1

We consider the class variable $Y \in \{0, 1\}$ with a Bernoulli prior. We assume the data conditional probability density (or posterior) $X|Y$ to be distributed with a distinct gaussian mean m-vector μ_y for the different classes but a shared covariance matrix Σ. We suppose $\mathbf{x}_i$ to be a m-vector where $i = 1, \dots, n$ (the dataset $\mathcal{D}$ has m features and n examples).

$$Y \sim \text{Bernoulli}(\pi)$$

$$p(\mathbf{x}_i \mid y_i; \theta) \sim \mathcal{N}(\mathbf{x}_i | \mu_{y_i}, \Sigma)$$

Let θ be the parameter vector

$$\theta^T = (\mu_0, \mu_1, \Sigma, \pi)$$

Then the generative approach to solve for θ is to model the joint probability

$$p(\mathbf{x}_i, y_i; \theta) = p(y_i|\pi)p(\mathbf{x}_i|y_i; \theta)$$

And then do MLE for the log-likelihood which is proportional the joint $\log p(\theta \mid \mathcal{D}) \propto \log p(\mathcal{D}, \theta)$. By assuming all data points are independent we get ¡!– $\log p(\theta \mid \mathcal{D})$ which is proportional up to a constant to the marginal of the log of the joint over the dataset. –¿

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log p(\mathbf{x}_i, y_i; \theta)$$

From our assumptions

$$p(\mathbf{x}_i, y_i; \theta) = \frac{\pi^{y_i}(1-\pi)^{1-y_i}}{\sqrt{(2\pi)^m \det \Sigma}} \exp\left\{ -\frac{1}{2}(\mathbf{x}_i - \mu_{y_i})^T \Sigma^{-1} (\mathbf{x}_i - \mu_{y_i}) \right\}$$

Such that

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} y_i \log \pi + (1 - y_i) \log(1 - \pi) - \frac{m}{2} \log(2\pi) \tag{1.1}$$

$$+ \frac{1}{2} \log \det \Sigma^{-1} - \frac{1}{2}(\mathbf{x}_i - \mu_{y_i})^T \Sigma^{-1} (\mathbf{x}_i - \mu_{y_i}) \tag{1.2}$$

Now we take the derivative of the RHS with respect to the individual parameters and set them to zero to get the MLE estimation

$$\frac{\partial}{\partial \pi}(\cdot) = \sum_{i=1}^{n} \frac{y_i}{\pi} - \frac{1 - y_i}{1 - \pi}$$

Setting this derivative to zero,

$$\frac{\pi^*}{1 - \pi^*} = \frac{\sum_{i=1}^n y_i}{\sum_{i=1}^n 1 - y_i} = \frac{\sum_{i=1}^n y_i}{n - \sum_{i=1}^n y_i}$$

$$\implies \boxed{\hat{\pi}_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} y_i}$$

We verify that this is a maximum at the extremum,

$$\frac{\partial^2}{\partial \pi^2}(\cdot) = -\frac{1}{\pi^2} \sum_{i=1}^{n} y_i - \frac{n - \sum_{i=1}^n y_i}{(1 - \pi)^2}$$

$$\frac{\partial^2}{\partial \pi^2}(\cdot)\bigg|_{\pi = \pi^*} = -n^2 \left(\frac{\sum_{i=1}^n y_i}{(\sum_{i=1}^n y_i)^2} + \frac{n - \sum_{i=1}^n y_i}{(n - \sum_{i=1}^n y_i)^2} \right) < 0$$

for any dataset. We do not simplify in case a term is singular. Therefore $\hat{\pi}_{\text{MLE}}$ maximizes the likelihood.

We solve for the class mean

$$\frac{\partial}{\partial \mu_{y_j}}(\cdot) = -\frac{1}{2} \sum_{i=1}^{n} \frac{\partial}{\partial \mu_{y_j}} (\mathbf{x}_i - \mu_{y_i})^T \Sigma^{-1} (\mathbf{x}_i - \mu_{y_i})$$

Here, we notice that the function to the right of the gradient is a function $g : \mathbb{R}^m \to \mathbb{R}$. Therefore we expect the derivative to be a matrix $dg_{x_0} \in \mathbb{R}^{m\times 1} = \mathbb{R}^m$ which is also a vector. Outside of that, we can use the fact that the quadratic form is equal to its transpose to prove that

$$\partial_{\mathbf{x}} \mathbf{x}^T A\mathbf{x} = \mathbf{x}^T(A + A^T) = (A + A^T)\mathbf{x}$$

By composing the function $f(\mathbf{x}_i) = (\mathbf{x}_i - \mu_{y_i})$ inside the identity above

$$\frac{\partial}{\partial \mu_{y_j}}(\cdot) = \frac{1}{2}\sum_{i=1}^{n}(\Sigma^{-1} + (\Sigma^{-1})^T)(\mathbf{x}_i - \mu_{y_j})\delta_{y_i y_j}$$

But, since Σ is symmetric, its inverse must also be symmetric and we get

$$\frac{\partial}{\partial \mu_{y_j}}(\cdot) = \sum_{i=1}^{n}\Sigma^{-1}(\mathbf{x}_i - \mu_{y_j})\delta_{y_i y_j}$$

We have used the Kronecker delta symbol to replace the gradient $\frac{\partial \mu_{y_i}}{\partial \mu_{y_j}}$ since it will be 1 when $y_i = y_j$ and zero otherwise. Setting this to zero, we get

$$\sum_{i=1}^{n}\Sigma\Sigma^{-1}(\mathbf{x}_i - \mu^*_{y_i})\delta_{y_i y_j} = \sum_{i=1}^{n} I_{m\times m}(\mathbf{x}_i - \mu^*_{y_j})\delta_{y_i y_j} = 0$$

To simpify further, we can use the labels $y_i = \delta_{i1}$ to replace the kronecker delta when $y_j = 1$ and $1 - y_i = \delta_{i0}$ when $y_j = 0$. From this simplification, we get

$$\boxed{\hat{\mu_1}_{\text{MLE}} = \frac{1}{\sum_{i=1}^{n} y_i}\sum_{i=1}^{n} y_i \mathbf{x}_i}$$

and

$$\boxed{\hat{\mu_0}_{\text{MLE}} = \frac{1}{n - \sum_{i=1}^{n} y_i}\sum_{i=1}^{n}(1 - y_i)\mathbf{x}_i}$$

The second order gradients are negative sum of Kronecker deltas

$$\frac{\partial^2}{\partial \mu^2_{y_j}}(\cdot) = -\sum_{i=1}^{n}\delta^2_{y_j y_i}$$

and are thus negative for dataset that have at least 1 representative of each class. The MLE estimator maximize the likelihood in that situation.

Finally, we look for the covariance estimator. Here, we optimize w.r.t the matrix inverse of the covariance. Let $\Lambda = \Sigma^{-1}$, then

$$\frac{\partial}{\partial \Lambda}(\cdot) = \frac{1}{2}\sum_{i=1}^{n}\frac{\partial}{\partial \Lambda}\log\det\Lambda - \frac{\partial}{\partial \Lambda}(\mathbf{x}_i - \mu_{y_i})^T\Lambda(\mathbf{x}_i - \mu_{y_i}) \tag{1.3}$$

We know that $\frac{\partial}{\partial \Lambda}\log\det\Lambda = \Lambda^{-1} = \Sigma$. Also, since the quadratic form (second term) is a function o fthe form $f : \mathbf{S}^m \to \mathbb{R}$, then its derivative must be a symmetric matrix $\in \mathbf{S}^m$. We find that the result must be the exterior product

$$\frac{\partial}{\partial \Lambda}(\mathbf{x}_i - \mu_{y_i})^T\Lambda(\mathbf{x}_i - \mu_{y_i}) = (\mathbf{x}_i - \mu_{y_i})(\mathbf{x}_i - \mu_{y_i})^T$$

Setting this derivative to zero, and choosing the maximum for th e class mean also,

$$\frac{\partial}{\partial \Lambda}(\cdot)\Big|_{\theta=\theta^*} = 0 \tag{1.4}$$

$$\implies \sum_{i=1}^{n}\Sigma^* - (\mathbf{x}_i - \mu^*_{y_i})(\mathbf{x}_i - \mu^*_{y_i})^T = 0 \tag{1.5}$$

$$\implies \Sigma^* = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \mu^*_{y_i})(\mathbf{x}_i - \mu^*_{y_i})^T \tag{1.6}$$

Therefore,

$$\hat{\Sigma}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\mu_{y_i}}_{\text{ MLE}})(\mathbf{x}_i - \hat{\mu_{y_i}}_{\text{ MLE}})^T$$

We can rewrite this conveniently with the class labels

$$\boxed{\hat{\Sigma}_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - (1-y_i)\hat{\mu_0}_{\text{ MLE}} - y_i\hat{\mu_1}_{\text{ MLE}})(\mathbf{x}_i - (1-y_i)\hat{\mu_0}_{\text{ MLE}} - y_i\hat{\mu_1}_{\text{ MLE}})^T}$$

2

From Baye's theorem, this conditional can be rewritten as

$$p(y=1 \mid \mathbf{x}, \theta) = \frac{p(\mathbf{x}|y=1,\theta)p(y=1 \mid \pi)}{p(\mathbf{x} \mid \theta)}$$

The denominator is the marginal over the class labels, which we write as

$$p(\mathbf{x} \mid \theta) = p(\mathbf{x} \mid y=1,\theta)p(y=1 \mid \pi) + p(\mathbf{x} \mid y=0,\theta)p(y=0 \mid \pi)$$

Using this definition and rewritting the conditional, we get

$$p(y=1 \mid \mathbf{x}, \theta) = \frac{1}{1 + \dfrac{p(\mathbf{x} \mid y=0,\theta)p(y=0 \mid \pi)}{p(\mathbf{x} \mid y=1,\theta)p(y=1 \mid \pi)}}$$

Using the definition of a conditional, we can rewrite this as a ratio of joint distributions

$$p(y=1 \mid \mathbf{x}, \theta) = \frac{1}{1 + \dfrac{p(\mathbf{x}, y=0;\theta)}{p(\mathbf{x}, y=1;\theta)}}$$

Using our definitions from the previous problem, and simplifying the common factor, we get

$$p(y=1 \mid \mathbf{x}, \theta) = \frac{1}{1 + \dfrac{1-\pi}{\pi}\exp\left\{-\frac{1}{2}(\mathbf{x}-\mu_0)^T\Sigma^{-1}(\mathbf{x}-\mu_0) + \frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma^{-1}(\mathbf{x}-\mu_1)\right\}}$$

We send the factor in the exponent of the exponential and realize that this is the form of a sigmoid. This exponent can be simplified:

$$-\arg(\sigma) = -\log\frac{\pi}{1-\pi} - \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 + \frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mu_0 + \frac{1}{2}\mu_0^T\Sigma^{-1}\mathbf{x} \tag{2.1}$$

$$+ \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 - \frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mu_1 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mathbf{x} \tag{2.2}$$

Where we simplified the quadratic term. Using the fact that the transpose of a scalar is equal to itself, we can rearrenge the terms and exploit the following quadratic identity $(x-y)(x+y) = x^2 - y^2$

$$-\arg(\sigma) = -\log\frac{\pi}{1-\pi} - \frac{1}{2}(\mu_0-\mu_1)^T\Sigma^{-1}(\mu_0+\mu_1) - (\mu_1-\mu_0)^T\Sigma^{-1}\mathbf{x}$$

Therefore

$$p(y=1 \mid \mathbf{x}, \theta) = \sigma(\beta^T\mathbf{x} + \gamma)$$

where

$$\gamma \equiv \log\frac{\pi}{1-\pi} + \frac{1}{2}(\mu_0-\mu_1)^T\Sigma^{-1}(\mu_1+\mu_0)$$

and

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

The sigmoid is a logistic function of $\mathbf{x}$ and therefore the conditional has the correct form for logistic regression.