SHRUTI RAO
2636454
s2.rao@student.vu.nl

Mixture of Experts

1

The likelihood for the entire dataset is given by:

$$\begin{aligned} p(y_n|\boldsymbol{x}_n, \boldsymbol{\Theta}, \boldsymbol{\Phi}) &= \prod_{n=1}^{N} \sum_{k=1}^{k} p(y_n|\boldsymbol{x}_n, \boldsymbol{\theta}_k = \boldsymbol{\Theta}\boldsymbol{z}_n) p(z_n = k|\boldsymbol{x}_n, \boldsymbol{\Phi}) \\ &= \prod_{n=1}^{N} \sum_{k=1}^{k} \pi_{nk} Exponential(y_n|\lambda = exp(\boldsymbol{\theta}_k^T, \boldsymbol{x}_n)) \end{aligned}$$

Since $Exponential(y_n|\lambda = \lambda exp(-\lambda y))$:

$$= \prod_{n=1}^{N} \sum_{k=1}^{k} \pi_{nk} exp(\boldsymbol{\theta}_k^T \boldsymbol{x}_n) exp(-exp(\boldsymbol{\theta}_k^T \boldsymbol{x}_n) y_n)$$

Taking the log on both sides:

$$\begin{aligned} \ln p(y_n|\boldsymbol{x}_n, \boldsymbol{\Theta}, \boldsymbol{\Phi}) &= \ln \prod_{n=1}^{N} \sum_{k=1}^{k} \pi_{nk} exp(\boldsymbol{\theta}_k^T \boldsymbol{x}_n) exp(-exp(\boldsymbol{\theta}_k^T \boldsymbol{x}_n) y_n) \\ &= \sum_{n=1}^{N} \ln \sum_{k=1}^{k} \pi_{nk} e^{(\boldsymbol{\theta}_k^T \boldsymbol{x}_n)} exp(-e^{(\boldsymbol{\theta}_k^T \boldsymbol{x}_n) y_n}) \end{aligned}$$

2

The formula for posterior in general is given by:

$$P(w|\mathcal{D}) = \frac{P(\mathcal{D}|w)P(w)}{P(\mathcal{D})}$$

The posterior probability r_{ni} of expert i producing label y for datapoint n is given by:

$$\begin{aligned} r_{ni} &= \frac{p(z_n = i|\boldsymbol{x}_n, \boldsymbol{\Phi}) p(y_n|\boldsymbol{x}_n, \boldsymbol{\theta}_k = \boldsymbol{\Theta}\boldsymbol{z}_n)}{p(y_n|\boldsymbol{x}_n, \boldsymbol{\Theta}, \boldsymbol{\Phi})} \\ &= \frac{\pi_{ni}.exp(-exp(\boldsymbol{\theta}_k^T \boldsymbol{x}_n) y_n)}{\sum_{i=1} \pi_{ni}.exp(-exp(\boldsymbol{\theta}_k^T \boldsymbol{x}_n) y_n)} \end{aligned}$$

3

Rewriting the log-likelihood equation in terms of its probability functions we get:

$$\begin{aligned} \ln p(y|\boldsymbol{x}, \boldsymbol{\Theta}, \boldsymbol{\Phi}) &= \sum_{n=1}^{N} \ln \sum_{k=1}^{k} \pi_{nk} e^{(\boldsymbol{\theta}_k^T \boldsymbol{x}_n)} exp(-e^{(\boldsymbol{\theta}_k^T \boldsymbol{x}_n) y_n}) \\ &= \sum_{n=1}^{N} \ln \sum_{k=1}^{k} p(y_n|\boldsymbol{x}_n, z_n, \boldsymbol{\Theta}) p(z_n = k|\boldsymbol{x}_n, \boldsymbol{\Phi}) \end{aligned}$$

First, we take the derivative with respect to $\boldsymbol{\theta}_i$:

$$\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{\theta}_i} &= \frac{\partial}{\partial \boldsymbol{\theta}_i}\left[\sum_{n=1}^{N} \ln \sum_{k=1}^{k} p(y_n|\boldsymbol{x}_n, z_n, \boldsymbol{\Theta}) p(z_n = k|\boldsymbol{x}_n, \boldsymbol{\Phi})\right] \\
&= \sum_{n=1}^{N}\left[\frac{\frac{\partial}{\partial \boldsymbol{\theta}_i}\sum_{k=1}^{K} p(y_n|\boldsymbol{x}_n, z_n, \boldsymbol{\Theta}) p(z_n = k|\boldsymbol{x}_n, \boldsymbol{\Phi})}{\sum_{k=1}^{K} p(y_n|\boldsymbol{x}_n, z_n, \boldsymbol{\Theta}) p(z_n = k|\boldsymbol{x}_n, \boldsymbol{\Phi})}\right] \\
&= \sum_{n=1}^{N}\left[\frac{p(y_n|\boldsymbol{x}_n, z_n, \boldsymbol{\Theta}) p(z_n = i|\boldsymbol{x}_n, \boldsymbol{\Phi})}{\sum_{k=1}^{K} p(y_n|\boldsymbol{x}_n, z_n, \boldsymbol{\Theta}) p(z_n = k|\boldsymbol{x}_n, \boldsymbol{\Phi})}\right]
\end{aligned}$$

From 1.2, we know that: $r_{ni} = \frac{p(z_n=i|\boldsymbol{x}_n,\boldsymbol{\Phi})p(y_n|\boldsymbol{x}_n,z_n,\boldsymbol{\Theta})}{p(y_n|\boldsymbol{x}_n,\boldsymbol{\Theta},\boldsymbol{\Phi})}$. Moreover, using the fact that: $\frac{\partial f(x)}{\partial x} = f(x)\frac{\partial \ln f(x)}{\partial x}$, we can re-write the equation as:

$$\begin{aligned}
&= \sum_{n=1}^{N}\left[\frac{p(y_n|\boldsymbol{x}_n, z_n, \boldsymbol{\Theta}) p(z_n = i|\boldsymbol{x}_n, \boldsymbol{\Phi})}{\sum_{k=1}^{K} p(y_n|\boldsymbol{x}_n, z_n, \boldsymbol{\Theta}) p(z_n = k|\boldsymbol{x}_n, \boldsymbol{\Phi})}\right] \\
&= \sum_{n=1}^{N} r_{ni} \frac{\partial}{\partial \boldsymbol{\theta}_i}\left[\ln p(y_n|\boldsymbol{x}_n, z_n, \boldsymbol{\Theta})\right]
\end{aligned}$$

Next, we take the derivative with respect to $\boldsymbol{\phi}_i$:

$$\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{\phi_i}} &= \frac{\partial}{\partial \boldsymbol{\phi}_i}\left[\sum_{n=1}^{N} \ln \sum_{k=1}^{k} p(y_n|\boldsymbol{x}_n, z_n, \boldsymbol{\Theta}) p(z_n = k|\boldsymbol{x}_n, \boldsymbol{\Phi})\right] \\
&= \sum_{n=1}^{N} \frac{\partial}{\partial \boldsymbol{\phi_i}} \ln\left[\sum_{k=1}^{K} p(y_n|\boldsymbol{x}_n, z_n, \boldsymbol{\Theta}) p(z_n = k|\boldsymbol{x}_n, \boldsymbol{\Phi})\right] \\
&= \sum_{n=1}^{N}\left[\frac{\sum_{k=1}^{K} p(y_n|\boldsymbol{x}_n, z_n, \boldsymbol{\Theta}) \frac{\partial}{\partial \boldsymbol{\phi}_i} p(z_n = k|\boldsymbol{x}_n, \boldsymbol{\Phi})}{\sum_{k=1}^{K} p(y_n|\boldsymbol{x}_n, z_n, \boldsymbol{\Theta}) p(z_n = k|\boldsymbol{x}_n, \boldsymbol{\Phi})}\right] \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K}\left[\frac{p(y_n|\boldsymbol{x}_n, z_n, \boldsymbol{\Theta}) p(z_n = i|\boldsymbol{x}_n, \boldsymbol{\Phi})}{\sum_{k=1}^{K} p(y_n|\boldsymbol{x}_n, z_n, \boldsymbol{\Theta}) p(z_n = k|\boldsymbol{x}_n, \boldsymbol{\Phi})}\right] \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{ni} \frac{\partial}{\partial \boldsymbol{\phi}_i}\left[\ln p(z_n = i|\boldsymbol{x}_n, \boldsymbol{\Phi})\right]
\end{aligned}$$

4

For the derivative with respect to theta $\frac{\partial L}{\partial \boldsymbol{\theta_i}}$:

$$\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{\theta}_i} &= \sum_{n=1}^{N} r_{ni} \frac{\partial}{\partial \boldsymbol{\theta}_i}\left[\ln p(y_n|\boldsymbol{x}_n, z_n, \boldsymbol{\Theta})\right] \\
&= \sum_{n=1}^{N} r_{ni} \frac{\partial}{\partial \boldsymbol{\theta}_i}\left[\sum_{k=1}^{K} \ln e^{(\boldsymbol{\theta}_k^T \boldsymbol{x}_n)} exp(-e^{(\boldsymbol{\theta}_k^T \boldsymbol{x}_n)y_n})\right] \\
&= \sum_{n=1}^{N} r_{ni} \frac{\partial}{\partial \boldsymbol{\theta}_i}\left[\sum_{k=1}^{K} \boldsymbol{\theta}_k^T \boldsymbol{x}_n - e^{(\boldsymbol{\theta}_k^T \boldsymbol{x}_n)} y_n\right] \\
&= \sum_{n=1}^{N} r_{ni}\left[\frac{\partial}{\partial \boldsymbol{\theta}_i}\sum_{k=1}^{K} \boldsymbol{\theta}_k^T \boldsymbol{x}_n\right] - \left[\frac{\partial}{\partial \boldsymbol{\theta}_i}\sum_{k=1}^{K} e^{(\boldsymbol{\theta}_k^T \boldsymbol{x}_n)} y_n\right] \\
&= \sum_{n=1}^{N} r_{ni}\left[\boldsymbol{x}_n^T - \boldsymbol{x}_n^T e^{(\boldsymbol{\theta}_k^T \boldsymbol{x}_n)} y_n\right] \\
&= \sum_{n=1}^{N} r_{ni} \boldsymbol{x}_n^T\left[1 - e^{(\boldsymbol{\theta}_k^T \boldsymbol{x}_n)} y_n\right]
\end{aligned}$$

For the derivative with respect to $\boldsymbol{phi}_i$:

$$
\begin{aligned}
\frac{\partial L}{\partial \boldsymbol{\phi}_i} &= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{ni}\frac{\partial}{\partial \boldsymbol{\phi}_i}\left[\ln p(z_n = i|\boldsymbol{x}_n, \boldsymbol{\Phi})\right] \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{ni}\frac{\partial}{\partial \boldsymbol{\phi}_i}\left[\ln \frac{e^{\boldsymbol{\Phi_i^T x_n}}}{\sum_j e^{\boldsymbol{\Phi_i^T x_n}}}\right] \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{ni}\frac{\partial}{\partial \boldsymbol{\phi}_i}\left[\boldsymbol{\Phi_i^T x_n} - \ln \sum_j e^{\boldsymbol{\Phi_i^T x_n}}\right] \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{ni}\frac{\partial}{\partial \boldsymbol{\phi}_i}\left[\boldsymbol{\Phi_i^T x_n}\right] - \frac{\partial}{\partial \boldsymbol{\phi}_i}\left[\ln \sum_j e^{\boldsymbol{\Phi_i^T x_n}}\right] \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{ni}\left[x_n^T\right] - \frac{\partial}{\partial \boldsymbol{\phi}_i}\left[\ln \sum_j e^{\boldsymbol{\Phi_i^T x_n}}\right] \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{ni}\left[x_n^T\right] - \left[\frac{e^{\boldsymbol{\Phi_i^T x_n}}}{\sum_j e^{\boldsymbol{\Phi_j^T x_n}}}\right]\left[\boldsymbol{x}_n^T\right] \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{ni}\left[x_n^T\right]\left[1 - \frac{e^{\boldsymbol{\Phi_i^T x_n}}}{\sum_j e^{\boldsymbol{\Phi_j^T x_n}}}\right] \text{ or} \\
&= \sum_{n=1}^{N}\sum_{k=1}^{K} r_{ni}\left[x_n^T\right]\left[1 - p(z_n = i|\boldsymbol{x}_n, \boldsymbol{\Phi})\right]
\end{aligned}
$$

Quadratic Discriminant Analysis

1

The probability density function for a multivariate gaussian is given by:

$$f(x) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|_k}} exp\left\{\frac{-(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)}{2}\right\}$$

The prior $p(\mathcal{C}) = \pi_k$ and thus the joint probability is altogether given by:

$$p(\boldsymbol{x}_n|\boldsymbol{\mathcal{C}_k}) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|_k}} exp\left\{\frac{-(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)}{2}\right\}\pi_k$$

2

The likelihood is given by:

$$p(\boldsymbol{T}, \boldsymbol{X}|\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_K, \boldsymbol{\mu}_1, .., \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_K) = \prod_{n=1}^{N}\prod_{k=1}^{K}\left(\frac{1}{\sqrt{(2\pi)^d|\Sigma|_k}} exp\left\{\frac{-(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)}{2}\right\}\pi_k\right)^{\mathcal{I}(t_n=k)}$$

The log likelihood is then given by:

$$
\begin{aligned}
\ln p(\boldsymbol{T}, \boldsymbol{X}|\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_K, \boldsymbol{\mu}_1, .., \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_K) &= \ln \prod_{n=1}^{N} \prod_{k=1}^{K} \left(\frac{1}{\sqrt{(2\pi)^d |\Sigma|_k}} exp \left\{ \frac{-(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)}{2} \right\} \pi_k \right)^{I(t_n=k)} \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \ln \left(\frac{1}{\sqrt{(2\pi)^d |\Sigma|_k}} exp \left\{ \frac{-(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)}{2} \right\} \pi_k \right) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \ln \left(\frac{1}{\sqrt{(2\pi)^d |\Sigma|_k}} \right) - \left(\frac{(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)}{2} \right) + \ln(\pi_k)
\end{aligned}
$$

3

The equality constraint is given by: $\sum_k^K \pi_k = 1$ The Lagrangian is therefore given by:

$$
\begin{aligned}
\mathcal{L} &= \ln p(\boldsymbol{T}, \boldsymbol{X}|\boldsymbol{\pi}_1, ..., \boldsymbol{\pi}_K, \boldsymbol{\mu}_1, .., \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_K) + \lambda[1 - \sum_k^K \pi_k] \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \ln \left(\frac{1}{\sqrt{(2\pi)^d |\Sigma|_k}} \right) \left(\frac{-(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)}{2} \right) + \ln(\pi_k) + \lambda \left[1 - \sum_k^K \pi_k \right] \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \ln \left(\frac{1}{\sqrt{(2\pi)^d |\Sigma|_k}} \right) \left(\frac{-(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)}{2} \right) + \ln(\pi_k) + \lambda - \lambda \sum_k^K \pi_k
\end{aligned}
$$

4

The partial derivative of the Lagrangian with respect to π_k is:

$$
\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \pi_k} &= \frac{\partial}{\partial \pi_k} \sum_{n=1}^{N} \sum_{k=1}^{K} \ln \left(\frac{1}{\sqrt{(2\pi)^d |\Sigma|_k}} \right) \left(\frac{-(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)}{2} \right) + \ln(\pi_k) + \lambda - \lambda \sum_k^K \pi_k \\
&= \frac{\partial}{\partial \pi_k} \sum_{n=1}^{N} \sum_{k=1}^{K} \ln(\pi_k) - \lambda \sum_k^K \pi_k \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} (\frac{1}{\pi_k}) - \lambda K \\
0 &= \frac{NK}{\pi_k} - \lambda K \\
\frac{NK}{\pi_k} &= \lambda K \\
\pi_k &= \frac{N}{\lambda}
\end{aligned}
$$

5

The partial derivative of the Lagrangian with respect to μ_k is:

$$\frac{\partial \mathcal{L}}{\partial \mu_k} = \frac{\partial}{\partial \mu_k} \sum_{n=1}^{N} \sum_{k=1}^{K} \left(\frac{-(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)}{2} \right)$$
$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{\partial}{\partial \mu_k} \left(\frac{-(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)}{2} \right)$$

Using $\frac{\partial}{\partial x} x^T A x = x^T (A^T + A)$

$$0 = \sum_{n=1}^{N} \sum_{k=1}^{K} -\frac{1}{2} (x-\mu_k)^T (\Sigma_k^{-T} + \Sigma_k^{-1})$$
$$0 = -\sum_{n=1}^{N} \sum_{k=1}^{K} (x^T - \mu_k^T)(\Sigma_k^{-T} + \Sigma_k^{-1})$$
$$0 = -\sum_{n=1}^{N} \sum_{k=1}^{K} x^T (\Sigma_k^{-T} + \Sigma_k^{-1}) - \mu_k^T (\Sigma_k^{-T} + \Sigma_k^{-1})$$
$$= -\sum_{n=1}^{N} \sum_{k=1}^{K} x^T \Sigma_k^{-T} + x^T \Sigma_k^{-1} - \mu_k^T \Sigma_k^{-T} - \mu_k^T \Sigma_k^{-1}$$
$$= -\sum_{n=1}^{N} \sum_{k=1}^{K} x^T \Sigma_k^{-T} - \sum_{n=1}^{N} \sum_{k=1}^{K} x^T \Sigma_k^{-1} + \sum_{n=1}^{N} \sum_{k=1}^{K} \mu_k^T \Sigma_k^{-T} + \sum_{n=1}^{N} \sum_{k=1}^{K} \mu_k^T \Sigma_k^{-1}$$
$$\sum_{n=1}^{N} \sum_{k=1}^{K} \mu_k^T \Sigma_k^{-T} + \sum_{n=1}^{N} \sum_{k=1}^{K} \mu_k^T \Sigma_k^{-1} = \sum_{n=1}^{N} \sum_{k=1}^{K} x^T \Sigma_k^{-T} + \sum_{n=1}^{N} \sum_{k=1}^{K} x^T \Sigma_k^{-1}$$
$$\sum_{n=1}^{N} \sum_{k=1}^{K} \mu_k^T (\Sigma_k^{-T} + \Sigma_k^{-1}) = \sum_{n=1}^{N} \sum_{k=1}^{K} x^T (\Sigma_k^{-T} + \Sigma_k^{-1})$$
$$\mu_k = x$$

6

The partial derivative of the Lagrangian with respect to Σ_k is:

$$\frac{\partial \mathcal{L}}{\partial \Sigma_k} = \frac{\partial}{\partial \Sigma_k} \sum_{n=1}^{N} \sum_{k=1}^{K} \ln \left(\frac{1}{\sqrt{(2\pi)^d |\Sigma|_k}} \right) \left(\frac{-(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)}{2} \right)$$
$$= \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{\partial}{\partial \Sigma_k} \ln \left((2\pi)^d |\Sigma|_k \right)^{-\frac{1}{2}} - \frac{\partial}{\partial \Sigma_k} \left(\frac{(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)}{2} \right)$$
$$= \sum_{n=1}^{N} \sum_{k=1}^{K} -\frac{1}{2} (\ln (2\pi)^d |\Sigma|_k)^{\frac{-1}{2}} - \frac{1}{2} (x-\mu_k)^T (x-\mu_k)$$
$$= -\sum_{n=1}^{N} \sum_{k=1}^{K} \frac{1}{2} (\ln (2\pi)^d |\Sigma|_k)^{\frac{-1}{2}} - \sum_{n=1}^{N} \sum_{k=1}^{K} \frac{1}{2} (x-\mu_k)^T (x-\mu_k)$$
$$\sum_{n=1}^{N} \sum_{k=1}^{K} \frac{1}{2} (\ln (2\pi)^d |\Sigma|_k)^{\frac{-1}{2}} = -\sum_{n=1}^{N} \sum_{k=1}^{K} \frac{1}{2} (x-\mu_k)^T (x-\mu_k)$$

Using equation 2.122 from Bishop 2.3.4:

$$\Sigma_k = \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} (x-\mu_k)(x-\mu_k)^T$$

7

π_k equals to the total N divided by the lagrange multplier. μ_k the mean of the estimate is the observed data point itself. Finally, the Σ_k depends on the symmetric difference between the observed data points and the means.

Principal Component Analysis

1

The projection z_{ni} is given by:

$$z_{ni} = u_i^T x_n$$

2

The empirical mean of the projection z_i across all points is given by:

$$\bar{z}_i = u_i^T x_n$$

3

The empirical variance of the projection z_i in terms of covariance matrix $\mathcal{S}$ is given by:

$$\begin{aligned} Var(z_i) &= \frac{1}{N} \sum_{n=1}^{N} \{u_i^T x_n - u_i^T \bar{z}_i\} \\ &= u_i^T \mathcal{S} u_i \end{aligned}$$

4

$$\begin{aligned} Var(z_i) &= u_i^T \mathcal{S} u_i \\ \text{Since } \mathcal{S} = U \Lambda U^T & \\ Var(z_i) &= u_i^T U \Lambda U^T u_i \\ &= \lambda_i \end{aligned}$$

5

Reducing dimensionality from D to K such that 99% of variance is maintained by picking a K such that the proportion of variance explained is the highest:

$$\frac{\sum_{i=1}^{K} \lambda_i}{Tr(\mathcal{S})} = \frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{D} \lambda_i} > 0.99$$