Longxiang Wang
12710172
wlongxiang1119@gmail.com

1 Mixture of Experts

a) Write down the data likelihood of $p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{\Phi})$ and its log form

Since we have N data points in the training dataset (i.i.d.), data likelihood can be written out as product of each data point:

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{\Phi}) = \prod_{n=1}^{N} p(\boldsymbol{y_n}|\boldsymbol{x_n}, \boldsymbol{\Theta}, \boldsymbol{\Phi}) \tag{1}$$

According to product rule and sum rule of probability, Equation 1 can be written as:

$$p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{\Phi}) = \prod_{n=1}^{N} p(\boldsymbol{y_n}|\boldsymbol{x_n}, \boldsymbol{\Theta}, \boldsymbol{\Phi}) = \prod_{n=1}^{N}\sum_{k=1}^{K} p(\boldsymbol{y_n}|\boldsymbol{x_n}, \boldsymbol{\theta_k}, \boldsymbol{z_n} = \boldsymbol{k})p(\boldsymbol{z_n} = \boldsymbol{k}|\boldsymbol{x_n}, \boldsymbol{\Phi}) \tag{2}$$

And the log-likelihood is given by

$$\begin{aligned} logp(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{\Phi}) &= log\Big(\prod_{n=1}^{N}\sum_{k=1}^{K} p(\boldsymbol{y_n}|\boldsymbol{x_n}, \boldsymbol{\theta_k}, \boldsymbol{z_n} = \boldsymbol{k})p(\boldsymbol{z_n} = \boldsymbol{k}|\boldsymbol{x_n}, \boldsymbol{\Phi})\Big) \\ &= \sum_{n=1}^{N} log\Big(\sum_{k=1}^{K} p(\boldsymbol{y_n}|\boldsymbol{x_n}, \boldsymbol{\theta_k}, \boldsymbol{z_n} = \boldsymbol{k})p(\boldsymbol{z_n} = \boldsymbol{k}|\boldsymbol{x_n}, \boldsymbol{\Phi})\Big) \end{aligned} \tag{3}$$

b) Write down the posterior probability r_{ni} of expert i producing the label y for datapoint n

Posterior probability r_{ni} of expert i producing the label y for datapoint n, is also referred as the responsibility of expert i for datapoint n, thus:

$$r_{ni} = p(z_n = i|y_n) = \frac{p(y_n|z_n = i)p(z_n = i)}{p(y_n)} = \frac{p(y_n|z_n = i)p(z_n = i)}{\sum_{j=1}^{K} p(y_n|z_n = j)p(z_n = j)} \tag{4}$$

Now we just need to plug the conditional variables into Equation 4, therefore:

$$r_{ni} = \frac{p(y_n|x_n, \theta_i, z_n = i)p(z_n = i|x_n, \Phi)}{\sum_{j=1}^{K} p(y_n|x_n, \theta_j, z_n = j)p(z_n = j|x_n, \Phi)} \tag{5}$$

c) Take the derivative of the log-likelihood w.r.t. the parameters of each expert θ_i and the parameters of the routing mechanism for each expert ϕ_i

We first build some preliminaries:

$$\frac{\partial f(x)}{\partial x} = f(x)\frac{\partial log f(x)}{\partial x} \rightarrow \frac{\partial log f(x)}{\partial x} = \frac{1}{f(x)}\frac{\partial f(x)}{\partial x} \tag{6}$$

We let

$$f(\boldsymbol{\theta}) = \sum_{k=1}^{K} p(\boldsymbol{y_n}|\boldsymbol{x_n}, \boldsymbol{\theta_k}, \boldsymbol{z_n} = \boldsymbol{k})p(\boldsymbol{z_n} = \boldsymbol{k}|\boldsymbol{x_n}, \boldsymbol{\Phi}) \tag{7}$$

$$f(\boldsymbol{\phi}) = \sum_{k=1}^{K} p(\boldsymbol{y_n}|\boldsymbol{x_n}, \boldsymbol{\theta_k}, \boldsymbol{z_n} = \boldsymbol{k})p(\boldsymbol{z_n} = \boldsymbol{k}|\boldsymbol{x_n}, \boldsymbol{\Phi}) \tag{8}$$

$$\mathcal{Z} = logp(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\Theta}, \boldsymbol{\Phi}) = \sum_{n=1}^{N} log f(\boldsymbol{\theta}) \tag{9}$$

Based on the above preliminaries, we can solve for the derivative of the log-likelihood w.r.t. the parameters of each expert $\boldsymbol{\theta_i}$ as below:

$$
\begin{aligned}
\frac{\partial \mathcal{Z}}{\partial \theta_i} &= \sum_{n=1}^{N} \frac{1}{f(\boldsymbol{\theta})} \frac{\partial f(\boldsymbol{\theta})}{\partial \theta_i} \\
&= \sum_{n=1}^{N} \frac{1}{\sum_{k=1}^{K} p(\boldsymbol{y_n}|\boldsymbol{x_n}, \boldsymbol{\theta_k}, z_{\boldsymbol{n}} = \boldsymbol{k}) p(z_{\boldsymbol{n}} = \boldsymbol{k}|\boldsymbol{x_n}, \boldsymbol{\Phi})} \frac{p(z_{\boldsymbol{n}} = \boldsymbol{i}|\boldsymbol{x_n}, \boldsymbol{\Phi}) \partial p(\boldsymbol{y_n}|\boldsymbol{x_n}, \boldsymbol{\theta_i}, z_{\boldsymbol{n}} = \boldsymbol{i})}{\partial \theta_i} \\
&= \sum_{n=1}^{N} \frac{p(z_{\boldsymbol{n}} = \boldsymbol{i}|\boldsymbol{x_n}, \boldsymbol{\Phi}) p(\boldsymbol{y_n}|\boldsymbol{x_n}, \boldsymbol{\theta_i}, z_{\boldsymbol{n}} = \boldsymbol{i})}{\sum_{k=1}^{K} p(\boldsymbol{y_n}|\boldsymbol{x_n}, \boldsymbol{\theta_k}, z_{\boldsymbol{n}} = \boldsymbol{k}) p(z_{\boldsymbol{n}} = \boldsymbol{k}|\boldsymbol{x_n}, \boldsymbol{\Phi})} \frac{\partial logp(\boldsymbol{y_n}|\boldsymbol{x_n}, \boldsymbol{\theta_i}, z_{\boldsymbol{n}} = \boldsymbol{i})}{\partial \theta_i} \\
&= \sum_{n=1}^{N} r_{ni} \frac{\partial logp(\boldsymbol{y_n}|\boldsymbol{x_n}, \boldsymbol{\theta_i}, z_{\boldsymbol{n}} = \boldsymbol{i})}{\partial \theta_i} \qquad (10)
\end{aligned}
$$

Based on the above preliminaries, we can solve for the derivative of the log-likelihood w.r.t. the parameters of the routing mechanism for each expert $\boldsymbol{\phi_i}$ as below:

$$
\begin{aligned}
\frac{\partial \mathcal{Z}}{\partial \phi_i} &= \sum_{n=1}^{N} \frac{1}{f(\boldsymbol{\phi})} \frac{\partial f(\boldsymbol{\phi})}{\partial \phi_i} \\
&= \sum_{n=1}^{N} \frac{1}{\sum_{k=1}^{K} p(\boldsymbol{y_n}|\boldsymbol{x_n}, \boldsymbol{\theta_k}, z_{\boldsymbol{n}} = \boldsymbol{k}) p(z_{\boldsymbol{n}} = \boldsymbol{k}|\boldsymbol{x_n}, \boldsymbol{\Phi})} \frac{\sum_{k=1}^{K} p(\boldsymbol{y_n}|\boldsymbol{x_n}, \boldsymbol{\theta_i}, z_{\boldsymbol{n}} = \boldsymbol{k}) \partial p(z_{\boldsymbol{n}} = \boldsymbol{k}|\boldsymbol{x_n}, \boldsymbol{\Phi})}{\partial \phi_i} \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{p(z_{\boldsymbol{n}} = \boldsymbol{k}|\boldsymbol{x_n}, \boldsymbol{\Phi}) p(\boldsymbol{y_n}|\boldsymbol{x_n}, \boldsymbol{\theta_i}, z_{\boldsymbol{n}} = \boldsymbol{k})}{\sum_{k=1}^{K} p(\boldsymbol{y_n}|\boldsymbol{x_n}, \boldsymbol{\theta_k}, z_{\boldsymbol{n}} = \boldsymbol{k}) p(z_{\boldsymbol{n}} = \boldsymbol{k}|\boldsymbol{x_n}, \boldsymbol{\Phi})} \frac{\partial logp(z_{\boldsymbol{n}} = \boldsymbol{k}|\boldsymbol{x_n}, \boldsymbol{\Phi})}{\partial \phi_i} \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \frac{\partial logp(z_{\boldsymbol{n}} = \boldsymbol{k}|\boldsymbol{x_n}, \boldsymbol{\Phi})}{\partial \phi_i} \qquad (11)
\end{aligned}
$$

d) Replace the expressions for each of the respective probability distributions and compute the final derivatives for θ_i, ϕ_i

We first look at the final derivatives for θ_i:

$$
logp(\boldsymbol{y_n}|\boldsymbol{x_n}, \boldsymbol{\theta_i}, z_{\boldsymbol{n}} = \boldsymbol{i}) = logexp(\boldsymbol{y_n}|\lambda = exp(\boldsymbol{\theta_i^T x_n})) = log\lambda exp(-\lambda y) = log\lambda - \lambda \boldsymbol{y_n} = \boldsymbol{\theta_i^T x_n} - \boldsymbol{y_n} exp(\boldsymbol{\theta_i^T x_n}) \qquad (12)
$$

$$
\frac{\partial logp(\boldsymbol{y_n}|\boldsymbol{x_n}, \boldsymbol{\theta_i}, z_{\boldsymbol{n}} = \boldsymbol{i})}{\partial \theta_i} = \boldsymbol{x_n^T} - \boldsymbol{y_n} exp(\boldsymbol{\theta_i^T x_n}) \boldsymbol{x_n^T} \qquad (13)
$$

$$
\frac{\partial \mathcal{Z}}{\partial \theta_i} = \sum_{n=1}^{N} r_{ni}(1 - \boldsymbol{y_n} exp(\boldsymbol{\theta_i^T x_n})) \boldsymbol{x_n^T} \qquad (14)
$$

We then look at the final derivatives for ϕ_i:

$$
logp(z_n = k|\boldsymbol{x_n}, \boldsymbol{\Phi}) = log\pi_{nk} = log \frac{exp(\boldsymbol{\phi_k^T x_n})}{\sum_{j=1}^{j=K} exp(\boldsymbol{\phi_j^T x_n})} = \boldsymbol{\phi_k^T x_n} - log \sum_{j=1}^{j=K} exp(\boldsymbol{\phi_j^T x_n}) \qquad (15)
$$

$$
\frac{\partial logp(z_n = k|\boldsymbol{x_n}, \boldsymbol{\Phi})}{\partial \phi_i} = \boldsymbol{x_n^T} - \frac{exp(\boldsymbol{\phi_i^T x_n}) \boldsymbol{x_n^T}}{\sum_{j=1}^{j=K} exp(\boldsymbol{\phi_j^T x_n})} = (\mathcal{I}[k = i] - \pi_{ni}) \boldsymbol{x_n^T} \qquad (16)
$$

$$
\frac{\partial \mathcal{Z}}{\partial \phi_i} = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk}(\mathcal{I}[k = i] - \pi_{ni}) \boldsymbol{x_n^T} = \sum_{n=1}^{N} (r_{nk} - \pi_{ni}) \boldsymbol{x_n^T} \qquad (17)
$$

2 Quadratic Discriminant Analysis

a) Write down the joint probability $p(\boldsymbol{x_n}, C_k)$ for a single datapoint using a Gaussian class-conditional density and prior

According to the product rule and Gaussian distribution for class observations:

$$p(\boldsymbol{x_n}, C_k) = p(\boldsymbol{x_n}|C_k)p(C_k) = \mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})\pi_k \tag{18}$$

b) Write down the likelihood function and its log form

Using the result obtained in a) for a single datapoint and i.i.d. asumption:

$$p(\boldsymbol{t_n}, \boldsymbol{x_n}|\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K) = \prod_{n=1}^{N}\prod_{k=1}^{K}(\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})\pi_k)^{t_{nk}} \tag{19}$$

Using the property of log function $log(xy) = logx + logy$, we have:

$$\begin{aligned} logp(\boldsymbol{t_n}, \boldsymbol{x_n}|\pi_1, \ldots, \pi_K, \mu_1, \ldots, \mu_K, \Sigma_1, \ldots, \Sigma_K) &= \sum_{n=1}^{N}\sum_{k=1}^{K} log\Big((\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})\pi_k)^{t_{nk}}\Big) && (20)\\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} log(\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})\pi_k) && (21)\\ &= \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\Big(log\pi_k + log\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})\Big) && (22) \end{aligned}$$

c) Write down the Lagrangian function using the log likelihood, a Lagrange multiplier and the equality constraint $\sum_{k=1}^{K}\pi_k = 1$

The equality constrain can be written as:

$$\sum_{k=1}^{K}\pi_k - 1 = 0 \tag{23}$$

Therefore the Lagrange function can be written as:

$$L(\pi_k, \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}, \lambda) = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\Big(log\pi_k + log\mathcal{N}(\boldsymbol{x_n}|\boldsymbol{\mu_k}, \boldsymbol{\Sigma_k})\Big) + \lambda(\sum_{k=1}^{K}\pi_k - 1) \tag{24}$$

where λ is the Lagrange multiplier.

d) Write down π_{ML}

We only have to look at the terms related to π_k in the Lagrange function, and take the derivative w.r.t. to it to zero:

$$\frac{\partial L}{\partial \pi_k} = \frac{\partial\Big(\sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}log\pi_k + +\lambda\sum_{k=1}^{K}\pi_k\Big)}{\partial \pi_k} = \lambda + \sum_{n=1}^{N}\frac{t_{nk}}{\pi_k} = \lambda + \frac{N_k}{\pi_k} = 0 \rightarrow \pi_k = \frac{-N_k}{\lambda} \tag{25}$$

where N_k represents number of samples in class k.
Now if we plug in π_k into Equation 23, we can solve for λ:

$$\sum_{k=1}^{K}\frac{-N_k}{\lambda} - 1 = 0 \rightarrow \lambda = -N \tag{26}$$

where N represents total number of samples in input space.
Therefore, we plug in λ back, we can obtain the final form of π_k:

$$\pi_{k,ML} = \frac{-N_k}{\lambda} = \frac{N_k}{N} \tag{27}$$

e) Write down μ_{ML}

We only have to look at the terms related to μ_k in the Lagrange function, and take the derivative w.r.t. to it to zero:

$$\frac{\partial L}{\partial \mu_k} = \frac{\partial}{\partial \mu_k} \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} log \mathcal{N}(\boldsymbol{x_n} | \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) \tag{28}$$

As we only care about the quadratic form of Gaussian when setting the derivative to zero, we have:

$$\begin{aligned} \frac{\partial L}{\partial \mu_k} &= \frac{\partial}{\partial \mu_k} \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} log \mathcal{N}(\boldsymbol{x_n} | \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) \\ &= \frac{\partial}{\partial \mu_k} \{ \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{-t_{nk}}{2} (\boldsymbol{x_n} - \boldsymbol{\mu_k})^{\boldsymbol{T}} \boldsymbol{\Sigma_k^{-1}} (\boldsymbol{x_n} - \boldsymbol{\mu_k}) + const \} \\ &= \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} (\boldsymbol{x_n} - \boldsymbol{\mu_k})^{\boldsymbol{T}} \boldsymbol{\Sigma_k^{-1}} = 0 \rightarrow \sum_{n=1}^{N} (t_{nk} \boldsymbol{x_n^T} - t_{nk} \boldsymbol{\mu_k^T}) = 0 \rightarrow \sum_{n=1}^{N} t_{nk} \boldsymbol{x_n} = \sum_{n=1}^{N} t_{nk} \boldsymbol{\mu_k} = N_k \boldsymbol{\mu_k} \end{aligned} \tag{29}$$

Therefore, we have the final form:

$$\boldsymbol{\mu_{k,ML}} = \frac{1}{N_k} \sum_{n=1}^{N} t_{nk} \boldsymbol{x_n} \tag{30}$$

f) Write down Σ_{ML}

According to Equation 2.122 in Bishop book, we know the MLE solution Σ_{ML} for a multivariate Gaussian distribution is given by

$$\Sigma_{ML} = \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{x_n} - \boldsymbol{\mu_{ML}})(\boldsymbol{x_n} - \boldsymbol{\mu_{ML}})^{\boldsymbol{T}} \tag{31}$$

In the case of class conditional Gaussian distribution, we only need to replace the mean and sample counts by the mean and sample counts of a class, therefore, we have:

$$\begin{aligned} \frac{\partial L}{\partial \Sigma_k} &= \frac{\partial}{\partial \Sigma_k} \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} log \mathcal{N}(\boldsymbol{x_n} | \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}) \\ &= \frac{\partial}{\partial \Sigma_k} \{ \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{-t_{nk}}{2} (\boldsymbol{x_n} - \boldsymbol{\mu_k})^{\boldsymbol{T}} \boldsymbol{\Sigma_k^{-1}} (\boldsymbol{x_n} - \boldsymbol{\mu_k}) + const \} = 0 \\ &\rightarrow \boldsymbol{\Sigma_{k,ML}} = \frac{N_k}{N} \sum_{n=1}^{N} (\boldsymbol{x_n} - \boldsymbol{\mu_{k,ML}})(\boldsymbol{x_n} - \boldsymbol{\mu_{k,ML}})^{\boldsymbol{T}} \end{aligned} \tag{32}$$

g) Write a single-sentence interpretation for each of the solutions: π_{ML}, μ_{ML} and Σ_{ML}

- for π_{ML}, it can be interpreted as the frequency of class k in the entire input data set.
- for μ_{ML}, it is the mean of all the input vector assigned to class K.
- for Σ_{ML}, as we can see from $\boldsymbol{\Sigma_{k,ML}}$, the final covariance is a weighted average of the covariance for each class k.

3 Principal Component Analysis

a) Projection z_{ni} of given data point $\boldsymbol{x_n}$ onto eigen vector $\boldsymbol{u_i}$

It is easy to see that:

$$z_{ni} = \boldsymbol{u_i^T} \boldsymbol{x_n} \tag{33}$$

where $\boldsymbol{u_i^T} = (u_1, u_2, \ldots, u_D)$ and $\boldsymbol{x_n^T} = (x_1, x_2, \ldots, x_D)$.

b) Empirical mean of the projection z_i across all points $\boldsymbol{x_n}$

Based on the assumption that all input dimensions have zero mean, we know that $\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x_n} = 0$, therefore, we have

$$\mathbb{E}(z_i) = \frac{1}{N}\sum_{n=1}^{N} z_{ni} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{u_i^T}\boldsymbol{x_n} = \frac{\boldsymbol{u_i^T}}{N}\sum_{n=1}^{N}\boldsymbol{x_n} = 0 \tag{34}$$

c) Empirical variance of the projection z_i across all points $\boldsymbol{x_n}$

According to the definition of variance $\mathbb{VAR}(X) = \mathbb{E}[(X-\mu)^2]$, we have

$$\mathbb{VAR}(z_i) = \frac{1}{N}\sum_{n=1}^{N}(z_{ni} - \mathbb{E}(z_i))^2 = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{u_i^T}\boldsymbol{x_n})^2 = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{u_i^T}\boldsymbol{x_n}\boldsymbol{x_n^T}\boldsymbol{u_i} = \boldsymbol{u_i^T}\left(\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x_n}\boldsymbol{x_n^T}\right)\boldsymbol{u_i} = \boldsymbol{u_i^T}\boldsymbol{S}\boldsymbol{u_i} \tag{35}$$

where $\boldsymbol{S} = \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{x_n}\boldsymbol{x_n^T}$

e) Replace covariance matriax S with its eigen decomposition and simply the variance above

We just need to plug in the eigen decomposition:

$$\mathbb{VAR}(z_i) = \boldsymbol{u_i^T}\boldsymbol{S}\boldsymbol{u_i} = \boldsymbol{u_i^T}\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U^T}\boldsymbol{u_i} = (\boldsymbol{U^T}\boldsymbol{u_i})^{\boldsymbol{T}}\boldsymbol{\Lambda}(\boldsymbol{U^T}\boldsymbol{u_i}) = \boldsymbol{e_i^T}\boldsymbol{\Lambda}\boldsymbol{e_i} = \lambda_i \tag{36}$$

where $\boldsymbol{e_i^T} = (0,\ldots,\boldsymbol{u_i^T}\boldsymbol{u_i},\ldots,0) = (0,\ldots,1,\ldots,0)$

d) Select proper K (K < D) such that 99% of variance is captured

We can do the following:

- Do eigen decomposition of input data $\boldsymbol{X}$ to get eigen values of $\boldsymbol{\Lambda} = [\lambda_1, \lambda_2, \ldots, \lambda_D]$
- Sort $\boldsymbol{\Lambda}$ by descending order
- Select eigen values from sorted $\boldsymbol{\Lambda}$ until $\sum_{i=1}^{K} >= 0.99$
- K has been found