

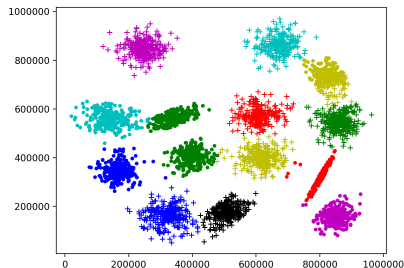
Mixture of Gaussians Clustering

Shrey Gupta

Applied Machine Learning (HOUSECS 59-01), Duke University

December 5, 2018

Unsupervised Learning



- ▶ Data is unlabeled (no “ground truth”).
- ▶ Problems: clustering, density estimation, and pattern detection.

Clustering

- ▶ Most common unsupervised problem is clustering.
- ▶ Can we separate the data into different clusters, each with a given (but not necessarily the same class) distribution?
- ▶ We can then analyze the underlying properties of each cluster.

Hard vs. Soft Clustering

- ▶ Hard clustering: assign each point to a cluster.
- ▶ Soft clustering: assign a probability γ_{ik} that each point x_i belongs to the k^{th} cluster.

Example

- ▶ Crime is happening on the streets of Gotham City!
- ▶ There are $n = 2$ criminals: Bane and the Joker. Suppose every night, one of the two decides to commit a series of crimes.
- ▶ Bane succeeds 50% of the time, and the Joker 70% (the Joker is more skilled). That is, $P(\text{success}) = 0.5$ for Bane and $P(\text{success}) = 0.7$ for the Joker on each attempt.
- ▶ Neither criminal is identified nor caught during each attempt.
- ▶ Over a series of $m = 100$ nights, $j = 10$ crimes are attempted by one of the two criminals.

Example

- ▶ Suppose we know there are $n = 2$ criminals, and the number of crimes (from $j = 10$ attempts) succeeded during each of $m = 100$ nights.
- ▶ However, we don't know which criminal committed the series of crimes each night, and $P(\text{success})$ for each criminal.
- ▶ Can we recover a probability that a night's crimes were committed by a given criminal, and $P(\text{success})$ for each criminal?
- ▶ Yes! Utilize the expectation-maximization (EM) algorithm.

EM Algorithm

- ▶ Powerful soft clustering algorithm that allows us to maximize likelihood (of clusters) even with several missing data.

EM Algorithm

- ▶ Randomly initialize the parameters θ of the n distributions (clusters).
- ▶ E-step: compute the probabilities γ_{ik} that each point x_i belongs to the k^{th} cluster:

$$\gamma_{ik} = P(z_i = k | x_i, \theta^{(t)}).$$

- ▶ M-step: maximize a lower bound on the likelihood of an estimate of the new parameters $\theta^{(t+1)}$.
- ▶ Repeat until convergence.

Mixture of Gaussians

- ▶ Specific case of EM-algorithm: assumes each cluster is normally distributed with mean μ and variance Σ .
- ▶ E-step: compute $\gamma_{ik}^{(t+1)} = \frac{P(X_i=x_i|z_i=k, \theta^{(t)})P(z_i=k|\theta^{(t)})}{P(X_i=x_i|\theta^{(t)})}$.
- ▶ M-step: update w_k , μ_k , and Σ_k :

$$w_k^{(t+1)} = \frac{\sum_i \gamma_{ik}^{(t+1)}}{n}, \mu_k^{(t+1)} = \frac{\sum_i x_i \gamma_{ik}^{(t+1)}}{\sum_i \gamma_{ik}^{(t+1)}}, \text{ and}$$

$$\Sigma_k^{(t+1)} = \frac{\sum_i \gamma_{ik}^{(t+1)} (x_i - \mu_k^{(t+1)})(x_i - \mu_k^{(t+1)})^T}{\sum_i \gamma_{ik}^{(t+1)}}.$$

Mixture of Gaussians

- ▶ Understanding the derivations of the formulas is beyond the scope of this class.
- ▶ Instead, understand the interpretation of the EM algorithm, with GMM as a specific case, and when it can be utilized.
- ▶ We can also apply it to our example.

Example

- ▶ Let's return to our example: $n = 2$ criminals, and $j = 10$ attempted crimes from $m = 100$ nights.
- ▶ We don't know which criminal committed the series of crimes each night, and $P(\text{success})$ for each criminal.
- ▶ Today's notebook contains a snippet of code to generate this data. Let's assume it is normally distributed.

Example

- ▶ We can implement a mixture of Gaussians model. Let's do that using the notebook.
- ▶ Observe γ_{ik} , w_k , μ_k over the k criminals. What is the interpretation of each of these values?

Interpretation

- ▶ γ_{ik} : probability criminal k committed crime on night i .
- ▶ w_k : proportion of nights criminal k commits a crime.
- ▶ μ_k : average number of successful crimes per night for criminal k .

Applications

- ▶ But we still don't have labels on the clusters! We don't know whether it was Bane or the Joker who is criminal k .
- ▶ In general, this is a problem for unlabeled data. What (or who) do the clusters represent?
- ▶ We can correlate the parameters of each cluster (distribution). Say the data also included the locations (in coordinates) where the crime was committed, or the hour at which it was committed.
- ▶ Questions we can ask: at what hour does criminal k commit a crime, and where?

Applications

- ▶ Say over a few of the j nights, a witness comes out to identify the criminal who performed the crime on that night.
- ▶ We can then assign a probability that Bane and the Joker are criminal k , and probabilities that they committed crimes over each of the j nights.
- ▶ Predictions of assigning Bane or the Joker to criminal k get stronger the more identifications (labels) we have. But the more labeled data we have, the less the need for the EM algorithm.

Limitations

- ▶ Assumes clusters arise from the same distribution (in the mixture of Gaussians case, from the normal distribution).
- ▶ Can be extremely slow: works well for low-dimensional data with few missing data.