Ali Taïga
Ali

Homework 1: Graphical Models

Exercice 2: Linear classification

1.(a.) The log-likelihood l is given by:

$$l(\pi, \mu_1, \mu_2, \Sigma) = \sum_{i=1}^{n} \log(p(y_i, x_i ; \mu_1, \mu_2, \pi, \Sigma))$$

$$= \sum_{i=1}^{n} \log(p(y_i)) + \log(p(x_i | y_i))$$

$$= \sum_{i=1}^{n} y_i \times \log(\pi) + (1-y_i)\log(1-\pi) + \sum_{i=1}^{n} \log(p(x_i | y_i ; \mu_1, \mu_2, \Sigma))$$

We obtain two seperate terms, one for the marginal distribution of Y and the other for the conditional distribution of X_j given Y. Maximizing with respect to π involves only the former term, and for π we therefore obtain:

$$\hat{\pi}_{ML} = \arg\max_{\pi} \sum_{i=1}^{n} \log(y_i | \pi)$$

$$= \arg\max_{\pi} \sum_{i=1}^{n} \{y_i \times \log\pi + (1-y_i)\log(1-\pi)$$

Which is equal to $\hat{Y} = \frac{1}{n} \times \sum_{i=1}^{n} y_i$ (maximum likelihood for the bernoulli model)

The second term is equal to:

$$(1)\quad \sum_{\substack{i=1\\y_i=1}}^{n} -\log(2\pi) - \frac{1}{2}\log(\det(\Sigma)) - \frac{1}{2}{}^t(x_i-\mu_1)\Sigma^{-1}(x_i-\mu_1)$$

$$+ \sum_{\substack{i=1\\y_i=2}}^{n} -\log(2\pi) - \frac{1}{2}\log(\det\Sigma) - \frac{1}{2}{}^t(x_i-\mu_0)\Sigma^{-1}(x_i-\mu_0)$$

With $m_1 = \sum_{\substack{i=1\\y_i=1}}^{n} 1$ and $m_0 = \sum_{\substack{i=1\\y_i=2}}^{n} 1$, we have:

$$= -\log(2\pi)(m_1+m_0) - \frac{1}{2}\log(\det(\Sigma)) \times (m_1+m_0)$$

$$-\frac{1}{2}\mathrm{Tr}\Big(\Sigma^{-1}\Big(\underbrace{\sum_{\substack{i=1\\y_i=1}}^{n}(x_i-\mu_1)^t(x_i-\mu_1)}_{\tilde{\Sigma}_1} + \underbrace{\sum_{\substack{i=1\\y_i=0}}^{n}(x_i-\mu_0)^t(x_i-\mu_0)}_{\tilde{\Sigma}_0}\Big)$$

$$(2)\quad = -n\log(2\pi) + \frac{n}{2}\underbrace{\log(\det(\Sigma^{-1}))}_{(*)} - \frac{1}{2}\mathrm{Tr}\big(\Sigma^{-1}\times(\tilde{\Sigma}_1+\tilde{\Sigma}_0)\big)$$

In (1) the expression is concave in μ_1:

$$\nabla_{\mu_1} l(\mu_1,\mu_0,\Sigma) = 0 \iff \sum_{\substack{i=1\\y_i=1}}^{n} \Sigma^{-1}(\mu_1 - x_i) = 0$$

$$\iff \Sigma^{-1}\Big(m_1\mu_1 - \sum_{\substack{i=1\\y_i=1}}^{n} x_i\Big) = 0$$

$$\iff \hat{\mu}_1 = \frac{1}{m_1} \times \sum_{\substack{i=1\\y_i=1}}^{n} x_i$$

We also have by the same way: $\hat{\mu}_0 = \frac{1}{m_0} \times \sum_{\substack{i=1\\y_i=0}}^{n} x_i$

(*) $\log(\det(\Sigma)) = \log(1/\det(\Sigma^{-1})) = -\log(\det(\Sigma^{-1}))$

Expression (2) is also concave in Σ:

$$\nabla_{\Sigma}\, l(\mu_0, \mu_1, \Sigma) = + \frac{m}{2} \cdot \Sigma - \frac{1}{2} \times (\tilde{\Sigma}_1 + \tilde{\Sigma}_0) = 0$$

$$\Longrightarrow \quad \hat{\Sigma} = \frac{1}{m} \times \Big(\sum_{\substack{i=1 \\ y_i=1}}^{m} (x_i - \mu_1)^t (x_i - \mu_1) + \sum_{\substack{i=1 \\ y_i=0}}^{m} (x_i - \mu_0)^t (x_i - \mu_0) \Big)$$

(b) $$p(y=1 \mid x) = \frac{p(x \mid y=1) \times p(y=1)}{p(x)}$$

$$= \frac{p(x \mid y=1) \times p(y)}{p(x \mid y=1) \times p(y=1) + p(x \mid y=0) \times p(y=0)}$$

$$= \frac{\pi \times \exp\left(-1/2\, {}^t(x-\mu_1) \cdot \Sigma^{-1} (x-\mu_1)\right)}{\pi \times \exp\left(-1/2 \cdot {}^t(x-\mu_1) \Sigma^{-1} (x-\mu_1)\right) + (1-\pi) \exp\left(-1/2\, {}^t(x-\mu_0) \Sigma^{-1} (x-\mu_0)\right)}$$

$$= \frac{1}{1 + \exp\left(-\log\left(\frac{\pi}{1-\pi}\right) + \frac{1}{2} \times {}^t(x-\mu_1) \Sigma^{-1} (x-\mu_1) - \frac{1}{2}\, {}^t(x-\mu_0) \Sigma^{-1} (x-\mu_0)\right)}$$

$$= \frac{1}{1 + \exp(-{}^t\beta x - \gamma)}$$

with $\beta = \Sigma^{-1} (\mu_1 - \mu_0)$

$\gamma = -\frac{1}{2} \cdot {}^t(\mu_1 - \mu_0) \cdot \Sigma^{-1} (\mu_1 + \mu_0) + \log\left(\frac{\pi}{1-\pi}\right)$

(c) $p(y=1|x) = \frac{1}{2} \iff {}^t\beta . x + \gamma = 0$

The decision boundary is a line in the feature space

4.

5.

5. (a). This time the second member of the log-likelihood is separated in two independent parts responsible for each gaussian:

$$\sum_{\substack{i=1 \\ y_i=1}}^{m} -\log(2\pi) - \frac{1}{2}\log(\det \Sigma_1) - \frac{1}{2} . {}^t(x_i - \mu_1) . \Sigma_1^{-1} . (x_i - \mu_1)$$

$$+ \sum_{\substack{i=1 \\ y_i=0}}^{m} -\log(2\pi) - \frac{1}{2}\log(\det \Sigma_0) - \frac{1}{2} . {}^t(x_i - \mu_0) . \Sigma_0^{-1} (x_i - \mu_0)$$

The maximum likelihood estimate for each gaussian is given by:

$$\mu_1 = \sum_{\substack{i=1 \\ y_i=1}}^{m} x_i \Big/ m_1 \qquad \Sigma_1 = \frac{\sum_{\substack{i=1 \\ y_i=1}}^{m} (x_i - \mu_1) . {}^t(x_i - \mu_1)}{m_1}$$

$$\mu_0 = \frac{\sum_{\substack{i=1 \\ y_i=0}}^{m} x_i}{m_0} \qquad \Sigma_0 = \frac{\sum_{\substack{i=1 \\ y_i=0}}^{m} (x_i - \mu_0) . {}^t(x_i - \mu_0)}{m_0}$$

(b) This time:

$$p(y=1 \mid x) = \frac{1}{1 + \exp\left(-\log\left(\frac{\pi}{1-\pi}\right) + \frac{1}{2} \cdot {}^t(x-\mu_1) \cdot \Sigma_1^{-1}(x-\mu_1) - \frac{1}{2}{}^t(x-\mu_0)\Sigma_0^{-1}(x-\mu_0)\right)}$$

$$= \frac{1}{1 + \exp\left(-{}^t x . A . x - {}^t\beta x - \gamma\right)}$$

with $A = \frac{\Sigma_0^{-1} - \Sigma_1^{-1}}{2}$ $\quad \beta = \Sigma_1^{-1} . \mu_1 - \Sigma_0^{-1} . \mu_0$

$$\gamma = \log\left(\frac{\pi}{1-\pi}\right) - \frac{1}{2}\left(\mu_1 . \Sigma_1^{-1} . \mu_1 - \mu_0 . \Sigma_0^{-1} . \mu_0\right)$$

and the décision boundary is given by ${}^t x . A . x + {}^t\beta x + \gamma = 0$

Exercice 1: Learning in discrete graphical models

The log likelihood for the sample $((x_i, y_i))_{1 \leq i \leq n}$ is given by:

$$\ell(\pi_1, \dots, \pi_M, \theta_{11}, \dots \theta_{MK}) = \sum_{i=1}^{n} \log(p(x_i, z_i)) = \sum$$

$$= \sum_{i=1}^{n} \log(p(z_i)) + \sum_{i=1}^{n} \log(p(x_i \mid z_i))$$

Where we see that we obtain two separate terms, one for the marginal distribution of Z and the other for the conditional distribution of X_j given Z, maximizing with respect to $\pi_1, \dots, \pi_M$ involve only the first term

$$f_1(\pi_1, \dots, \pi_M) = \sum_{i=1}^{n} \log(z_i) = \sum_{i=1}^{n} \sum_{m=1}^{M} (z_i = m) \times \log(\pi_m)$$

with $(z_i = m) = 1$ if $z_i = 0$ otherwise 0

$$= \sum_{m=1}^{M} n_m \log(\pi_m) \quad \text{with} \quad n_m = \sum_{i=1}^{n} (z_i = m)$$

It's a multinomial model; the maximum likelihood estimator is the solution of the problem:

$$\text{minimize} \quad -\sum_{m=1}^{M} n_m \log(\pi_m)$$

$$\text{with} \quad \sum_{m=1}^{M} \pi_m = 1$$

$\forall m \quad n_m \geq 0$, it's a convex optimization problem

There exist $\pi_1, \dots, \pi_M$ such that $\sum_m \pi_m = 1$, hence Slater's constraint qualification is verified and this problem verify the strong duality property : $p^* = d^*$.

The lagrangian of the problem is given by:

$$L(\pi_1, \dots, \pi_M, \nu) = -\sum_{m=1}^{M} n_m \log(\pi_m) + \nu\left(\sum_{m=1}^{M} \pi_m - 1\right)$$

L is convex in $\pi_1, \dots, \pi_M$ hence we can compute its derivatives to find its minimum:

$$\forall k \quad \frac{\partial L}{\partial \pi_k} = -\frac{n_k}{\pi_k} + \nu = 0$$

$$\Rightarrow \quad \pi_k = \frac{n_k}{\nu} \quad \forall k$$

Besides $\sum_{m=1}^{M} \pi_m = 1 \Rightarrow \nu = n$

$$\text{hence} \quad \forall m \in [1; M] \quad \hat{\pi}_m = \frac{\sum_{i=1}^{n} (z_i = m)}{n}$$

For the second term of the log-likelihood:

$$l_2(\theta_{11}, \dots, \theta_{MK}) = \sum_{i=1}^{m} \log(p(x_i | z_i))$$

$$= \sum_{i=1}^{m} \log\left(\prod_{\substack{1 \leq m \leq M \\ 1 \leq k \leq K}} \theta_{mk}\right) = \sum_{i=1}^{m} \sum_{m=1}^{M} \sum_{k=1}^{K} (z_i = m, x_i = k) \times \log(\theta_{mk})$$

$$= \sum_{m=1}^{M} \sum_{k=1}^{K} \alpha_{mk} \times \log(\theta_{mk}) \quad \text{with} \quad \alpha_{mk} = \sum_{i=1}^{m} (x_i = k, z_i = m)$$

$$\text{and} \sum_{\substack{1 \leq m \leq M \\ 1 \leq k \leq K}} \theta_{mk} = 1$$

As previously we can show that:

$$\forall m \in [1, M], k \in [1, K] \quad \theta_{mk} = \frac{\sum_{i=1}^{m} (z_i = m, x_i = k)}{m}$$