CSCI567 HW4 Bonus Report

Ritu Pravakar and Charlene Yuen

November 2022

Problem 2: Gaussian Mixture Model and EM

2.2

We can show that with assignments $\{\sigma_j \to 0, \pi_j \in U(0,1)\}_{j=1}^{k}$ we can reduce GMM to k-means.

$$P(z_i = j|\mathbf{x}_i) = \frac{P(z_i = j, \mathbf{x}_i)}{\sum_k P(z_i = k, \mathbf{x}_i)} = \frac{\pi_j N(\mathbf{x}_i \mid \mu_k, \Sigma_k)}{\sum_k \pi_k N(\mathbf{x}_i \mid \mu_k, \Sigma_k)}$$

For the E-step:

$$= \frac{\pi_j \ln \frac{1}{(\sqrt{2\pi})^d |\Sigma_j|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \mu_j)^T \Sigma_j^{-1} (\mathbf{x}_i - \mu_j)\right)}{\sum_k \pi_k \ln \frac{1}{(\sqrt{2\pi})^d |\Sigma_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \mu_k)^T \Sigma_k^{-1} (\mathbf{x}_i - \mu_k)\right)}$$

Let $\epsilon = (\mathbf{x}_i - \mu_j)$ and $\epsilon^* = \min_j(\mathbf{x}_i - \mu_j)$

$$= \frac{\pi_j \ln \frac{1}{(\sqrt{2\pi})^d |\sigma_j^2 I|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\epsilon^* - \epsilon)^T (\sigma_j^2 I)^{-1} (\epsilon^* - \epsilon)\right)}{\sum_k \pi_k \ln \frac{1}{(\sqrt{2\pi})^d |\sigma_k^2 I|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\epsilon^* - \epsilon)^T (\sigma_k^2 I)^{-1} (\epsilon^* - \epsilon)\right)}$$

If x_i isn't closest to μ_j and $\epsilon^* \neq \epsilon$ then $\epsilon^* - \epsilon < 0$:

$$\lim_{\sigma \to 0} \frac{\pi_j \ln \frac{1}{(\sqrt{2\pi})^d |\sigma_j^2 I|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\epsilon^* - \epsilon)^T (\frac{1}{\sigma_j^2} I)(\epsilon^* - \epsilon)\right)}{\sum_k \pi_k \ln \frac{1}{(\sqrt{2\pi})^d |\sigma_k^2 I|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\epsilon^* - \epsilon)^T (\frac{1}{\sigma_k^2} I)(\epsilon^* - \epsilon)\right)}$$

$$= \frac{\pi_j \ln \frac{1}{(\sqrt{2\pi})^d |\sigma_j^2 I|^{\frac{1}{2}}} \exp(-\infty)}{\pi_j \ln \frac{1}{(\sqrt{2\pi})^d |\sigma_j^2 I|^{\frac{1}{2}}} + \sum_{k \neq j} \pi_k \ln \frac{1}{(\sqrt{2\pi})^d |\sigma_k^2 I|^{\frac{1}{2}}} \exp(-\infty)}$$

$$= 0 \text{ since } \lim_{x \to -\infty} e^x = 0$$

Else x_i is closest to μ_j and $\epsilon^* = \epsilon$ then $\epsilon^* - \epsilon = 0$:

$$\lim_{\sigma \to 0} \frac{\pi_j \ln \frac{1}{(\sqrt{2\pi})^d |\sigma_j^2 I|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\epsilon^* - \epsilon)^T (\frac{1}{\sigma_j^2} I)(\epsilon^* - \epsilon)\right)}{\sum_k \pi_k \ln \frac{1}{(\sqrt{2\pi})^d |\sigma_k^2 I|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\epsilon^* - \epsilon)^T (\frac{1}{\sigma_k^2} I)(\epsilon^* - \epsilon)\right)}$$

$$= \frac{\pi_j \ln \frac{1}{(\sqrt{2\pi})^d |\sigma_j^2 I|^{\frac{1}{2}}}}{\pi_j \ln \frac{1}{(\sqrt{2\pi})^d |\sigma_j^2 I|^{\frac{1}{2}}} + \sum_{k \neq j} \pi_k \ln \frac{1}{(\sqrt{2\pi})^d |\sigma_k^2 I|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\epsilon^* - \epsilon)^T (\frac{1}{\sigma_k^2} I)(\epsilon^* - \epsilon)\right)}$$

$$= \frac{\pi_j \infty}{\pi_j \infty} = 1$$

Then: $\gamma_{ij} = 1$ if j is the ith datapoint's closest cluster

$\gamma_{ij} = 0$ otherwise, which reduces the E-step to the k-means cluster assignments.

Since the assignments γ_{ij} are reduced to that of k-means, the cluster center μ_j for GMM calculated in the M-step should remain the same for both models as expected.

Problem 4: PCA for Learning Word Embeddings

4.5: Bonus

The overall accuracy is 0.5534467323187109, which seems barely better than just guessing but decent when taking into consideration that the classifier has to choose among many words to complete the analogy, so narrowing all those possible words down to slightly better than a 50/50 chance isn't too bad. It seems to do poorly on the topics of countries, verbs, and adverbs due to some English conventions that may not logically match up grammatically as PCA finds. For example with countries, for analogy "brazil brazilian netherlands dutch", PCA predicts danish when it expects dutch, when many countries describe their nationalities with a suffix of "ish" (eg. British, Irish) or "an" (eg. Brazilian, American, Indian), but "dutch" doesn't follow those patterns and also doesn't build off the country's name. It also struggles with verbs (ex. "feeding fed sitting sat"), which predicts "seated" instead of "sat", following "fed" with the suffix "ed", but in this case "sat" is the correct past tense term. For a similar reason, it struggles with adverbs. However, it does generally decently with nouns, since there aren't too many variances in singular vs plural form, such as most plural forms simply having a suffix of "s" (building, buildings) or "en" (child, children).