}_t(z)} \log \pi(z) - \sum_{i=1}^{n}\sum_{z=1}^{m} P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) \left|\left|\boldsymbol{x}^{(i)} - \mu_z\right|\right|_2^2$$

Expected Log Likelihood of Labeled Data Under $\Phi^{(t)}$

$$\sum_{i=1}^{n}\sum_{z=1}^{m} P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) \times \log P_{\Phi}(\boldsymbol{x}^{(i)}, z)$$

$$= \sum_{i=1}^{n}\sum_{z=1}^{m} P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) \times \left(\log \pi(z) - \frac{1}{2}\left|\left|\boldsymbol{x}^{(i)} - \mu_z\right|\right|_2^2 - \log\sqrt{2\pi}\right)$$

$$= \sum_{z=1}^{m} \underbrace{\sum_{i=1}^{n} P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)})}_{\widehat{\mathbf{count}}_t(z)} \log \pi(z) - \sum_{i=1}^{n}\sum_{z=1}^{m} P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) \left|\left|\boldsymbol{x}^{(i)} - \mu_z\right|\right|_2^2$$

MLE in the t-th iteration of EM

$$\pi(z) = \frac{\widehat{\mathbf{count}}_t(z)}{n} \qquad \mu_z = \frac{\sum_{i=1}^{n} P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) x^{(i)}}{\widehat{\mathbf{count}}_t(z)}$$

EM for GMMs

1. Initialize GMM parameters $\Phi^{(0)}$.
2. For $t = 0 \dots T-1$,
   2.1 For $i = 1 \dots n$ and $y = 1 \dots m$, calculate current posterior

$$P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) = \frac{\pi^{(t)}(z) \times \mathcal{N}(\boldsymbol{x}|\mu_z^{(t)}, I_d)}{\sum_{z=1}^{m} \pi^{(t)}(z) \times \mathcal{N}(\boldsymbol{x}|\mu_z^{(t)}, I_d)}$$

   2.2 Set $\Phi^{(t+1)} = \left\{\pi^{(t+1)}(z),\ \mu_z^{(t+1)}\right\}$ by

$$\pi(z) = \frac{\widehat{\mathbf{count}}_t(z)}{n} \qquad \mu_z = \frac{\sum_{i=1}^{n} P_{\Phi^{(t)}}(z|\boldsymbol{x}^{(i)}) x^{(i)}}{\widehat{\mathbf{count}}_t(z)}$$

3. Return $\Phi^{(T)}$.

Overview

- EM for Naive Bayes
  - Maximum Likelihood Estimation with Labeled Data
  - Maximum Likelihood Estimation with Unlabeled Data
- EM for Gaussian Mixture Models
  - Maximum Likelihood Estimation with Labeled Data
  - Maximum Likelihood Estimation with Unlabeled Data
- k-Means Clustering

Non-Probabilistic Clustering

- You can train a GMM Φ with k clusters with EM and obtain a "soft" k-clustering given by the posterior

$$P_\Phi(z|\boldsymbol{x}^{(i)}) = \frac{\pi(z) \times \mathcal{N}(\boldsymbol{x}|\mu_z, I_d)}{\sum_{z=1}^{k} \pi(z) \times \mathcal{N}(\boldsymbol{x}|\mu_z, I_d)}$$

- If all you want is to cluster n points $\boldsymbol{x}^{(1)} \dots \boldsymbol{x}^{(n)} \in \mathbb{R}^d$ into k clusters, you can do k-**means clustering**.

k-Means Clustering

Input: points $U = \left\{\boldsymbol{x}^{(i)}\right\}_{i=1}^{n}$ in $\mathbb{R}^d$, number of clusters k, T
Output: cluster assignments $a_1 \dots a_n \in \{1 \dots k\}$

1. Initialize centroids: $\boldsymbol{\nu}_1^{(0)} \dots \boldsymbol{\nu}_k^{(0)} \in \mathbb{R}^d$.
2. For $t = 0 \dots T-1$,
   2.1 Assign each point to its closest centroid:

$$a_i^{(t)} \leftarrow \arg\min_{j=1}^{k} \left|\left|\boldsymbol{x}^{(i)} - \boldsymbol{\nu}_j^{(t)}\right|\right|_2^2$$

   2.2 Update centroids: denoting $C_j^{(t)} := \left\{\boldsymbol{x}^{(i)} : a_i^{(t)} = j\right\}$,

$$\boldsymbol{\nu}_j^{(t+1)} \leftarrow \frac{1}{\left|C_j^{(t)}\right|} \sum_{\boldsymbol{x} \in C_j^{(t)}} \boldsymbol{x}$$

3. Return $a_i^{(T)} \leftarrow \arg\min_{j=1}^{k} \left|\left|\boldsymbol{x}^{(i)} - \boldsymbol{\nu}_j^{(T)}\right|\right|_2^2$.

Loss of k-Means Clustering

Using indicator $[[A]]$ which is 1 if A is true and 0 otherwise,

$$L(\boldsymbol{\nu}_1 \dots \boldsymbol{\nu}_k, a_1 \dots a_n) := \sum_{i=1}^{n} \sum_{j=1}^{k} [[a_i = j]] \left|\left| \boldsymbol{x}^{(i)} - \boldsymbol{\nu}_j \right|\right|_2^2$$

Loss of k-Means Clustering

Using indicator $[[A]]$ which is 1 if A is true and 0 otherwise,

$$L(\boldsymbol{\nu}_1 \dots \boldsymbol{\nu}_k, a_1 \dots a_n) := \sum_{i=1}^{n} \sum_{j=1}^{k} [[a_i = j]] \left|\left| \boldsymbol{x}^{(i)} - \boldsymbol{\nu}_j \right|\right|_2^2$$

k-means is an **alternating minimization** algorithm for this loss.

1. Fix centroids $\boldsymbol{\nu}_1 \dots \boldsymbol{\nu}_k$, optimize over assignments $a_1 \dots a_n$:

$$a_i \leftarrow \arg\min_{j=1}^{k} \left|\left| \boldsymbol{x}^{(i)} - \boldsymbol{\nu}_j \right|\right|_2^2$$

Loss of k-Means Clustering

Using indicator $[[A]]$ which is 1 if A is true and 0 otherwise,

$$L(\boldsymbol{\nu}_1 \dots \boldsymbol{\nu}_k, a_1 \dots a_n) := \sum_{i=1}^{n} \sum_{j=1}^{k} [[a_i = j]] \left|\left| \boldsymbol{x}^{(i)} - \boldsymbol{\nu}_j \right|\right|_2^2$$

k-means is an **alternating minimization** algorithm for this loss.

1. Fix centroids $\boldsymbol{\nu}_1 \dots \boldsymbol{\nu}_k$, optimize over assignments $a_1 \dots a_n$:

$$a_i \leftarrow \arg\min_{j=1}^{k} \left|\left| \boldsymbol{x}^{(i)} - \boldsymbol{\nu}_j \right|\right|_2^2$$

2. Fix assignments $a_1 \dots a_n$, optimize over centroids $\boldsymbol{\nu}_1 \dots \boldsymbol{\nu}_k$:

$$\boldsymbol{\nu}_j \leftarrow \arg\min_{\boldsymbol{w} \in \mathbb{R}^d} \sum_{i=1:\, a_i = j}^{n} \left|\left| \boldsymbol{x}^{(i)} - \boldsymbol{w} \right|\right|_2^2 = \frac{1}{|C_j|} \sum_{\boldsymbol{x} \in C_j} \boldsymbol{x}$$

Thus k-means can only decrease the loss in each step.

Generalization of k-Means

Choose a "distortion" function $D(\boldsymbol{x}, \boldsymbol{y}) \geq 0$ and do alternating minimization of

$$L(\boldsymbol{\nu}_1 \ldots \boldsymbol{\nu}_k, a_1 \ldots a_n) := \sum_{i=1}^{n} \sum_{j=1}^{k} [[a_i = j]] \, D\left(\boldsymbol{x}^{(i)}, \boldsymbol{\nu}_j\right)$$

1. Fix centroids $\boldsymbol{\nu}_1 \ldots \boldsymbol{\nu}_k$, optimize over assignments $a_1 \ldots a_n$:

$$a_i \leftarrow \arg\min_{j=1}^{k} D\left(\boldsymbol{x}^{(i)}, \boldsymbol{\nu}_j\right)$$

2. Fix assignments $a_1 \ldots a_n$, optimize over centroids $\boldsymbol{\nu}_1 \ldots \boldsymbol{\nu}_k$:

$$\boldsymbol{\nu}_j \leftarrow \arg\min_{\boldsymbol{w} \in \mathbb{R}^d} \sum_{i=1:\, a_i = j}^{n} D\left(\boldsymbol{x}^{(i)}, \boldsymbol{w}\right)$$

Choice of Distortion Function

- The standard k-means clustering uses squared Euclidean distance $D(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} - \boldsymbol{y}||_2^2$ and

$$\arg\min_{\boldsymbol{w} \in \mathbb{R}^d} \sum_{i=1:\, a_i = j}^{n} D\left(\boldsymbol{x}^{(i)}, \boldsymbol{w}\right)$$

  is given by the **mean** of C_j.

- It turns out that for a wide class of distortion functions called the **Bregman divergence**, this optimization is always given by the mean of C_j.

- Examples of Bregman divergence: squared Euclidiean norm, KL divergence (this only makes sense if data points are probability distributions).

- So we can swap in any Bregman divergence and perform exactly the same updates.

k-Medians Clustering

- Use the Manhattan distance in the algorithm:

$$D(\boldsymbol{x}, \boldsymbol{y}) = ||\boldsymbol{x} - \boldsymbol{y}||_1 := \sum_{l=1}^{d} |x_l - y_l|$$

- The solution of

$$\underset{\boldsymbol{w} \in \mathbb{R}^d}{\arg\min} \sum_{i=1:\, a_i = j}^{n} \left|\left| \boldsymbol{x}^{(i)} - \boldsymbol{w} \right|\right|_1$$

  is given by the **element-wise median** of C_j.