ML Fundamentals: k-means

April 14, 2021

1 k-means

k-means group training data $(x^{(1)}, x^{(2)}, .., x^{(m)})$ into k cohesive clusters in an unsupervised manner. It works as follow

- Initialize cluster centroids $\mu_1, \mu_2, .., \mu_k$ randomly.
- Repeat until convergence:
  - Assign training example to closest centroid. For every i, set

$$c^{(i)} := \arg\min_j \left\| x^{(i)} - \mu_j \right\|^2$$

  - Moving each centroid to the mean of new cluster. For each j, map

$$\mu_j := \frac{\sum_{i=1}^m 1\left\{c^{(i)} = j\right\} x^{(i)}}{\sum_{i=1}^m 1\left\{c^{(i)} = j\right\}}$$

Q. Does K-Means guaranteed to converge? A. Let us define a distortion function,

$$J(c, \mu) = \sum_{i=1}^m \left\| x^{(i)} - \mu_{c^{(i)}} \right\|^2$$

. In first step, it minimizes J by moving c keeping μ fixed, and in the second step it keeps c fixed and choose optimal value of μ to minimize J. So, it is guaranteed to converge.

k-means suffers from the problem of getting trapped **local minima**. Multiple runs would help here, as it would initialize μ randomly.

2 Gaussian Mixture Model and EM

We are given training set $(x^{(1)}, x^{(2)}, .., x^{(n)})$. Lets assume, each $x^{(i)}$ was drawn from one of the k-Gaussians depending on $z^{(i)}$.

- $z^{(i)} \sim$ Multinomial(ϕ), which means, $\phi_j \geq 0, \sum_{j=1}^k \phi_j = 1$
- parameter ϕ_j gives the probability of $z^{(i)} = j$
- Given the state of latent variable $z^{(i)}$, the variable $x^{(i)}$ is drawn from Gaussian determined by $z^{(i)}$.

$$x^{(i)} \mid (z^{(i)} = j) \sim \mathcal{N}(\mu_j, \Sigma_j)$$

- So the parameters are ϕ, μ, Σ
- Likelihood function to estimate these parameters can be written as,

$$\begin{aligned} \ell(\phi, \mu, \Sigma) &= \sum_{i=1}^m \log p\left(x^{(i)}; \phi, \mu, \Sigma\right) \\ &= \sum_{i=1}^m \log \sum_{z^{(i)}=1}^k p\left(x^{(i)} \mid z^{(i)}; \mu, \Sigma\right) p\left(z^{(i)}; \phi\right) \end{aligned}$$

- For the above function $\ell(\phi, \mu, \Sigma)$ it is not possible to find Maximum Likelihood Estimations in the closed form.
- If we make assignment of $z^{(i)}$ fixed, than it becomes easy find MLE estimates. As the equation would be reduced to
$$\ell(\phi, \mu, \Sigma) = \sum_{i=1}^{m} \log p\left(x^{(i)} \mid z^{(i)}; \mu, \Sigma\right) + \log p\left(z^{(i)}; \phi\right)$$
- So, we need to pre-assign $z^{(i)}$ using parameters, $x^{(i)}, \phi, \mu, \Sigma$
- So, it becomes expectation and maximization problem as follow:
- Repeat until convergence
  - E-Step: For each i, j, set:
$$w_j^{(i)} := p\left(z^{(i)} = j \mid x^{(i)}; \phi, \mu, \Sigma\right)$$
  - M-Step: Update the parameters by maximizing
$$\ell(\phi, \mu, \Sigma) = \sum_{i=1}^{m} \log p\left(x^{(i)} \mid z^{(i)}; \mu, \Sigma\right) + \log p\left(z^{(i)}; \phi\right)$$
    , which would be as follow:
$$\begin{aligned}
\phi_j &:= \frac{1}{m} \sum_{i=1}^{m} w_j^{(i)} \\
\mu_j &:= \frac{\sum_{i=1}^{m} w_j^{(i)} x^{(i)}}{\sum_{i=1}^{m} w_j^{(i)}} \\
\Sigma_j &:= \frac{\sum_{i=1}^{m} w_j^{(i)} \left(x^{(i)} - \mu_j\right)\left(x^{(i)} - \mu_j\right)^T}{\sum_{i=1}^{m} w_j^{(i)}}
\end{aligned}$$
- E-step, can be computed using posterior distribution as follow:
$$p\left(z^{(i)} = j \mid x^{(i)}; \phi, \mu, \Sigma\right) = \frac{p\left(x^{(i)} \mid z^{(i)} = j; \mu, \Sigma\right) p\left(z^{(i)} = j; \phi\right)}{\sum_{l=1}^{k} p\left(x^{(i)} \mid z^{(i)} = l; \mu, \Sigma\right) p\left(z^{(i)} = l; \phi\right)}$$
- **EM-algorithm** is reminiscent of k**-means algorithm**, but with one major difference of assigning **"soft"** clusters in E-step, and using Gaussian probabilities in M-step. And, hence, does not make any assumption on probability distribution with in clusters.