Université du Sud Toulon - Var

Master 2 Informatique

Probabilistic Learning and Data Anlysis

TD1 : Estimation of Gaussians and logistic regression for classification

by Faicel CHAMROUKHI

Solution

1 Learning of a multivariate Gaussian density model

The likelehood to be maximized is given as the joint probability density function for sample $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ of n independent identically distributed normal random variables

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \prod_{i=1}^{n} f(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)\right). \quad (1)$$

Maximizing this likelihood is equivalent to maximizing the following log-likelihood function

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) &= \log \prod_{i=1}^{n} \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}_k|^{1/2}} \exp\left(-\frac{1}{2} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)\right) \\
&= -\frac{nd}{2} \log 2\pi - \frac{n}{2} \log |\boldsymbol{\Sigma}_k| - \frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k). \quad (2)
\end{aligned}$$

1.1 ML estimation for the expectation (mean vector)

Taking the derivative with respect to $\boldsymbol{\mu}_k$ is straightforward and is given by

$$
\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\partial \boldsymbol{\mu}_k} &= \frac{\partial - \frac{1}{2}\sum_{i=1}^{n} \left(\mathbf{x}_i - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu}_k\right)}{\partial \boldsymbol{\mu}_k} = -\frac{1}{2}\frac{\sum_{i=1}^{n} \partial \left(\mathbf{x}_i - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu}_k\right)}{\partial \boldsymbol{\mu}_k} \\
&= -\frac{1}{2}\sum_{i=1}^{n} \frac{\partial \left(\mathbf{x}_i - \boldsymbol{\mu}_k\right)^T}{\partial \boldsymbol{\mu}_k} \boldsymbol{\Sigma}_k^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu}_k\right) + \left(\mathbf{x}_i - \boldsymbol{\mu}_k\right)^T \frac{\partial \boldsymbol{\Sigma}_k^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu}_k\right)}{\partial \boldsymbol{\mu}_k} \qquad (3) \\
&= -\frac{1}{2}\sum_{i=1}^{n} \left(\boldsymbol{\Sigma}_k^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu}_k\right)\right)^T \frac{\partial \left(\mathbf{x}_i - \boldsymbol{\mu}_k\right)}{\partial \boldsymbol{\mu}_k} + \left(\mathbf{x}_i - \boldsymbol{\mu}_k\right)^T \frac{\partial \boldsymbol{\Sigma}_k^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu}_k\right)}{\partial \boldsymbol{\mu}_k} \qquad (4) \\
&= -\frac{1}{2}\sum_{i=1}^{n} \left(\mathbf{x}_i - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1} \frac{\partial \left(\mathbf{x}_i - \boldsymbol{\mu}_k\right)}{\partial \boldsymbol{\mu}_k} + \left(\mathbf{x}_i - \boldsymbol{\mu}_k\right)^T \frac{\partial \boldsymbol{\Sigma}_k^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu}_k\right)}{\partial \boldsymbol{\mu}_k} \qquad (5) \\
&= -\frac{1}{2}\sum_{i=1}^{n} - \left(\mathbf{x}_i - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1} - \left(\mathbf{x}_i - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1} \qquad (6) \\
&= \sum_{i=1}^{n} \left(\mathbf{x}_i - \boldsymbol{\mu}_k\right)^T \boldsymbol{\Sigma}_k^{-1} \qquad (7)
\end{aligned}
$$

Here we used standard results as follows
(3) $\frac{\partial YZ}{X} = \frac{\partial Y}{\partial X} Z + Y \frac{\partial Z}{\partial X}$
(4) a scaler is equal to its transpose, thatis for the scalar $\mathbf{a}^T\mathbf{b}$ we have $\mathbf{a}^T\mathbf{b} = (\mathbf{a}^T\mathbf{b})^T$
(5) $(AB)^T = B^T A$
(6) $(\boldsymbol{\Sigma}_k^{-1})^T = \boldsymbol{\Sigma}_k^{-1}$, $\boldsymbol{\Sigma}_k^{-1}$ being symmetric

Setting to 0, multiplying both sides by $\boldsymbol{\Sigma}_k$ and we get the ML estimate for $\boldsymbol{\mu}_k$, that is

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i$$

which is the sample mean. This is an unbiased estimator for $\boldsymbol{\mu}_k$ since

$$\mathbb{E}[\widehat{\boldsymbol{\mu}_k}] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\boldsymbol{X}_i] = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\mu}_k = \boldsymbol{\mu}_k$$

1.2 ML estimation for the covariance matrix

The derivative with respect to $\mathbf{\Sigma}_k^{-1}$ is given by

$$\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)}{\partial \mathbf{\Sigma}_k^{-1}} &= \frac{\partial \left(-\frac{n}{2} \log |\mathbf{\Sigma}_k| - \frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right)}{\partial \mathbf{\Sigma}_k^{-1}} \\
&= \frac{n}{2} \frac{\partial \log |\mathbf{\Sigma}_k^{-1}|}{\partial \mathbf{\Sigma}_k^{-1}} - \frac{1}{2} \sum_{i=1}^{n} \frac{\partial (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k)}{\partial \mathbf{\Sigma}_k^{-1}} && (8) \\
&= \frac{n}{2} \frac{1}{\mathbf{\Sigma}_k^{-1}} - \frac{1}{2} \sum_{i=1}^{n} \frac{\partial trace \left((\mathbf{x}_i - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}_k^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_k) \right)}{\partial \mathbf{\Sigma}_k^{-1}} && (9) \\
&= \frac{n}{2} \frac{1}{\mathbf{\Sigma}_k^{-1}} - \frac{1}{2} \sum_{i=1}^{n} \frac{\partial trace \left((\mathbf{x}_i - \boldsymbol{\mu}_k) (\mathbf{x}_i - \boldsymbol{\mu}_k)^T \mathbf{\Sigma}_k^{-1} \right)}{\partial \mathbf{\Sigma}_k^{-1}} && (10) \\
&= \frac{n}{2} \mathbf{\Sigma}_k - \frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu}_k) (\mathbf{x}_i - \boldsymbol{\mu}_k)^T && (11)
\end{aligned}$$

Here we used standard results as follows

(8) $\frac{\partial \log |A|}{\partial A} = \frac{1}{A}$

(9) $\mathbf{x}^T A \mathbf{x} = trace(\mathbf{x}^T A \mathbf{x})$

(10) $trace(\mathbf{x}^T A \mathbf{x}) = trace(\mathbf{x}\mathbf{x}^T A)$

(11) $\frac{\partial trace(BA)}{A} = B^T$

Finally, setting to zero and using the ML estimate $\hat{\boldsymbol{\mu}}_k$ for $\boldsymbol{\mu}_k$ yields the maximum likelihood estimate

$$\hat{\mathbf{\Sigma}}_k = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k) (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k)^T \qquad (12)$$

which is simply the sample covariance matrix. We can see that the ML estimator of $\mathbf{\Sigma}_k$

$$\widehat{\mathbf{\Sigma}}_k = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_k) (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_k)^T$$

is biased. Indeed, we can show similarly as we have seen for the variance (last year), that $\mathbb{E}[\widehat{\mathbf{\Sigma}}_k] = \frac{n-1}{n} \mathbf{\Sigma}_k$. An unbiased estimator is therefore

$$\widehat{\mathbf{\Sigma}}_k = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_k) (\boldsymbol{X}_i - \widehat{\boldsymbol{\mu}}_k)^T$$

2 Learning of a binary logistic regression model

The conditional likelihood of $\boldsymbol{w}$ for the binary logistic regression model given the labeled training set of independent observations $((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n))$ is given by

$$p(y_1, \ldots, y_n | \mathbf{x}_1, \ldots, \mathbf{x}_n; \boldsymbol{w}) = \prod_{i=1}^{n} p(y_i | \mathbf{x}_i; \boldsymbol{w}) \tag{13}$$

By using the fact that

$$\begin{aligned} p(y_i = 1 | \mathbf{x}_i; \boldsymbol{w}) &= \pi(\mathbf{x}_i; \boldsymbol{w}) \\ p(y_i = 0 | \mathbf{x}_i; \boldsymbol{w}) &= 1 - \pi(\mathbf{x}_i; \boldsymbol{w}) \end{aligned}$$

we can write

$$p(y_i | \mathbf{x}_i; \boldsymbol{w}) = \pi(\mathbf{x}_i; \boldsymbol{w})^{y_i} \left(1 - \pi(\mathbf{x}_i; \boldsymbol{w})\right)^{1-y_i}$$

whith $y_i \in \{0, 1\}$.

The likelihood (13) car therefore be rewritten as

$$p(y_1, \ldots, y_n | \mathbf{x}_1, \ldots, \mathbf{x}_n; \boldsymbol{w}) = \prod_{i=1}^{n} \pi(\mathbf{x}_i; \boldsymbol{w})^{y_i} \left(1 - \pi(\mathbf{x}_i; \boldsymbol{w})\right)^{1-y_i} \tag{14}$$

and the log-likelihood is then given by

$$\begin{aligned} \mathcal{L}(\boldsymbol{w}) &= \log p(y_1, \ldots, y_n | \mathbf{x}_1, \ldots, \mathbf{x}_n; \boldsymbol{w}) \\ &= \sum_{i=1}^{n} \log \left[\pi(\mathbf{x}_i; \boldsymbol{w})^{y_i} \left(1 - \pi(\mathbf{x}_i; \boldsymbol{w})\right)^{1-y_i} \right] \\ &= \sum_{i=1}^{n} y_i \log \pi(\mathbf{x}_i; \boldsymbol{w}) + (1 - y_i) \log \left(1 - \pi(\mathbf{x}_i; \boldsymbol{w})\right) \end{aligned} \tag{15}$$

This log-likelihood is convex but can not be maximized in a closed form. The Newton-Raphson (NR) algorithm is generally used to maximize it. A single NR update is given by :

$$\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \left[\frac{\partial^2 \mathcal{L}(\boldsymbol{w})}{\partial \boldsymbol{w} \partial \boldsymbol{w}^T} \right]^{-1} \frac{\partial \mathcal{L}(\boldsymbol{w})}{\partial \boldsymbol{w}} \tag{16}$$

where the Hessian and the gradient of $\mathcal{L}(\boldsymbol{w})$ (which are respectively the second and first derivative of $\mathcal{L}(\boldsymbol{w})$) are evaluated at $\boldsymbol{w} = \boldsymbol{w}^{(t)}$, l being the iteration number.

The gradient is calculated as follows.

$$\begin{aligned} \frac{\partial \mathcal{L}(\boldsymbol{w})}{\partial \boldsymbol{w}} &= \sum_{i=1}^{n} y_i \frac{\partial \log \pi(\mathbf{x}_i; \boldsymbol{w})}{\partial \boldsymbol{w}} + (1 - y_i) \frac{\partial \log \left(1 - \pi(\mathbf{x}_i; \boldsymbol{w})\right)}{\partial \boldsymbol{w}} \\ &= \sum_{i=1}^{n} y_i \frac{\partial \log \pi(\mathbf{x}_i; \boldsymbol{w})}{\partial \boldsymbol{w}} + (1 - y_i) \frac{\partial \log \left(1 - \pi(\mathbf{x}_i; \boldsymbol{w})\right)}{\partial \boldsymbol{w}} \\ &= \sum_{i=1}^{n} y_i \frac{\frac{\partial \pi(\mathbf{x}_i; \boldsymbol{w})}{\partial \boldsymbol{w}}}{\pi(\mathbf{x}_i; \boldsymbol{w})} + (1 - y_i) \frac{\frac{\partial (1 - \pi(\mathbf{x}_i; \boldsymbol{w}))}{\partial \boldsymbol{w}}}{(1 - \pi(\mathbf{x}_i; \boldsymbol{w}))} \end{aligned} \tag{17}$$

The partial derivative of the logistic function $\pi(\mathbf{x}_i; \boldsymbol{w})$ with respect to $\boldsymbol{w}$ is given by

$$\begin{aligned}
\frac{\partial \pi(\mathbf{x}_i; \boldsymbol{w})}{\partial \boldsymbol{w}} &= \frac{\partial \frac{\exp(\boldsymbol{w}^T \mathbf{x}_i)}{1+\exp(\boldsymbol{w}^T \mathbf{x}_i)}}{\partial \boldsymbol{w}} \\
&= \frac{\frac{\partial \exp(\boldsymbol{w}^T \mathbf{x}_i)}{\partial \boldsymbol{w}} \left(1 + \exp(\boldsymbol{w}^T \mathbf{x}_i)\right) - \exp(\boldsymbol{w}^T \mathbf{x}_i) \frac{\partial 1+\exp(\boldsymbol{w}^T \mathbf{x}_i)}{\partial \boldsymbol{w}}}{(1 + \exp(\boldsymbol{w}^T \mathbf{x}_i))^2} \\
&= \frac{\mathbf{x}_i \exp(\boldsymbol{w}_k^T \mathbf{x}_i) \left(1 + \exp(\boldsymbol{w}^T \mathbf{x}_i)\right) - \exp(\boldsymbol{w}^T \mathbf{x}_i) \mathbf{x}_i \exp(\boldsymbol{w}_\ell^T \mathbf{x}_i)}{(1 + \exp(\boldsymbol{w}^T \mathbf{x}_i))^2} \\
&= \frac{\exp(\boldsymbol{w}^T \mathbf{x}_i)}{1 + \exp(\boldsymbol{w}^T \mathbf{x}_i)} \mathbf{x}_i - \frac{\exp(\boldsymbol{w}^T \mathbf{x}_i) \exp(\boldsymbol{w}^T \mathbf{x}_i)}{(1 + \exp(\boldsymbol{w}^T \mathbf{x}_i))^2} \mathbf{x}_i \\
&= \pi(\mathbf{x}_i; \boldsymbol{w}) \mathbf{x}_i - \pi(\mathbf{x}_i; \boldsymbol{w}) \pi(\mathbf{x}_i; \boldsymbol{w}) \mathbf{x}_i \\
&= \pi_k(\mathbf{x}_i; \boldsymbol{w}) \left(1 - \pi(\mathbf{x}_i; \boldsymbol{w})\right) \mathbf{x}_i
\end{aligned} \tag{18}$$

The gradient (17) is therefore given by

$$\begin{aligned}
\frac{\partial \mathcal{L}(\boldsymbol{w})}{\partial \boldsymbol{w}} &= \sum_{i=1}^{n} y_i \frac{\pi_k(\mathbf{x}_i; \boldsymbol{w}) \left(1 - \pi(\mathbf{x}_i; \boldsymbol{w})\right) \mathbf{x}_i}{\pi(\mathbf{x}_i; \boldsymbol{w})} - (1 - y_i) \frac{\pi_k(\mathbf{x}_i; \boldsymbol{w}) \left(1 - \pi(\mathbf{x}_i; \boldsymbol{w})\right) \mathbf{x}_i}{(1 - \pi(\mathbf{x}_i; \boldsymbol{w}))} \\
&= \sum_{i=1}^{n} y_i \left(1 - \pi(\mathbf{x}_i; \boldsymbol{w})\right) \mathbf{x}_i - (1 - y_i) \pi_k(\mathbf{x}_i; \boldsymbol{w}) \mathbf{x}_i \\
&= \sum_{i=1}^{n} \left(y_i \left(1 - \pi(\mathbf{x}_i; \boldsymbol{w})\right) - (1 - y_i) \pi_k(\mathbf{x}_i; \boldsymbol{w})\right) \mathbf{x}_i \\
&= \sum_{i=1}^{n} \left(y_i - \pi(\mathbf{x}_i; \boldsymbol{w})\right) \mathbf{x}_i
\end{aligned} \tag{19}$$

which can be formulated in a matrix form as

$$\frac{\partial \mathcal{L}(\boldsymbol{w})}{\partial \boldsymbol{w}} = \mathbf{X}^T (\mathbf{y} - \mathbf{p})$$

where $\mathbf{X}$ is the $n \times (d+1)$ matrix whose rows are the input vectors $\mathbf{x}_i$, $\mathbf{y}$ is the $n \times 1$ column vector whose elements are the indicator variables y_i

$$\mathbf{y} = (y_1, \ldots, y_n)^T$$

and $\mathbf{p}$ is the $n \times 1$ column vector of logistic probabilities corresponding to the ith input

$$\mathbf{p} = (\pi(\mathbf{x}_1; \boldsymbol{w}), \ldots, \pi(\mathbf{x}_n; \boldsymbol{w}))^T.$$

The hessian (matrix of second partial derivatives) is then calculated as

$$\begin{aligned}
\frac{\partial^2 \mathcal{L}(\boldsymbol{w})}{\partial \boldsymbol{w} \partial \boldsymbol{w}^T} &= \frac{\partial \sum_{i=1}^{n} \left(y_i - \pi(\mathbf{x}_i; \boldsymbol{w})\right) \mathbf{x}_i}{\partial \boldsymbol{w}^T} \\
&= -\sum_{i=1}^{n} \mathbf{x}_i \frac{\partial \pi(\mathbf{x}_i; \boldsymbol{w})}{\partial \boldsymbol{w}^T} \\
&= -\sum_{i=1}^{n} \mathbf{x}_i \pi_k(\mathbf{x}_i; \boldsymbol{w}) \left(1 - \pi(\mathbf{x}_i; \boldsymbol{w})\right) \mathbf{x}_i^T \\
&= -\sum_{i=1}^{n} \pi_k(\mathbf{x}_i; \boldsymbol{w}) \left(1 - \pi(\mathbf{x}_i; \boldsymbol{w})\right) \mathbf{x}_i \mathbf{x}_i^T
\end{aligned} \tag{20}$$

which can be formulated in a matrix form as

$$\frac{\partial^2 \mathcal{L}(\boldsymbol{w})}{\partial \boldsymbol{w} \partial \boldsymbol{w}^T} = -\mathbf{X}^T \mathbf{W} \mathbf{X}$$

where $\mathbf{W}$ is the $n \times n$ diagonal matrix whose diagonal elements are $\pi(\mathbf{x}_i; \boldsymbol{w})\,(1 - \pi(\mathbf{x}_i; \boldsymbol{w}))$ for $i = 1, \ldots, n$. The NR algorithm (16) in this case can therefore be reformulated from the Equations (20) and (21) as

$$\begin{aligned}
\boldsymbol{w}^{(t+1)} &= \boldsymbol{w}^{(t)} + (\mathbf{X}^T \mathbf{W}^{(t)} \mathbf{X})^{-1} \mathbf{X}^T (\mathbf{y} - \mathbf{p}^{(t)}) \\
&= (\mathbf{X}^T \mathbf{W}^{(t)} \mathbf{X})^{-1} \left[\mathbf{X}^T \mathbf{W}^{(t)} \mathbf{X} \boldsymbol{w}^{(t)} + \mathbf{X}^T (\mathbf{y} - \mathbf{p}^{(t)}) \right] \\
&= (\mathbf{X}^T \mathbf{W}^{(t)} \mathbf{X})^{-1} \mathbf{X}^T \left[\mathbf{W}^{(t)} \mathbf{X} \boldsymbol{w}^{(t)} + (\mathbf{y} - \mathbf{p}^{(t)}) \right] \\
&= (\mathbf{X}^T \mathbf{W}^{(t)} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W}^{(t)} \mathbf{y}^*
\end{aligned}$$

where $\mathbf{y}^* = \mathbf{X}\boldsymbol{w}^{(t)} + (\mathbf{W}^{(t)})^{-1}(\mathbf{y} - \mathbf{p}^{(t)})$ which yields the Iteratively Reweighted Least Squares (IRLS) algorithm.