Machine Learning

CSCI 567 Spring 2021

Discussion: EM

Q1. EM Algorithm We will derive an expectation-maximization (EM) algorithm for clustering a set of images (assuming black and white pixels, no color). The inputs $\mathbf{x}^{(i)}$ can be thought of as vectors of binary values corresponding to black and white pixel values of an image i.

(a) Consider a vector of binary random variables, $\mathbf{X} \in \{0,1\}^D$. Assume each variable X_d is drawn independently from a Bernoulli distribution with parameter p_d. Express probability $P(\mathbf{X} = \mathbf{x}; \mathbf{p})$ in terms of x_d and p_d, where $\mathbf{p} = (p_1, p_2, \cdots p_D)$.

(b) For clustering the set of images, we assume that there are K clusters (groups of images) i.e., images are drawn from a mixture of K Bernoulli distributions, each with parameter $\mathbf{p}^{(k)}$. Let the prior probability of belonging to the k^{th} cluster be w_k. (You may use θ to denote the collection of parameters w_k and $\mathbf{p}^{(k)}$ for all K).

- Write down the log-likelihood of a set of images $(\mathbf{x}^{(1)}, \cdots \mathbf{x}^{(N)})$.
- Write down the complete log-likelihood of the set of images assuming $z^{(i)} \in \{1, 2 \cdots K\}$ as the value of the latent variable for the cluster the image belongs to.

(c) Write the E-step i.e, the posterior $P(Z = k | \mathbf{X} = \mathbf{x}; \theta^{(t)})$ for iteration t of EM.

(d) Write down the M-step for EM using the posterior computed in the E-step.