# Probabilistic Graphical Models MVA 2018/2019 HWK1

Louis GUO, Laurent LIN

## Exercise 1: Learning in discrete graphical models

Let $(z_i, x_i)_{i=1,...,n}$ an i.i.d. sample of observations. Let $\boxed{n_{k,m} = \sum_{i=1}^{n} 1_{z_i=m, x_i=k}}$ and $\boxed{n_m = \sum_{i=1}^{n} 1_{z_i=m}}$ for $m \in \{1, ..., M\}$ and $k \in \{1, ..., K\}$.

The maximum likelihood estimator of $\pi$ and $\theta$ are found by respectively considering the distribution of $z$ and $x|z$:

$$\boxed{\hat{\pi_m} = \frac{n_m}{n}}, \boxed{\hat{\theta_{k,m}} = \frac{n_{k,m}}{n}}$$

## Exercise 2.1.(a): LDA formulas

Let $(x_i, y_i)_{i=1,...,n}$ an i.i.d. sample of observations. For estimating $\pi$, we consider the distribution of $y \sim Bernoulli(p)$, for $\Sigma$, $\mu_1$ and $\mu_0$, we consider the negative log-likelihood of $x|y$ distribution knowing that $p(X|Y = j) = Normal(\mu_j, \Sigma)$ for $j \in \{0, 1\}$. Let $\boxed{n_j = \sum_{i=1}^{n} 1_{y_i=j}}$, $\boxed{\widetilde{\Sigma_j} = \frac{\sum_i^n 1_{y_i=j}(x_i - \mu_j)(x_i - \mu_j)^T}{n_j}}$ for $j \in \{0, 1\}$.

The maximum likelihood estimators are :

$$\boxed{\hat{\pi} = \frac{\sum_{i=1}^n y_i}{n}}, \boxed{\hat{\mu_j} = \frac{\sum_i^n x_i 1_{y_i=j}}{n_j}}, \boxed{\hat{\Sigma} = \frac{n_1\widetilde{\Sigma_1} + n_0\widetilde{\Sigma_0}}{n}}$$

With Bayes formula, we have $p(y = 1|x) = \frac{\pi p(x|y=1)}{\pi p(x|y=1)+(1-\pi)p(x|y=0)} = \frac{1}{1+\frac{1-\pi}{\pi}\frac{p(x|y=0)}{p(x|y=1)}} = \sigma(\omega^T x + C)$ with $\omega = \Sigma^{-1}(\mu_0 - \mu_1)$, C a constant wrt to $x$ and $\sigma$ the sigmoid function. We recognize here the logistic regression model.

## Exercise 2.5.(a): QDA formulas

We will keep the same notations as for 2.1.(a). For estimating $\pi$, we consider the distribution of $y \sim Bernoulli(p)$, for $\Sigma_1$, $\Sigma_0$, $\mu_1$ and $\mu_0$, we consider the negative log-likelihood of $x|y$ distribution knowing that $p(X|Y = j) = Normal(\mu_j, \Sigma_j)$ for $j \in \{0, 1\}$.

The maximum likelihood estimator of $\pi$, $\Sigma_1$, $\Sigma_0$, $\mu_1$ and $\mu_0$ are :

$$\boxed{\hat{\pi} = \frac{\sum_{i=1}^n y_i}{n}}, \boxed{\hat{\mu_j} = \frac{\sum_{i=1}^n x_i 1_{y_i=j}}{n_j}}, \boxed{\hat{\Sigma_j} = \widetilde{\Sigma_j}}$$

More detailed proves on page 5.

## Exercise 1: Learning in discrete graphical models

Let $(z_i, x_i)_{i=1,...,n}$ an i.i.d. sample of observations.

The negative log-likelihood of the distribution of $z$ is given by:

$$-l(\pi) = -log(p(z_1, ..., z_n)) = -log(\prod_{i=1}^{n} p(z_i)) = -\sum_{i=1}^{n}\sum_{m=1}^{M} log(\pi_m^{1_{z_i=m}}) = -\sum_{m=1}^{M} n_m log(\pi_m) = -n\sum_{m=1}^{M} \frac{n_m}{n} log(\pi_m)$$

with $\boxed{n_m = \sum_{i=1}^{n} 1_{z_i=m}}$ We want to minimize the negative log-likelihood over $\pi$ with constraints $\sum_{m=1}^{M} \pi_m = 1$ and $\pi > 0$, which can be seen as the cross-entropy between discrete distributions $z$ and $\tilde{z}$. ($\tilde{z}$ the empirical discrete variable with $p(\tilde{z} = m) = \frac{n_m}{n}$ such that $-l(\pi) = nH(z, \tilde{z})$. Here $\tilde{z}$ is fixed, cross-entropy takes on its minimal value for $z = \tilde{z}$ i.e. $\boxed{\hat{\pi_m} = \frac{n_m}{n}}$ as the maximum likelihood estimator.

The negative log-likelihood of $x|z$ is given by:

$$-l(\theta) = -log(\prod_{i=1}^{n} p(x_i|z_i)) = -\sum_{i=1}^{n}\sum_{k,m} log(\theta_{k,m}^{1_{z_i=m,x_i=k}}) = -\sum_{k,m} n_{k,m} log(\theta_{k,m})$$

with $\boxed{n_{k,m} = \sum_{i=1}^{n} 1_{z_i=m,x_i=k}}$ . As $\sum_{k,m} n_{k,m} = 1$. With the same cross-entropy argument, we can derive the maximum likelihood estimator $\boxed{\hat{\theta_{k,m}} = \frac{n_{k,m}}{n}}$

## Exercise 2.1.(a): LDA formulas

Let $(x_i, y_i)_{i=1,...,n}$ an i.i.d. sample of observations, we want to find the maximum likelihood estimator of $\pi$, $\Sigma$, $\mu_1$ and $\mu_0$. Let $\boxed{n_j = \sum_{i=1}^{n} 1_{y_i=j}}$, $\boxed{\widetilde{\Sigma_j} = \frac{\sum_i^n 1_{y_i=j}(x_i - \mu_j)(x_i - \mu_j)^T}{n_j}}$ for $j \in \{0, 1\}$. For estimating $\pi$, we consider the distribution of $y \sim Bernoulli(p)$, hence $\boxed{\hat{\pi} = \frac{\sum_{i=1}^{n} y_i}{n}}$. For the rest, we consider the negative log-likelihood of $x|y$ distribution knowing that $p(X|Y = j) = Normal(\mu_j, \Sigma)$ :

$$\begin{aligned}
-l(\mu_0, \mu_1, \Sigma) &= -log(\prod_i p(x = x_i|y = 0)^{1_{y_i=0}} p(x = x_i|y = 1)^{1_{y_i=1}}) \\
&= \sum_{i=1}^{n} -log(\frac{\exp(\frac{-(x_i-\mu_1)^T\Sigma^{-1}(x_i-\mu_1)}{2})^{y_i} \exp(\frac{-(x_i-\mu_0)^T\Sigma^{-1}(x_i-\mu_0)}{2})^{1-y_i}}{2\pi\sqrt{\det(\Sigma)}}) \\
&= n\log(2\pi) + \frac{n}{2}\log(\det(\Sigma)) + \frac{1}{2}\sum_{i,y_i=0}(x_i - \mu_0)^T\Sigma^{-1}(x_i - \mu_0) + \frac{1}{2}\sum_{i,y_i=1}(x_i - \mu_1)^T\Sigma^{-1}(x_i - \mu_1) \\
&= n\log(2\pi) - \frac{n}{2}\log(\det(\Sigma^{-1})) + \frac{n_1}{2}Tr(\Sigma^{-1}\widetilde{\Sigma_1}) + \frac{n_0}{2}Tr(\Sigma^{-1}\widetilde{\Sigma_0})
\end{aligned}$$

For $j \in \{0, 1\}$, $\nabla_{\mu_j}(-l) = \sum_{i,y_i=j} \Sigma^{-1}(\mu_j - x_i) = \Sigma^{-1}(\sum_{i,y_i=j} \mu_j - x_i) = 0$, if and only if $\boxed{\hat{\mu_j} = \frac{\sum_i^n x_i 1_{y_i=j}}{n_j}}$, it's minimum as $-l$ is strictly convex w.r.t to $\mu_j$ since $\nabla^2_{\mu_j}(-l) = \Sigma^{-1}n_j$ is symmetric positive-semidefinite. Since $\nabla_\Sigma(\frac{n_j}{2}Tr(\Sigma^{-1}\widetilde{\Sigma_j})) = \frac{n_j}{2}\widetilde{\Sigma_j}$ and $\nabla_\Sigma(-\frac{n}{2}\log(\det(\Sigma^{-1}))) = -\frac{n}{2}\Sigma$, $\nabla_\Sigma(-l) = \frac{n}{2}\Sigma - \frac{n_1\widetilde{\Sigma_1}+n_0\widetilde{\Sigma_0}}{2} = 0$ if and only if $\boxed{\hat{\Sigma} = \frac{n_1\widetilde{\Sigma_1} + n_0\widetilde{\Sigma_0}}{n}}$, it's the minimum as $-l$ is strictly convex w.r.t to $\Sigma$ since $\nabla^2_\Sigma(-l) = \frac{n}{2}I_n$ is symmetric positive-semidefinite.

With Bayes formula, we have $p(y = 1|x) = \frac{\pi p(x|y=1)}{\pi p(x|y=1)+(1-\pi)p(x|y=0)} = \frac{1}{1+\frac{1-\pi}{\pi}\frac{p(x|y=0)}{p(x|y=1)}}$

$$\begin{aligned}
\frac{1-\pi}{\pi}\frac{p(x|y = 0)}{p(x|y = 1)} &= \frac{1-\pi}{\pi}\exp(-\frac{||x - \mu_0||^2_{\Sigma^{-1}} - ||x - \mu_1||^2_{\Sigma^{-1}}}{2}) \\
&= \exp(\log(\frac{1-\pi}{\pi}) - \frac{-2 < x|\mu_0 - \mu_1 >_{\Sigma^{-1}} +(||\mu_0||^2_{\Sigma^{-1}} - ||\mu_1||^2_{\Sigma^{-1}})}{2}) \\
&= \exp(\omega^T x + C)
\end{aligned}$$

with $\omega = \Sigma^{-1}(\mu_0 - \mu_1)$, C a constant wrt to $x$ and $\sigma$ the sigmoid function. Hence $\boxed{p(y=1|x) = \sigma(\omega^T x + C)}$ : we recognize the logistic regression model. The assumption of equal covariance matrices cancel the quadratic part in the exponents. The decision boundary are then linear in $x$: classifications regions will be separated by hyperplanes.

## Exercise 2.5.(a): QDA formulas

We will keep the same notations and we want to find the maximum likelihood estimator of $\pi$, $\Sigma_1$, $\Sigma_0$, $\mu_1$ and $\mu_0$. We still have $\boxed{\hat{\pi} = \frac{\sum_{i=1}^n y_i}{n}}$. For the rest, we consider the negative log-likelihood of $x|y$ distribution knowing that $p(X|Y=j) = Normal(\mu_j, \Sigma_j)$ :

$$\begin{aligned} -l(\mu_0, \mu_1, \Sigma_0, \Sigma_1) &= -log(\prod_i p(x = x_i|y=0)^{1_{y_i=0}} p(x=x_i|y=1)^{1_{y_i=1}}) \\ &= n\log(2\pi) - \frac{n_1}{2}\log(\det(\Sigma_1^{-1})) - \frac{n_0}{2}\log(\det(\Sigma_0^{-1})) + \frac{n_1}{2}Tr(\Sigma_1^{-1}\widetilde{\Sigma_1}) + \frac{n_0}{2}Tr(\Sigma_0^{-1}\widetilde{\Sigma_0}) \end{aligned}$$

For $j \in \{0,1\}$, we also obtain $\boxed{\hat{\mu_j} = \frac{\sum_{i=1}^n x_i 1_{y_i=j}}{n_j}}$, it's the minimum as $-l$ is strictly convex w.r.t to $\mu_j$ since $\nabla^2_{\mu_j}(-l) = \Sigma_j^{-1} n_j$ is symmetric positive-semidefinite. $\nabla_{\Sigma_j}(\frac{n_j}{2}Tr(\Sigma_j^{-1}\widetilde{\Sigma_j})) = \frac{n_j}{2}\widetilde{\Sigma_j}$ and $\nabla_{\Sigma_j}(-\frac{n_j}{2}\log(\det(\Sigma_j^{-1}))) = -\frac{n_j}{2}\Sigma_j$, $\nabla_{\Sigma_j}(-l) = \frac{n_j}{2}\Sigma_j - \frac{n_j}{2}\widetilde{\Sigma_j} = 0$, if and only if $\boxed{\hat{\Sigma_j} = \widetilde{\Sigma_j}}$, it's the minimum as $-l$ is strictly convex w.r.t to $\Sigma_j$ since $\nabla^2_{\Sigma_j}(-l) = \frac{n_j}{2} I_n$ is symmetric positive-semidefinite.