Recitation 7

Zhenzhen Weng

Oct 15, 2015

1 Semi-Supervised Learning

Consider learning a Gaussian Naive Bayes Classifier with both labeled and unlabeled data. Suppose we have labeled data $(x^1, y^1), ..., (x^l, y^l)$, and unlabeled data $x^{l+1}, ..., x^{l+u}$, where each $x^n \in \mathcal{R}^f$ is a sample in f dimensional space.

- For the labeled data $n = 1, ..., l$, we have $y^n \in \{1, 2, ..., K\}$ which are the labels that determine the class of x^n, i.e., $z_n^k = P(y^n = k|x^n; \theta) = \delta(y^n = k)$.
- For the unlabeled data $n = l+1, ..., l+u$, we want to assign soft class membership $\langle z_n^k \rangle = P(z_n = k|x^n; \theta)$ for each x^n. Note that we handle the unlabeled data using the Gaussian Mixture Model, where the soft class membership corresponds to the latent component in GMM that generates the observations.

After obtaining both memberships we can improve the parameter estimation of $\theta = \{\pi_k, \mu_{ik}, \sigma_{ik}^2\}$ for each feature i and each class k, where π_k is the probability of class k, and μ_{ik}, σ_{ik}^2 represent the Gaussian mean and variance for the i_{th} feature and k_{th} class.

1. Write the objective function for semi-supervised EM.
   For the labeled data, We want to maximize the log likelihood and for the unlabeled data, we want to maximize the expected complete

data log likelihood. Therefore, our objective is

$$\begin{aligned}
&\max_{\theta} \sum_{i=1}^{l} \log \prod_{k} \pi_k^{z_i^k} + \sum_{i=1}^{l} \log \prod_{k} \mathcal{N}(x^i; \mu_k, \sigma_k^2)^{z_k^i} \\
&\quad + \sum_{i=l+1}^{l+u} \log \prod_{k} \pi_k^{\langle z_i^k \rangle} + \sum_{i=l+1}^{l+u} \log \prod_{k} \mathcal{N}(x^i; \mu_k, \sigma_k^2)^{\langle z_k^i \rangle} \\
= &\max_{\theta} \sum_{i=1}^{l} \sum_{k=1}^{K} z_i^k \log \pi_k + \sum_{i=1}^{l} \sum_{k=1}^{K} z_k^i \log \mathcal{N}(x^i; \mu_k, \sigma_k^2) \\
&\quad + \sum_{i=l+1}^{l+u} \sum_{k=1}^{K} \langle z_i^k \rangle \log \pi_k + \sum_{i=l+1}^{l+u} \sum_{k=1}^{K} \langle z_k^i \rangle \log \mathcal{N}(x^i; \mu_k, \sigma_k^2)
\end{aligned}$$

2. In the E-step, we assign probabilistic labels $\langle z_n^k \rangle$ to the unlabeled data using parameters from the previous M-step. Derive the formula for estimating $\langle z_n^k \rangle$ for the unlabeled data.

$$\begin{aligned}
\langle z_n^k \rangle =& P(z_n = k | x^n; \hat{\theta}) \\
=& \frac{P(x^n | z_n = k; \hat{\theta}) P(z_n = k)}{\sum_{j=1}^{K} P(x^n | z_n = j; \hat{\theta}) P(z_n = j)} \\
=& \frac{\mathcal{N}(x^n; \hat{\mu}_k, \hat{\sigma}_k^2) \hat{\pi}_k}{\sum_{j=1}^{K} \mathcal{N}(x^n; \hat{\mu}_j, \hat{\sigma}_j^2) \hat{\pi}_j}
\end{aligned}$$

where

$$\mathcal{N}(x^n; \hat{\mu}_j, \hat{\sigma}_j^2) = \frac{1}{\sqrt{2\pi}\hat{\sigma}_j} e^{-\frac{(x^n - \hat{\mu}_j)^2}{2\hat{\sigma}_j^2}}$$

and $\hat{\theta} = \{\hat{\pi}_k, \hat{\mu}_k, \hat{\sigma}_j^2\}$ are the parameters from the previous M-step.

3. In the M-step, derive the formula for updating parameters $\{\pi_k, \mu_{ik}, \sigma_{ik}^2\}$. In the M-step, we retrain the classifier using both the labeled data and unlabeled data. So we find the MLE estimates of $\{\pi_k, \mu_{ik}, \sigma_{ik}^2\}$ by

maximizing the objective function.

$$\begin{aligned}
\hat{\pi}_{k,new} =& \frac{\sum_{i=1}^{l} z_i^k + \sum_{i=l+1}^{l+u} \langle z_i^k \rangle}{l+u} \\
\hat{\mu}_{k,new} =& \frac{\sum_{i=1}^{l} z_i^k x_i + \sum_{i=l+1}^{l+u} \langle z_i^k \rangle x_i}{\sum_{i=1}^{l} z_i^k + \sum_{i=l+1}^{l+u} \langle z_i^k \rangle} \\
\hat{\sigma}_{k,new}^2 =& \frac{\sum_{i=1}^{l} z_i^k (x_i - \hat{\mu}_{k,new})^2 + \sum_{i=l+1}^{l+u} \langle z_i^k \rangle (x_i - \hat{\mu}_{k,new})^2}{\sum_{i=1}^{l} z_i^k + \sum_{i=l+1}^{l+u} \langle z_i^k \rangle}
\end{aligned}$$

where $\langle z_i^k \rangle$ is calculated in the previous E-step.

2 Conditional Independence

$$
\begin{aligned}
& X \perp Y | Z \\
\Leftrightarrow\ & P(X|Z)P(Y|Z) = P(X,Y|Z) \\
\Leftrightarrow\ & P(X|Z) = P(X|Y,Z)
\end{aligned}
$$

2.0.1 Example

X,Y are independent flips of fair coin. Z = XOR(X, Y). Are X and Y independent conditional on Z?

$$
\begin{aligned}
P(X=0|Z=0) &= \frac{P(X=0,Z=0)}{P(Z=0)} && (1)\\
&= \frac{P(X=0,Z=0)}{P(X=0,Z=0)+P(X=1,Z=0)} && (2)\\
&= \frac{0.25}{0.5} && (3)\\
&= \frac{1}{2} && (4)\\
P(Y=0|Z=0) &= \frac{P(Y=0,Z=0)}{P(Z=0)} && (5)\\
&= \frac{P(Y=0,Z=0)}{P(Y=0,Z=0)+P(Y=1,Z=0)} && (6)\\
&= \frac{0.25}{0.5} && (7)\\
&= \frac{1}{2} && (8)\\
P(X=0,Y=0|Z=0) &= \frac{P(X=0,Y=0,Z=0)}{P(Z=0)} && (9)\\
&= \frac{P(X=0,Y=0,Z=0)}{P(Y=0,Z=0)+P(Y=1,Z=0)} && (10)\\
&= \frac{0.25}{0.5} && (11)\\
&= \frac{1}{2} && (12)
\end{aligned}
$$

Therefore,

$$P(X|Z=0)P(Y|Z=0) \neq P(X,Y|Z=0)$$

Independence does not imply conditional independence.

2.0.2 Example

A box contains two coins: a regular coin and one fake two-headed coin (P(H)=1). Choose a coin at random and toss it twice. Define the following events.

• A= First coin toss results in an H.

• B= Second coin toss results in an H.

• C= Coin 1 (regular) has been selected.

Note that A and B are not independent, but they are conditionally independent given C. Therefore, conditional independence does not imply independence.