Probabilistic Graphical Model: Homework 1

Paul Chauvin and Louis Tamames

November 2020

1 Exercise 1: Linear classification

In this exercice, the dependant variables $x1, x2, .., x_n$ are columns vectors. We denote $x = (x_1, ..., x_n)$. Here $n = 2$.

1.1 Generative model

a) Let $y\tilde{}B(\pi)$, the likelihood of the model is The MLE of π is the usual MLE of the Bernouilli which is $\hat{\pi} = \frac{1}{n}\sum_{i=1}^{n} y_i$.
We denote $\theta = (\pi, \mu_1, \mu_2, \Sigma)$

$$\begin{aligned}
L(\theta) &= \sum_{i=1}^{n} \log p(x_i, y_i|\theta) \\
&= \sum_{i=1}^{n} \log([\pi N(x_i; \mu_1, \Sigma_1)]^{y_i} [(1-\pi) N(x_i; \mu_0, \Sigma_0)]^{1-y_i}) \\
&= \sum_{i=1}^{n} y_i(\log(\pi) - \frac{1}{2}(d\log(2pi) + \log(|\Sigma|) + (x_i - \mu_1)^T \Sigma^{-1}(x_i - \mu_1) \\
&+ (1-y_i)(\log(1-\pi) - \frac{1}{2}(d\log(2pi) + \log(|\Sigma|) + (x_i - \mu_0)^T \Sigma^{-1}(x_i - \mu_0))
\end{aligned}$$

MLE for μ:

$$L(\theta) = -\frac{1}{2}\sum_{i=1}^{n} y_i(x_i - \mu_1)^T \Sigma^{-1}(x_i - \mu_1) + C_{\pi,\mu_0,\Sigma} = -\frac{1}{2}\sum_{i=1}^{n} y_i[\mu_1^T \Sigma^{-1}\mu_1 - 2\mu_1^T \Sigma^{-1} x_i] + C_{\pi,\mu_0,\Sigma}$$

This function is concav in μ_1.

$$\frac{\partial L}{\partial \mu_1}(\theta) = -\frac{1}{2}\sum_{i=1}^{n} y_i[2\Sigma^{-1}\mu_1 - 2\Sigma^{-1}x_i] = 0 \iff \Sigma^{-1}\sum_{i=1}^{n} y_i\mu_1 = \Sigma^{-1}\sum_{i=1}^{n} y_i x_i \iff \hat{\mu}_1 = \frac{\sum_{i=1}^{n} y_i x_i}{\sum_{i=1}^{n} y_i}$$

With symmetric computations we get that

$$\hat{\mu}_0 = \frac{\sum_{i=1}^{n}(1-y_i)x_i}{n - \sum_{i=1}^{n} y_i}$$

MLE for Σ:

$$L(\theta) = -\frac{n}{2}\log|\Sigma| - \frac{1}{2}\sum_{i=1}^{n} y_i(x_i - \mu_1)^T \Sigma^{-1}(x_i - \mu_1) + (1-y_i)(x_i - \mu_0)^T \Sigma^{-1}(x_i - \mu_0)) + C$$

This function is concav with respect to Σ.

As $\nabla_\Sigma \log|\Sigma| = (\Sigma^{-1})^T = (\Sigma^{-T})^{-1} = \Sigma^{-1}$

$$\frac{\partial L}{\partial \Sigma}(\theta) = -\frac{n}{2}\Sigma^{-1} - \frac{1}{2}\sum_{i=1}^{n} \Sigma^{-1} y_i(x_i - \mu_1)(x_i - \mu_1)^T \Sigma^{-1} + \Sigma^{-1}(1-y_i)(x_i - \mu_0)(x_i - \mu_0)^T \Sigma^{-1} = 0$$

Multiplying by Σ on both side of the expression we get that

$$\frac{\partial L}{\partial \Sigma}(\theta) = 0 \iff -\frac{n}{2}\Sigma - \frac{1}{2}\sum_{i=1}^{n} y_i(x_i-\mu_1)(x_i-\mu_1)^T + (1-y_i)(x_i-\mu_0)(x_i-\mu_0)^T = 0$$

$$\iff \hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} y_i(x_i-\mu_1)(x_i-\mu_1)^T + (1-y_i)(x_i-\mu_0)(x_i-\mu_0)^T$$

b)

$$\begin{aligned} p(y=1|x) &= \frac{p(x|y=1)\pi}{p(x)} \\ &= \frac{p(x|y=1)\pi}{p(x|y=1)\pi + p(x|y=0)(1-\pi)} \\ &= \frac{1}{1+\frac{p(x|y=0)(1-\pi)}{p(x|y=1)\pi}} \\ &= \frac{1}{1+\frac{(1-\pi)}{\pi}\exp(-\frac{1}{2}a)} \\ &= f(a) \end{aligned}$$

where f is the logistic function and

$$a = [(x-\mu_0)^T\Sigma^{-1}(x-\mu_0) - (x-\mu_1)^T\Sigma^{-1}(x-\mu_1)] = x^T\beta + b$$

with

$$\begin{aligned} \beta &= \Sigma^{-1}(\mu_1-\mu_0) \\ b &= \frac{1}{2}(\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1) - \log(\frac{\pi}{1-\pi}) \end{aligned}$$

1.2 Logistic regression

In the next exercises we denote $w = (w1, w2)^T$

a) For logistic regression we have the following log-likelyhood:

$$l(w) = \sum_{i=1}^{n} y_i log(h(w^T x_i + b)) + (1-y_i)log(1-h(w^T x_i + b))$$

With $h(z) = \frac{1}{1+exp(-z)}$. To simplify the computations, we set $X = (x, 1)$ where 1 denote the column vector of ones and $W = (w_1, w_2, b)$. Hence, with these notation we have $(XW)_i = w^T x_i + b$ for $i = 1, ..., n$.

b) The line defined by the equation $P(y=1|x) = 1/2$, is equivalent to:

$$P(y=1|x_i) = h((WX^T)_i) = \frac{1}{1+exp(-(WX^T)_i)} = \frac{1}{2}$$

Hence,

$$(WX^T)_i = 0$$

1.3 Linear regression

We want to solve the linear regression:

$$Y = w$$

by solving the normal equation :

$$XW + b = Y$$

So we have the solution: $\hat{W} = (X^T X)^{-1} XY$, with the constant term in the last term of $\hat{W}$.

2 Exercise 2: Gaussian mixtue models and EM

The log-likelihood of the gaussian mixture model is :

$$L_{x_1,...x_n}(\theta) = \sum_{i=1}^{n} log(\sum_{k=1}^{K} \pi_k N(x_i;\mu_k,\sigma_k))$$

With $\theta = (\pi, \mu, \sigma)$.
The isseue is that maximise this function is not easy, so we introduce the iid variables, $z_i \sim M(i,\pi)$, such that $X_i|z_{ik} = 1 \sim N(\mu_k, \sigma_k)$

The log-likelihood become:

$$L_{x_1,...x_n}(\theta) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} log(\pi_k N(x_i;\mu,\sigma_k))$$

Now we can use the EM algorithm to estimate the parameters of this model:
Input: data $X \in \mathbb{R}^{nxp}$, number of cluster K
Init : take random value for π_k, μ_k, σ_k
(1) E: Compute τ
(2) M: Compute $\hat{\pi}$, $\hat{\mu}$, $\hat{\sigma}$
Repeat until log-likelihood don't move or number of iteration is done

To do so we use the formulas we proof during the course :
$\forall i = 1, ..., n, \forall k = 1, ..., K$, we have:

$$\tau_{ik} = \frac{\pi_k N(x_i;\mu_k,\sigma_k)}{\sum_{l=1}^{K} \pi_l N(x_i;\mu_l,\sigma_l)}$$

And to update the parameters:
$\forall k = 1, ..., K$, we fix $n_k = \sum_{i=1}^{n} \tau_{ik}$, then we have:

$$\hat{\pi}_k = (1/n)\sum_{i=1}^{n} \tau_{ik}$$

$$\hat{\mu}_k = (1/n_k)\sum_{i=1}^{n} \tau_{ik} x_i$$

$$\hat{\sigma}_k = (1/n_k)\sum_{i=1}^{n} \tau_{ik}(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

The EM algorithm maximizes the log-likelihood of the Gaussian mixture model, let prove it.

Given θ_l, (θ_l is the value of θ at step l) we want to show that $L(\theta_l) \geq L(\theta_{l+1})$.

First, we can write the function as:

$$L(\theta) = log(p_\theta(x) = log(p_\theta(x,z) - log(p_\theta(z|x)$$

Since $log(p_\theta(x,z)) = log(p_\theta(x) log(p_\theta(z|x))$.

This hold for each value of z if $p_\theta(x,z) > 0$, so we can take the expectation with respect to the corresponding conditional distribution of the latent variables, $p_\theta(z|x)$, and we obtain:

$$L(\theta) = L_{\theta_l}(\theta) - \sum_z p_{\theta_l}(z|x) log(p_\theta(z|x))$$

with $L_{\theta_l}(\theta) = \sum_z p_{\theta_l}(z|x) log(p_\theta(x,z))$

Now the difference in log-likelihood is:

$$L(\theta) - L(\theta_l) = L_{\theta_l}(\theta) - L_{\theta_l}(\theta_l) + D_{KL}(\theta_l||\theta)$$

where $D_{KL}(\theta_l||\theta) = \sum_z p_{\theta_l}(z|x) log(\frac{p_{\theta_l}(z|x))}{p_\theta(z|x))}$, this quantity is non-negative.

So applying this to θ_l and θ_{l+1}, we find:

$$L(\theta_{l+1}) - L(\theta_l) = L_{\theta_l}(\theta_{l+1}) - L_{\theta_l}(\theta_l) + D_{KL}(\theta_l||\theta_{l+1}) \geq D_{KL}(\theta_l||\theta_{l+1}) \geq 0$$

We have $L_{\theta_l}(\theta_{l+1}) - L_{\theta_l}(\theta_l) \geq 0$ since θ_{l+1} maximise $L_{\theta_l}(.)$ after the M step.

Given that the sequence of log-likelihoods $(L(\theta_t))_{t\geq 0}$ is non-decreasing and thus converges.