IFT6269-A2018

Probabilistic Graphical Models

Assignment I

Frédéric Boileau
p0991440

Prof. Simon Lacoste-Julien

5th October 2018

1 Generative Model

$Y \sim \text{Bernouilli}(\pi), \quad X|Y = j \sim \mathcal{N}(\mu_j, \Sigma)$

First let us write the two joint distributions implied by the definition:

$$P(x_i, Y = 1) = P(x_i|Y = 1)P(Y = 1) = \pi\mathcal{N}(x_i|\mu_1, \Sigma)$$

$$P(x_i, Y = 0) = P(x_i|Y = 0)P(Y = 0) = (1 - \pi)\mathcal{N}(x_i|\mu_2, \Sigma)$$

Now taking the product of the observations for the likelihood function we have:

$$L(\theta) = P(X, Y|\mu, \Sigma, \pi) = \prod_{i=1}^{N} \{\pi\mathcal{N}(x_i|\mu_1, \Sigma)\}^{y_i} \{(1 - \pi)\mathcal{N}(x_i|\mu_2, \Sigma)\}^{1-y_i} \tag{1}$$

Where θ is just a surrogate for all the parameters to ease notation.

Taking the log we get the log-likelihood function and keeping only the terms that depend on π:

$$l_\pi(\theta) = \sum_{i=1}^{N} \{y_i \ln \pi + (1 - y_i) \ln(1 - \pi)\} \tag{2}$$

To maximize we simply take the derivative and set to zero:

$$l'_\pi = \sum_{i=1}^{N} \left\{ \frac{y_i}{\pi} - \frac{1 - y_i}{1 - \pi} \right\} = 0 \tag{3}$$

Whence we get that

$$\pi_{MLE} = \frac{1}{N} \sum_{y=1}^{N} = \frac{N_1}{N_1 + N_2} \tag{4}$$

Where $N_1 = |\{i : y_i = 1\}|$ and $N_2 = |\{i : y_i = 1\}|$

Now for μ_1:

$$l_{\mu_1} = -\frac{1}{2} \sum_{i=1}^{N} y_i (x_i - \mu_1)^{\mathsf{T}} \Sigma^{-1} (x_i - \mu_1) + const \tag{5}$$

Taking the derivative and setting to zero :

$$l'_{\mu_1} = -\frac{1}{2} \sum_{i=1}^{N} y_i (x_i - \mu_1)^{\mathsf{T}} (\Lambda + \Lambda^{\mathsf{T}})$$

Where $\Lambda = \Sigma^{-1}$ is the precision matrix which is symmetric as well :

$$0 = \sum_{i=1}^{N} y_i (x_i - \mu_1)^{\mathsf{T}} \Lambda = \sum_{i=1}^{N} y_i (x_i - \mu_1)$$

All in all we have that

$$\mu_{1_{MLE}} = \frac{1}{N_1} \sum_{i=1}^{N} y_i x_i \qquad \mu_{2_{MLE}} = \frac{1}{N_2} \sum_{i=1}^{N} (1 - y_i) x_i \tag{6}$$

Where the latter is obtained following the same steps for μ_2 but replacing y_i by $1 - y_i$. Indeed, in general, if we have some vector of mixture proportions α whose components sum to 1 the MLE for the respective means would be the weighted sum of the observed x_i divided by the number of data points in the corresponding classes.

For the MLE estimate of the covariance matrix we consider the relevant terms of the "sum expansion" of the log-likelihood which gives:

$$l_{\Sigma}(\theta) = \frac{N}{2}\log|\Sigma^{-1}| - \frac{1}{2}\sum_{I} y_i(x_i-\mu_1)^{\mathsf{T}}\Sigma^{-1}(x_i-\mu_1) - \frac{1}{2}\sum_{I}(1-y_i)(x_i-\mu_2)^{\mathsf{T}}\Sigma^{-1}(x_i-\mu_2) \tag{7}$$

Taking the derivative with respect to Σ:

$$\begin{aligned}
D_{\Sigma^{-1}} l_{\Sigma}(\theta) =& \frac{N}{2}\Sigma - \frac{1}{2}\sum_{I} y_i \frac{\partial}{\partial \Sigma^{-1}} tr[(x_i-\mu_1)^{\mathsf{T}}\Sigma^{-1}(x_i-\mu_1)] - \frac{1}{2}\sum_{I}(1-y_i)\frac{\partial}{\partial \Sigma^{-1}} tr[(x_i-\mu_2)^{\mathsf{T}}\Sigma^{-1}(x_i-\mu_2)] \\
=& \frac{N}{2}\Sigma - \frac{1}{2}\sum_{I} y_i \frac{\partial}{\partial \Sigma^{-1}} tr[(x_i-\mu_1)(x_i-\mu_1)^{\mathsf{T}}\Sigma^{-1}] - \frac{1}{2}\sum_{I}(1-y_i)\frac{\partial}{\partial \Sigma^{-1}} tr[(x_i-\mu_2)(x_i-\mu_2)^{\mathsf{T}}\Sigma^{-1}] \\
=& \frac{N}{2}\Sigma - \frac{1}{2}\sum_{I} y_i(x_i-\mu_1)(x_i-\mu_1)^{\mathsf{T}} - \frac{1}{2}\sum_{I}(1-y_i)(x_i-\mu_2)(x_i-\mu_2)^{\mathsf{T}}
\end{aligned}$$

Finally setting to zero we have:

$$\Sigma = \frac{1}{N}[\sum_{I} y_i(x_i-\mu_1)(x_i-\mu_1)^{\mathsf{T}} + \sum_{I}(1-y_i)(x_i-\mu_2)(x_i-\mu_2)^{\mathsf{T}}] \tag{8}$$

Note on notation: We have used subscript in x_i as indicating the ith sample of the random variable and not its ith component.

b) Let $\pi = \pi_1$ and $1 - \pi = \pi_2$ for notational convenience. Moreover let the events $Y = 1$ and $Y = 0$ be denoted C_1 and C_2 respectively for the same reason.
By Baye's theorem we have:

$$p(C_1|x) = \frac{p(x|C_1)P(C_1)}{p(x|C_1)P(C_1) + p(x|C_2)P(C_2)} \tag{9}$$

$$= \frac{\pi_1\mathcal{N}(\mu_1,\Sigma)}{\pi_1\mathcal{N}(\mu_1,\Sigma) + \pi_2\mathcal{N}(\mu_2,\Sigma)} \tag{10}$$

Now letting $\alpha \triangleq \log\frac{\pi_1\mathcal{N}(\mu_1,\Sigma)}{\pi_2\mathcal{N}(\mu_2,\Sigma)}$ we have that :

$$p(C_1|x) = \frac{1}{1+\exp(-\alpha)} \triangleq \sigma(\alpha) \tag{11}$$

Hence we have a form that looks a lot like linear regression, where the posterior $P(\text{Class}|X) = \sigma(f(x))$ and $f(x)$ some function of x. However the logistic regression is a discriminant classifier and $f(x)$ depends directly on the input data whereas the model we are analyzing is a generative one. Moreover the input to the logit function is not directly dependent on x but only through latent variables which have themselves to be estimated.