Mid-term review solutions
ECE 175
Electrical and Computer Engineering
University of California San Diego

Nuno Vasconcelos

1.

a) For this problem, the Bayesian decision rule is to guess *heads* when

$$\begin{aligned} P_{S|R}(heads|heads) &> P_{S|R}(tails|heads) && (1) \\ P_{R|S}(heads|heads)P_S(heads) &> P_{R|S}(heads|tails)P_S(tails) && (2) \\ (1-\theta_1)\alpha &> \theta_2(1-\alpha) && (3) \\ \alpha &> \frac{\theta_2}{1-\theta_1+\theta_2} && (4) \end{aligned}$$

and *tails* when

$$\alpha < \frac{\theta_2}{1-\theta_1+\theta_2}. \quad (5)$$

When

$$\alpha = \frac{\theta_2}{1-\theta_1+\theta_2} \quad (6)$$

any guess is equally good.

b)When $\theta_1 = \theta_2 = \theta$ the minimum probability of error decision is to declare *heads* if

$$\alpha > \theta \quad (7)$$

and *tails* otherwise. This means that you should only believe your friend's report if your prior for *heads* is greater than the probability that he lies. To see that this makes a lot of sense let's look at a few different scenarios.

- If your friend is a pathological lier ($\theta = 1$), then you know for sure that the answer is not *heads* and you should always say tails. This is the decision that (7) advises you to take.
- If he never lies ($\theta = 0$) you know that the answer is *heads*. Once again this is the decision that (7) advises you to take.
- If both $\alpha = 0$ and $\theta = 0$ we have a contradiction, i.e. you know for sure that the result of the toss is always *tails* but this person that never lies is telling you that it is *heads*. In this case Bayes just gives up and says "either way is fine". This is a sensible strategy, there is something wrong with the models, you probably need to learn something more about the problem.
- If your friend is completely random, $\theta = 1/2$, (7) tells you to go with your prior and ignore him. If you believe that that the coin is more likely to land on *heads* say *heads* otherwise say *tails*. Bayes has no problem with ignoring the observations, whenever these are completely uninformative.

- When you do not have prior reason to believe that one of the outcomes is more likely than the other, i.e. if you assume a fair coin ($\alpha = 1/2$), (7) advises you to reject the report whenever you think that your friend is more of a lier ($\theta > 1/2$) and to accept it when you believe that he is more on the honest side ($\theta < 1/2$). Once again this makes sense.

- In general, the optimal decision rule is to "modulate" this decision by your prior belief on the outcome of the toss: say *heads* if your prior belief that the outcome was really *heads* is larger than the probability that your friend is lying.

c) Denoting by R_i the i^{th} report and assuming that the sequence of reports $\{r_1, \ldots, r_n\}$ has $n_h heads$ and $n - n_h$ tails, the BDR is now to say *heads* if

$$
\begin{aligned}
P_{S|R_1,\ldots,R_n}(heads|r_1,\ldots,r_n) &> P_{S|R_1,\ldots,R_n}(tails|r_1,\ldots,r_n) && (8)\\
P_{R_1,\ldots,R_n|S}(r_1,\ldots,r_n|heads)P_S(heads) &> P_{R_1,\ldots,R_n|S}(r_1,\ldots,r_n|tails)P_S(tails) && (9)\\
(1-\theta_1)^{n_h}\theta_1^{n-n_h}\alpha &> \theta_2^{n_h}(1-\theta_2)^{n-n_h}(1-\alpha) && (10)\\
\alpha &> \frac{\theta_2^{n_h}(1-\theta_2)^{n-n_h}}{(1-\theta_1)^{n_h}\theta_1^{n-n_h} + \theta_2^{n_h}(1-\theta_2)^{n-n_h}} && (11)\\
\alpha &> \frac{1}{1 + (\frac{1-\theta_1}{\theta_2})^{n_h}(\frac{\theta_1}{1-\theta_2})^{n-n_h}} && (12)\\
& && (13)
\end{aligned}
$$

and *tails* otherwise.

d) When $\theta_1 = \theta_2 = \theta$ and the report sequence is all *heads* ($n_h = n$), the BDR becomes to declare *heads* if

$$\alpha > \frac{1}{1 + (\frac{1-\theta}{\theta})^n} \qquad (14)$$

and *tails* otherwise. As n becomes larger, i.e. $n \to \infty$, we have three situations.

- Your friend is more of a lier, $\theta > 1/2$. In this case, $(1 - \theta/\theta)^n \to 0$ and the decision rule becomes $\alpha > 1$. That is, you should always reject his report.

- Your friend is more of a honest person, $\theta < 1/2$. In this case, $(1 - \theta/\theta)^n \to \infty$ and the decision rule becomes $\alpha > 0$. That is, you should always accept his report.

- Your friend is really just random, $\theta = 1/2$. In this case, the decision rule becomes $\alpha > 1/2$ and you should go with your prior.

Once again this makes a lot of sense. Now you have a lot of observations so you are much more confident on the data and need to rely a lot less on your prior. It also takes a lot less work to figure out what you should do, since you do not have to make detailed probability comparisons. Because your friend seems to be so certain of the outcome (he always says *heads*), you either: 1) not trust him (θ somewhere in between 1/2 and 1) therefore believe that he is just trying to fool you and reject what he says, or 2) trust him (θ somewhere in between 0 and 1/2) and accept his report. It is only in the case that he is completely unpredictable that the prior becomes important. This looks like a really good strategy, and sounds a lot like the way people think. As you can see in this example, the optimal Bayesian decision can be something as qualitative as: *if you trust accept, if you doubt reject, otherwise ignore*.

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, as knowing the value of feature 2 does not provide any information as to which class the point could have come from.

c) **1.** As the decision boundary now coincides with the line $x_1 = x_2$, the two class means must lie on a line that is normal to the decision boundary, i.e. on $x_1 = -x_2$. It is easy to see then that 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.

3. Noting that

$$\begin{aligned} ||\mathbf{Tx} - \mathbf{Ty}||^2 &= (\mathbf{x} - \mathbf{y})^T \mathbf{T}^T \mathbf{T} (\mathbf{x} - \mathbf{y}) \\ &= (\mathbf{x} - \mathbf{y})^T (\mathbf{x} - \mathbf{y}) \\ &= ||\mathbf{x} - \mathbf{y}||^2 \end{aligned}$$

we see that the distance is still the Euclidean distance. This is due to the fact that $\mathbf{T}$ is a rotation, and rotations do not change distances between points.

3. **a)** The posterior is given by

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

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

with

$$
\mathbf{w} = \left[\begin{array}{c} \Sigma^{-1}(\mu_1-\mu_0) \\ \frac{\mu_0^T\Sigma^{-1}\mu_0-\mu_1^T\Sigma^{-1}\mu_1}{2} \end{array} \right]. \tag{15}
$$

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

$$
P_{\mathbf{Y}|\mathbf{X}}(\mathcal{D}_y|\mathcal{D}_x) = \prod_{i=1}^{n} P_{Y|\mathbf{X}}(y_i|\mathbf{x}_i) \tag{16}
$$

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