Mid-term review solutions
ECE 271A
Electrical and Computer Engineering
University of California San Diego

Nuno Vasconcelos — Fall 2008

1. a) The posterior is given by

$$\begin{aligned}
P_{Y|\mathbf{X}}(1|\mathbf{x}) &= \frac{P_{\mathbf{X}|Y}(\mathbf{x}|1)P_Y(1)}{P_{\mathbf{X}|Y}(\mathbf{x}|1)P_Y(1)+P_{\mathbf{X}|Y}(\mathbf{x}|0)P_Y(0)} \\
&= \frac{P_{\mathbf{X}|Y}(\mathbf{x}|1)}{P_{\mathbf{X}|Y}(\mathbf{x}|1)+P_{\mathbf{X}|Y}(\mathbf{x}|0)} \\
&= \frac{1}{1+\frac{P_{\mathbf{X}|Y}(\mathbf{x}|0)}{P_{\mathbf{X}|Y}(\mathbf{x}|1)}} \\
&= \frac{1}{1+\frac{e^{-\frac{1}{2}(\mathbf{x}-\mu_0)^T\Sigma^{-1}(\mathbf{x}-\mu_0)}}{e^{-\frac{1}{2}(\mathbf{x}-\mu_1)^T\Sigma^{-1}(\mathbf{x}-\mu_1)}}} \\
&= \frac{1}{1+\frac{e^{\mu_0^T\Sigma^{-1}\mathbf{x}-\frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0}}{e^{\mu_1^T\Sigma^{-1}\mathbf{x}-\frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1}}} \\
&= \frac{1}{1+e^{(\mu_0-\mu_1)^T\Sigma^{-1}\mathbf{x}-\frac{1}{2}(\mu_0^T\Sigma^{-1}\mu_0-\mu_1^T\Sigma^{-1}\mu_1)}} \\
&= \frac{1}{1+e^{-\mathbf{w}^T\mathbf{t}}}
\end{aligned}$$

with

$$\mathbf{w} = \left[\begin{array}{c} \Sigma^{-1}(\mu_1-\mu_0) \\ \frac{\mu_0^T\Sigma^{-1}\mu_0-\mu_1^T\Sigma^{-1}\mu_1}{2} \end{array}\right]. \qquad (1)$$

b) We start by noting that

$$\begin{aligned}
P_{Y|\mathbf{X}}(y_i|\mathbf{x}_i) &= \begin{cases} \frac{1}{1+e^{-\mathbf{w}^T\mathbf{t}_i}}, & y_i=1 \\ 1-\frac{1}{1+e^{-\mathbf{w}^T\mathbf{t}_i}}, & y_i=0 \end{cases} \\
&= \begin{cases} \frac{1}{1+e^{-\mathbf{w}^T\mathbf{t}_i}}, & y_i=1 \\ \frac{e^{-\mathbf{w}^T\mathbf{t}_i}}{1+e^{-\mathbf{w}^T\mathbf{t}_i}}, & y_i=0 \end{cases}
\end{aligned}$$

which can be written as

$$P_{Y|\mathbf{X}}(y_i|\mathbf{x}_i) = \left(\frac{1}{1+e^{-\mathbf{w}^T\mathbf{t}_i}}\right)^{y_i}\left(\frac{e^{-\mathbf{w}^T\mathbf{t}_i}}{1+e^{-\mathbf{w}^T\mathbf{t}_i}}\right)^{1-y_i}.$$

The fact that

$$P_{\mathbf{Y}|\mathbf{X}}(\mathcal{D}_y|\mathcal{D}_x) = \prod_{i=1}^{n} P_{Y|\mathbf{X}}(y_i|\mathbf{x}_i) \qquad (2)$$

is a straightforward consequence of the fact that the sample is iid.

c) To compute $\mathbf{w}^\star$ we take the usual route of considering the log of the function to maximize. This leads to

$$\mathbf{w}^\star \quad = \quad \arg\max_{\mathbf{w}} l(\mathbf{w})$$

with

$$\begin{aligned}
l(\mathbf{w}) &= \sum_{i=1}^{n} \log P_{Y|\mathbf{X}}(y_i|\mathbf{x}_i) \\
&= \sum_{i=1}^{n} y_i \log\left(\frac{1}{1+e^{-\mathbf{w}^T\mathbf{t}_i}}\right) + (1-y_i)\log\left(\frac{e^{-\mathbf{w}^T\mathbf{t}_i}}{1+e^{-\mathbf{w}^T\mathbf{t}_i}}\right) \\
&= \sum_{i=1}^{n} -y_i \log\left(1+e^{-\mathbf{w}^T\mathbf{t}_i}\right) - (1-y_i)\log\left(1+e^{-\mathbf{w}^T\mathbf{t}_i}\right) - (1-y_i)\mathbf{w}^T\mathbf{t}_i \\
&= -\sum_{i=1}^{n} \log\left(1+e^{-\mathbf{w}^T\mathbf{t}_i}\right) + (1-y_i)\mathbf{w}^T\mathbf{t}_i.
\end{aligned}$$

As usual, we compute the gradient of $l(\mathbf{w})$

$$\begin{aligned}
\nabla_{\mathbf{w}} l(\mathbf{w}) &= -\sum_{i=1}^{n} \frac{e^{-\mathbf{w}^T\mathbf{t}_i}}{1+e^{-\mathbf{w}^T\mathbf{t}_i}}(-\mathbf{t}_i) + (1-y_i)\mathbf{t}_i \\
&= \sum_{i=1}^{n} y_i\mathbf{t}_i - \sum_{i=1}^{n} \frac{1}{1+e^{-\mathbf{w}^T\mathbf{t}_i}}\mathbf{t}_i
\end{aligned}$$

and set it to zero to obtain

$$\sum_{i=1}^{n} y_i\mathbf{t}_i = \sum_{i=1}^{n} \frac{1}{1+e^{-\mathbf{w}^{\star T}\mathbf{t}_i}}\mathbf{t}_i.$$

Finally we note that

$$\nabla^2_{\mathbf{w}} l(\mathbf{w}) \quad = \quad -\sum_{i=1}^{n} \frac{e^{-\mathbf{w}^T\mathbf{t}_i}}{(1+e^{-\mathbf{w}^T\mathbf{t}_i})^2}\mathbf{t}_i\mathbf{t}_i^T$$

which is a sum of n negative semidefinite matrices and therefore negative semidefinite. Hence $\mathbf{w}^\star$ is a maximum.

d) Since this is a binary equal covariance Gaussian classification problem, if this approach is to work at all the solution should be the hyperplane associated with the BDR of this problem. Hence, we suspect that $\mathbf{w}^\star$ consists of the normal and bias of that plane. Checking in DHS, we verify that those are exactly given by (1), i.e.

$$\mathbf{w}^\star = \begin{bmatrix} \Sigma^{-1}(\mu_1 - \mu_0) \\ \frac{\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1}{2} \end{bmatrix}. \tag{3}$$

Note that in this case, from **a)**,

$$\frac{1}{1+e^{-\mathbf{w}^{\star T}\mathbf{t}}} = P_{Y|\mathbf{X}}(1|\mathbf{x})$$

and

$$\begin{aligned}
E_{\mathbf{X}}\left[\frac{1}{1+e^{-\mathbf{w}^{\star T}\mathbf{T}}}\mathbf{T}\right] &= E_{\mathbf{X}}\left[P_{Y|\mathbf{X}}(1|\mathbf{x})\mathbf{T}\right] \\
&= \int P_{Y|\mathbf{X}}(1|\mathbf{x})\mathbf{t}P_{\mathbf{X}}(\mathbf{x})d\mathbf{x} \\
&= \frac{1}{2}\int P_{\mathbf{X}|Y}(\mathbf{x}|1)\left[\begin{array}{c}\mathbf{x}\\1\end{array}\right]d\mathbf{x} \\
&= \frac{1}{2}\left[\begin{array}{c}\mu_1\\1\end{array}\right].
\end{aligned}$$

On the other hand,

$$\begin{aligned}
E_{Y,\mathbf{X}}[Y\mathbf{T}] &= E_{\mathbf{Y}}\{E_{\mathbf{X}|Y}[Y\mathbf{T}|Y]\} \\
&= \frac{1}{2}\{E_{\mathbf{X}|Y}[Y\mathbf{T}|Y=0]+E_{\mathbf{X}|Y}[Y\mathbf{T}|Y=1]\} \\
&= \frac{1}{2}E_{\mathbf{X}|Y}[\mathbf{T}|Y=1] \\
&= \frac{1}{2}\left[\begin{array}{c}\mu_1\\1\end{array}\right].
\end{aligned}$$

Hence,

$$E_{Y,\mathbf{X}}[Y\mathbf{T}] = E_{\mathbf{X}}\left[\frac{1}{1+e^{-\mathbf{w}^{\star T}\mathbf{T}}}\mathbf{T}\right].$$

2.

a) The BDR is to choose $Y = 0$ if

$$\begin{bmatrix} 1 & 0 \end{bmatrix} \mathbf{x} = x_1 > 0$$

where $\mathbf{x} = (x_1,\ x_2)^T$, and choose $Y = 1$ for "<".

b) The marginal distributions for the features are

1.

$$P_{X_1|Y}(x_1|0) = \mathcal{N}(1, \sigma^2),$$
$$P_{X_1|Y}(x_1|1) = \mathcal{N}(-1, \sigma^2),$$
$$P_{X_2|Y}(x_2|0) = P_{X_2|Y}(x_2|1) = \mathcal{N}(0, \sigma^2)$$

2. the plots are omitted.

3. feature 1 is more discriminant.

c) 1. The transformation matrix Γ is a clockwise rotation transformation of $\pi/4$,

$$\Gamma = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}$$

2. If the prior probability of class 0 was increased after transformation, then the decision boundary of BDR would still have the same normal as before, i.e., $\mathbf{w} = (1/\sqrt{2}, -1/\sqrt{2})^T$, but move toward the mean of class 1.

d) 1. The equations for point assignment is determined by BDR, i.e., assign a point to class $Y = 0$, if

$$\begin{aligned} \log P_{X|Y}(x|0) + \log P_Y(0) &> \log P_{X|Y}(x|1) + \log P_Y(1) \\ -\frac{1}{2}||\mathbf{x} - \mu_0^k||^2 + \log P_Y^k(0) &> -\frac{1}{2}||\mathbf{x} - \mu_1^k||^2 + \log P_Y^k(1), \end{aligned}$$

and assign a point to class $Y = 1$ for "<".

2. The parameter update equations are,

$$\begin{aligned} \mu_i^{k+1} &= \frac{1}{n}\sum_{l=1}^{n} \mathbf{x}_{i,l}^k \\ P_Y^{k+1}(i) &= \frac{|D_i^k|}{|D_0^k| + |D_1^k|}, \end{aligned}$$

where $|D|$ is the size of the point set D.

3. The assignments of the points of the three iterations are:
Iteration 1: class 0: $\{(-1.2, 1.1)^T\}$; class 1: other three points. Not successfully separated the two classes.
Iteration 2: class 0: $\{(-1, 0.9)^T, (-1.2, 1.1)^T\}$; class 1: $\{(1, 0.8)^T, (1.1, 1)^T\}$. Successfully separated the two classes.
Iteration 3: the same assignment as the iteration 2 and the two classes are, again, successfully separated.