CSC 311: Introduction to Machine Learning

Tutorial 11 - EM algorithm

Fall 2023

University of Toronto, Fall 2022

Bernoulli Distribution

- Bernoulli distribution: $\mathbf{X}$ is a random variable with two outcomes. We say that $\mathbf{X}$ follows $Ber(\mu)$ if:

$$P(\mathbf{X}=x)=\mu^x(1-\mu)^{1-x}, x\in\{0,1\}$$

- Example: A coin follows Bernoulli distribution.

Mixture of Bernoulli

- Mixture of Gaussians is defined over continuous variables. Mixture of Bernoulli can be seen as its counterpart for binary variables.

Mixture of Bernoulli

- Mixture of Gaussians is defined over continuous variables. Mixture of Bernoulli can be seen as its counterpart for binary variables.
- Assume a datapoint x is generated as follows:
  - Choose a cluster z from $\{1, ..., K\}$ such that $p(z = k) = \pi_k$
  - Given z, sample x from $Ber(\mu_k)$.
- We say x follows mixtures of Bernoulli distributions. It pmf can be expressed as:

$$P(x) = \sum_{i=1}^{k} \pi_i \mu_i^x (1 - \mu_i)^{1-x}$$

Maximum Likelihood

- We want to learn the parameters $\{\pi_k, \mu_k\}$ from the observations $\{x_i\}_{i=1}^{n}$.

Maximum Likelihood

- We want to learn the parameters $\{\pi_k, \mu_k\}$ from the observations $\{x_i\}_{i=1}^n$.
- Training objective: Maximize log likelihood

$$\max_{\pi_k, \mu_k} \sum_{n=1}^{N} \log \sum_{i=1}^{K} \pi_k \mu_k^{x_n} (1 - \mu_k)^{1 - x_n}$$

- $\underline{\text{Log}}$ inside $\underline{\text{sum}}$: EM algorithm

E step - compute the posterior

- Compute the posterior probability $z_{nk} = P(z_n = k | x_n)$ using Bayes' theorem:

E step - compute the posterior

- Compute the posterior probability $z_{nk} = P(z_n = k|x_n)$ using Bayes' theorem:

$$P(z_n = k|x_n) = \frac{P(z_n = k, x_n)}{P(x_n)}$$
$$= \frac{\pi_k \mu_k^{x_n}(1-\mu_k)^{1-x_n}}{\sum_{i=1}^{k} \pi_i \mu_i^{x_n}(1-\mu_i)^{1-x_n}}$$

- z_{nk} can be interpreted as how much we think a cluster k is responsible for generating a datapoint x_n.

M step - optimize the joint log likelihood

- Why do we need these responsibilities again?

$$\log p(D) = \sum_{n=1}^{N} \log p(x_n) = \sum_{n=1}^{N} \log \sum_{k=1}^{K} p(x_n | z_n = k)\, p(z_n = k)$$

$$p(x_n) = \sum_{k=1}^{K} p(x_n, z_n = k) = \underline{\sum_{k=1}^{K} p(x_n | z_n = k)\, p(z_n = k)}$$

- The expected joint likelihood can be expressed as:

$$\mathbb{E}_{p(z|x)}[\log p(\mathbf{X}, \mathbf{Z}; \mu, \pi)] = \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}_{p(z|x)}\left[\log\left(\pi_k \mu_k^{x_n} (1-\mu_k)^{(1-x_n)}\right)\right]$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} (\log \pi_k + x_n \log \mu_k + (1 - x_n) \log(1 - \mu_k))$$

- If we knew the mixture assignments (deterministically) our lives would be a lot easier!

$$\log p(X, Z; \mu, \pi) = \sum_{n=1}^{N} \log p(x_n, z_n; \mu, \pi) = \sum_{n=1}^{N} \log p(x_n | z_n) + \log p(z_n)$$

M step - optimize the joint log likelihood

- The $\overset{\text{expected}}{\vee}$ joint likelihood can be expressed as:

$$\mathbb{E}_{p(z|x)}\left[\log p(\mathbf{X}, \mathbf{Z}; \mu, \pi)\right]$$
$$= \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} (\log \pi_k + x_n \log \mu_k + (1 - x_n) \log(1 - \mu_k))$$

- Assume the responsibility is known, setting the derivative respect to μ_k to zero, we get:

$$\mu_k = \frac{\sum_{n=1}^{N} z_{nk} x_n}{\sum_{n=1}^{N} z_{nk}}$$

- Interpretation: The mean of component k is equal to the weighted mean of the data, with weighted coefficients proportional to the responsibility.

True or False

The EM algorithm optimizes a lower bound on its objective function, which is the marginal likelihood $\prod_i P(x_i)$ of the observed data points $x_1, x_2, ...x_N$.

True or False

The EM algorithm optimizes a lower bound on its objective function, which is the marginal likelihood $\prod_i P(x_i)$ of the observed data points $x_1, x_2, ...x_N$.

True. See slide 46 $\Rightarrow$ Jensen's Inequality

True or False

The EM algorithm is guaranteed to never decrease the value of its objective function on any iteration.

True or False

The EM algorithm is guaranteed to never decrease the value of its objective function on any iteration.

True $\Rightarrow$ Slides 52-54

Each step optimizes a lower bound!

True or False

The objective function optimized by the EM algorithm can also be optimized by a gradient descent algorithm which will find the global optimal solution, whereas EM finds its solution more quickly but may return only a locally optimal solution.

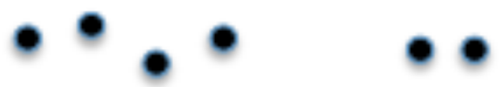
True or False

The objective function optimized by the EM algorithm can also be optimized by a gradient descent algorithm which will find the global optimal solution, whereas EM finds its solution more quickly but may return only a locally optimal solution.

False $\Rightarrow$ Slide 30

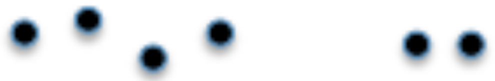
True or False

Consider the set of training data below, and two clustering algorithms: K-Means, and a Gaussian Mixture Model (GMM) trained using EM. These two clustering algorithms will produce the same cluster centers (means) for this data set.



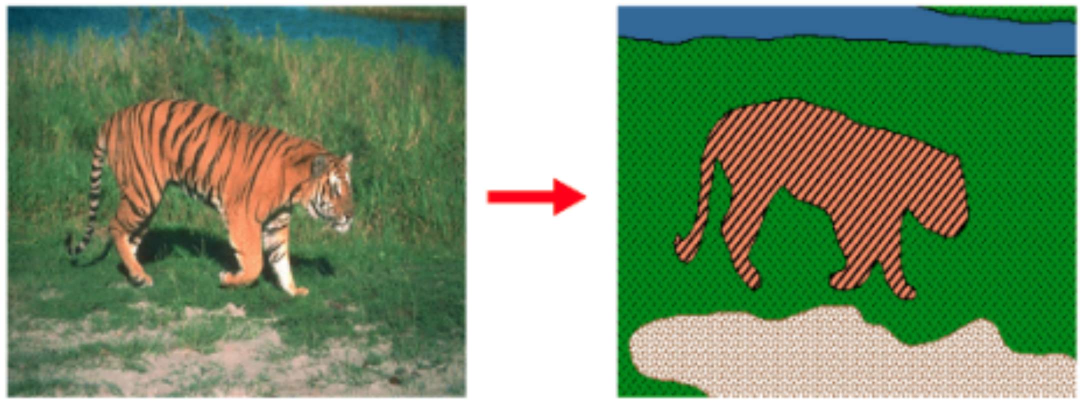
True or False

Consider the set of training data below, and two clustering algorithms: K-Means, and a Gaussian Mixture Model (GMM) trained using EM. These two clustering algorithms will produce the same cluster centers (means) for this data set.



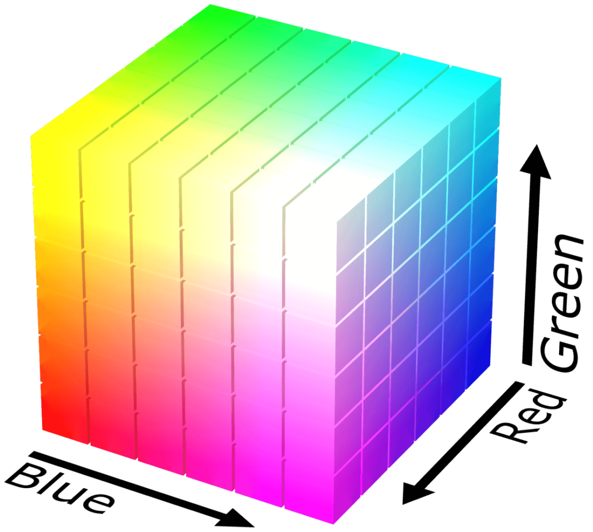
False. In k-means, the means of the clusters are determined by an average of the points assigned to that cluster, but in GMM the means of each cluster are (differently) weighted averages of all points.

Application: EM for image segmentation



Partition an image into regions each of which has a reasonably homogenous visual appearance

RGB image



- Each pixel in an RGB image is a point in 3-dimensional space comprising the intensities of the red, blue and green channels.
- We can think of image segmentation tasks as clustering problems on pixels.
- We can apply EM and k-means for image segmentation.