EECS 445: Machine Learning **Fall 2016**

Discussion 8: EM

Written by: Benjamin R. Bray and Chansoo Lee

This is a draft

Motivation: Partially observed data

8.1 Unsupervised Naive Bayes

We get unlabeled training data $\{\mathbf{x}^{(i)}\}_{i=1}^n$ where each data is a binary feature vector: $\mathbf{x}^{(i)} \in \{0,1\}^D$.

We don't have sufficient statistics (N_c and N_{cdm}) to compute the MLE for Naive Bayes, because we are missing class labels. A simple approach is to arbitrariliy or randomly fill in the labels and then compute the MLE. This approach, however, ignores the probabilistic relationship between features and the class label.

This problem is difficult because we are trying to solve two problems together: completing missing values and estimating the model parameters. Each problem, however, is easy when we have the solution to the other. Given complete data, we can estimate parameters. Given model parameters, we can find the distribution of the missing values $P(y_d = c|x;\theta,\pi) \propto \pi_c \prod_{d=1}^D \theta_{cdx_d}$. To solve this "chicken and egg" problem, we use an alternating optimization technique called EM.

8.1.1 Soft assignment EM

- Initiliaze the parameters to the values that will break the symmetry, and then repeat:
- In the E-step, we fill in the missing data. We often say that we compute the *expected sufficient statistics*, because we fill in the missing data only probabilistically.
- In the M-step, update the model parameters to be the MLE based on new expected sufficient statistics.

Suppose the two data points are $\mathbf{x}^{(1)} = (0,1)$ and $\mathbf{x}^{(2)} = (1,1)$.

Case Study: Binary Unsupervised Naive Bayes Initilaize $\pi_c = 0.7, \theta_{111} = 0.9, \theta_{011} = 0.3, \theta_{121} = 0.6, \theta_{021} = 0.2$.

Compute the probability for each assignment of missing values:

y^1	y^2	unnormalized probability	probability
1	1	$(0.7)(0.1)(0.6)(0.7)(0.9)(0.6) = 0.015876$	0.4773
1	0	$(0.7)(0.1)(0.6)(0.3)(0.3)(0.2) = 0.000756$	0.0227
0	1	$(0.3)(0.7)(0.2)(0.7)(0.9)(0.6) = 0.015876$	0.4773
0	0	$(0.3)(0.7)(0.2)(0.3)(0.3)(0.2) = 0.000756$	0.0227

Compute the sufficient statistics:
$\overline{N_1} = (0.4773)(2) + (0.0227) + (0.4773) = 1.4546$
$\overline{N_{111}} = (0.4773) + (0.4773) = 0.9546$
$\overline{N_{121}} = (0.4773)2 + (0.0227) + (0.4773) = 1.4546$
M-step:
$\pi_1 = 1.4546/2 = 0.7273.$
$\theta_{111} = N_{111}/N_1 = 0.6563.$
$\theta_{121} = N_{121}/N_1 = 1.$

Efficient E-step Instead of the joint distribution of missing values, we can consider them independently. Mathematically,

$$\begin{aligned} N_1 &= P(y^1 = 1, y^2 = 1) + P(y^1 = 1, y^2 = 1) + P(y^1 = 1, y^2 = 0) + P(y^1 = 0, y^2 = 1) \\ &= (P(y^1 = 1, y^2 = 1) + P(y^1 = 1, y^2 = 0)) + (P(y^1 = 1, y^2 = 1) + P(y^1 = 0, y^2 = 1)) \\ &= P(y^1 = 1) + P(y^2 = 1) \end{aligned}$$

With complete data, each instance contributes a full count of 1. With missing data, each instance contributes a fraction of the full count to each category such that it sums to 1.

Compute the probability for each assignment of missing values:

$$P(y^1 = 1) = \frac{(0.7)(0.1)(0.6)}{(0.7)(0.1)(0.6) + (0.3)(0.7)(0.2)} = 0.5$$

$$P(y^2 = 1) = \frac{(0.7)(0.9)(0.6)}{(0.7)(0.9)(0.6) + (0.3)(0.3)(0.2)} = 0.9546$$

8.1.2 Hard assignment EM

Sometimes it is difficult (due to mathematical complexity or limited computation time) to handle fractional contribution of each data point. In that case, we simply assign the missing value with the highest probability, breaking ties arbitrarily.

Example: k-means as a hard-assignment EM for GMM clustering.

Excercise 8.1. *What are the differences between soft and hard assignment EM?*

The hard-assignment EM explores the combinatorial space of missing variable assignments. The soft-assignment EM, on the other hand, explores the continuous space.

In clustering, the hard assignment EM tends to amplify the contrast among classes, while soft assignment EM attempts to model mixed-class memberships.

8.1.3 Random assignment EM

A slight variation of the hard assignment. We sample the missing values according to the computed distribution.