DS102 - Discussion 5

Wednesday, 2nd October, 2019

In lecture and lab we considered the situation where we were modelling a set of random variables that were drawn from a Gaussian mixture model (GMM) consisting of two distributions with a known variance. In this discussion we will extend Gaussian mixture models to consist of more than two Gaussians with unknown variance. Time permitting we will also review multivariate Gaussian distributions.

1. (Gaussian Mixture Models) Say you have an i.i.d dataset, $y_1, ..., y_n$ where y_i is the weight of a single fish in grams. You know the dataset was created by sampling from K lakes. Unfortunately the person was rather careless when collecting the data and forgot to log which lake each fish was sampled from or how many fish they sampled from each lake.

   (a) Let $X_i \in \{1, \ldots, K\}$ be the latent variable that represents which lake the i^{th} fish belongs to. Assuming that the weight of fish within lake j is Gaussian distributed with mean μ_j and variance σ_j^2 write down $\mathbb{P}(y_i | X_i = j)$.

   Solution: This is just the formula for a Gaussian distribution.

   $$\mathbb{P}(y_i | X_i = j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{(y_i - \mu_j)^2}{2\sigma_j^2}\right) = \mathcal{N}(y_i; \mu_j, \sigma_j)$$

   (b) Write down the likelihood $\mathbb{P}(y_1, \ldots, y_n; \theta)$ where $\theta = \{\mu_1, \ldots, \mu_K, \sigma_1, \ldots, \sigma_K, \pi_1, \ldots, \pi_K\}$ and π_j is the probability that any fish belongs to the j^{th} distribution.

   Solution: We have that

   $$\begin{aligned}\mathbb{P}(y_1, \ldots, y_n; \theta) &= \prod_{i=1}^{n} \sum_{j=1}^{K} \pi_j \mathbb{P}(y_i | X_i = j) \\ &= \prod_{i=1}^{n} \sum_{j=1}^{K} \pi_j \mathcal{N}(y_i; \mu_j, \sigma_j)\end{aligned}$$

   (c) Try to compute the derivative of the log likelihood with respect to μ_j, σ_j^2 and π_j can you find the maximum likelihood estimate by finding a closed form solution?

Solution: Computing the derivatives of the log likelihood gives us that

$$\frac{d}{d\mu_j}\log\mathbb{P}(y_1,\ldots,y_n|\theta)=\sum_{i=1}^{n}\frac{\pi_j\mathcal{N}(y_i;\mu_j,\sigma_j)}{\sum_{k=1}^{K}\pi_k\mathcal{N}(y_i;\mu_k,\sigma_k)}\frac{y_i-\mu_j}{\sigma_j^2}$$
$$\frac{d}{d\sigma_j^2}\log\mathbb{P}(y_1,\ldots,y_n|\theta)=\sum_{i=1}^{n}\frac{\pi_j\mathcal{N}(y_i;\mu_j,\sigma_j)}{\sum_{k=1}^{K}\pi_k\mathcal{N}(y_i;\mu_k,\sigma_k)}\frac{(y_i-\mu_j)^2-\sigma_j^2}{2\sigma_j^3}$$
$$\frac{d}{d\pi_j}\log\mathbb{P}(y_1,\ldots,y_n|\theta)=\sum_{i=1}^{n}\frac{\mathcal{N}(y_i;\mu_j,\sigma_j)}{\sum_{k=1}^{K}\pi_k\mathcal{N}(y_i;\mu_k,\sigma_k)}$$

Were we to set these quantities to 0 and try to solve for μ_j and σ_j we would find that there exists no closed form solution. Furthermore we need to constrain $\sum_{j=1}^{K}\pi_j=1$ so we can't just set the derivative to 0 in the third case. We need to use Lagrange multipliers instead. Don't worry if you haven't heard of them but it means that we need to set $\left[\frac{d}{d\pi_j}\log\mathbb{P}(y_1,\ldots,y_n|\theta)\right]-n$ to 0 here. In this case too there does not exist a closed form solution.

(d) Compute $\mathbb{P}(X_i=j|y_i)$ using Bayes' Theorem.

Solution: Using Bayes' Theorem gives us

$$\mathbb{P}(X_i=j|y_i)=\frac{\mathbb{P}(y_i|X_i=j)\mathbb{P}(X_i=j)}{\mathbb{P}(y_i)}=\frac{\pi_j\mathcal{N}(y_i;\mu_j,\sigma_j)}{\sum_{k=1}^{K}\pi_k\mathcal{N}(y_i;\mu_k,\sigma_k)}$$

(e) Now setup an expectation maximization procedure where you alternate between computing $\hat{\mathbb{P}}(X_i=j|y_i)$ and the parameters $\hat\mu_j,\hat\sigma_j,\hat\pi_j$.

Solution:

1. Expectation step: using our prior estimates $\hat\mu_j,\hat\sigma_j,\hat\pi_j$ we can approximate the result from part d as

$$\hat{\mathbb{P}}(X_i=j|y_i)=\frac{\hat\pi_j\mathcal{N}(y_i;\hat\mu_j,\hat\sigma_j)}{\sum_{k=1}^{K}\hat\pi_k\mathcal{N}(y_i;\hat\mu_k,\hat\sigma_k)}.$$

2. Maximization step: note that the true value of $\mathbb{P}(X_i=j|y_i)$ as written in part d occurs in all derivative from part c. We can substitute our estimate

$\hat{\mathbb{P}}(X_i = j|y_i)$ in those derivatives, set them to 0 and rearrange to get

$$\hat{\mu}_j = \frac{\sum_{i=1}^n \hat{\mathbb{P}}(X_i = j|y_i) y_i}{\hat{N}_j}$$
$$\hat{\sigma}_j^2 = \frac{\sum_{i=1}^n \hat{\mathbb{P}}(X_i = j|y_i)(y_i - \hat{\mu}_j)^2}{\hat{N}_j}$$
$$\hat{\pi}_j = \frac{N_j}{n},$$

where $N_j = \sum_{i=1}^n \hat{\mathbb{P}}(X_i = j|y_i)$ can be interpreted as our estimate of the number of fish associated to a specific distribution.

2. (Multivariate Gaussian) Recall that if a random vector $x \in \mathbb{R}^d$ is Gaussian distributed with mean vector μ and covariance matrix Σ then we can write

$$\mathbb{P}(x; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(-\frac{(x-\mu)^\top \Sigma^{-1}(x-\mu)}{2}\right).$$

Show that if we have independent samples $x_1, x_2, \dots, x_n$ where $x_i \sim \mathcal{N}(\mu_i, \sigma)$ then the aggregate random vector of all the samples $x = (x_1, x_2, \dots, x_n)^\top$ has distribution $\mathcal{N}(\mu, \sigma^2 I)$ where $\mu = (\mu_1, \mu_2, \dots \mu_n)^\top$.

Solution: Using the independence of the samples and the fact that they are Gaussian distributed we have

$$\begin{aligned}
\mathcal{P}(x; \mu, \sigma) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma^2}\right) \\
&= \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp\left(-\frac{\sum_{i=1}^d (x_i - \mu_i)^2}{2\sigma^2}\right) \\
&= \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp\left(-\frac{\sum_{i=1}^d (x - \mu)^\top (x - \mu)}{2\sigma^2}\right) \\
&= \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp\left(-\frac{\sum_{i=1}^d (x - \mu)^\top \sigma^{-2} I (x - \mu)}{2}\right) \\
&= \frac{1}{\sqrt{(2\pi)^d \det(\sigma^2 I)}} \exp\left(-\frac{\sum_{i=1}^d (x - \mu)^\top \sigma^{-2} I (x - \mu)}{2}\right)
\end{aligned}$$