Solutions to Homework Set One

ECE 271A
Electrical and Computer Engineering
University of California San Diego

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*. As we have seen in class, 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 MPE decision 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 MPE decision becomes to declare *heads* if

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

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. Problem 2.2.2. in DHS

a) We need to find the normalizing constant, i.e. the constant that multiplied by $e^{-|x-a_i|/b_i}$ will lead to a function that integrates to 1. Since

$$\begin{aligned}\int_{-\infty}^{\infty} e^{-|x-a_i|/b_i} dx &= \int_{-\infty}^{\infty} e^{-|y|/b_i} dy \ \ (y = x - a_i, dy = dx) && (15)\\ &= \int_{-\infty}^{0} e^{y/b_i} dy + \int_{0}^{\infty} e^{-y/b_i} dy && (16)\\ &= b_i e^{y/b_i}|_{-\infty}^{0} - b_i e^{-y/b_i}|_{0}^{\infty} && (17)\\ &= 2b_i, && (18)\end{aligned}$$

the constant is $1/2b_i$ and

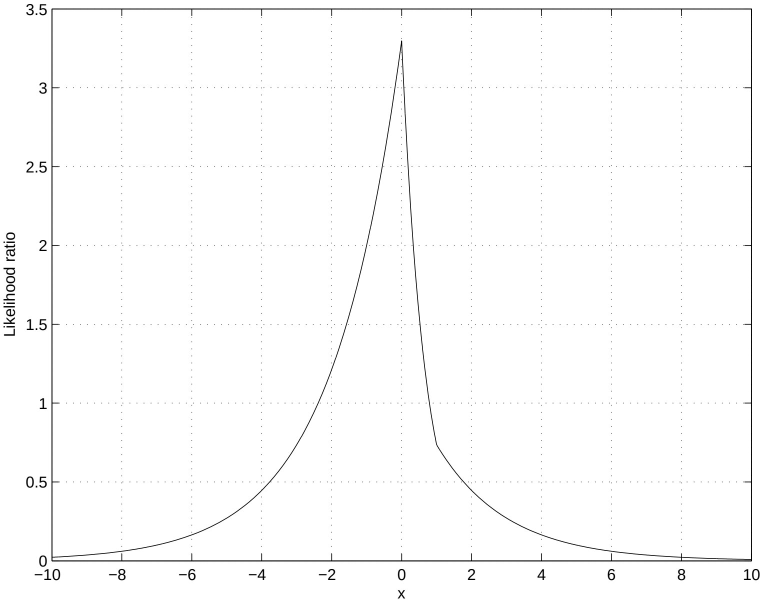
$$P_{X|Y}(x|i) = \frac{1}{2b_i} e^{-|x-a_i|/b_i}. \tag{19}$$

b) This is simply the function

$$\frac{P_{X|Y}(x|1)}{P_{X|Y}(x|2)} = \frac{b_2}{b_1} e^{-|x-a_1|/b_1 + |x-a_2|/b_2} \tag{20}$$

c) When $a_1 = 0, b_1 = 1, a_2 = 1$, and $b_2 = 2$

$$\frac{P_{X|Y}(x|1)}{P_{X|Y}(x|2)} = 2e^{-|x|+|x-1|/2} \tag{21}$$



3. Problem 2.5.23. in DHS

a)

$$P_{\mathbf{X}}(\mathbf{x}_0) = 0.0082$$

b)

$$\mathbf{\Phi} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \mathbf{\Lambda} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 7 \end{bmatrix} \mathbf{A}_w = \mathbf{\Phi}\mathbf{\Lambda}^{-1/2} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & -\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{14}} \\ 0 & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{14}} \end{bmatrix}$$

To find the transformation $\mathbf{T}$ that converts $\mathbf{X} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$ into $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ we start by subtracting the mean from $\mathbf{x}$. This does not change the covariance so, if we define $\mathbf{z} = \mathbf{x} - \mu$, we have $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$. Next, we look for the matrix $\mathbf{T}_1$ such that $\mathbf{y} = \mathbf{T}_1\mathbf{z}$ has identity covariance. Noting that the covariance of y is

$$\begin{aligned} \mathbf{\Sigma}_{\mathbf{y}} &= E[\mathbf{y}\mathbf{y}^T] \\ &= \mathbf{T}_1 E[\mathbf{z}\mathbf{z}^T]\mathbf{T}_1^T \\ &= \mathbf{T}_1\mathbf{\Sigma}\mathbf{T}_1^T \\ &= \mathbf{T}_1\mathbf{\Phi}\mathbf{\Lambda}\mathbf{\Phi}^T\mathbf{T}_1^T \end{aligned}$$

it becomes clear that the transformation $\mathbf{T}_1 = \Lambda^{-1/2}\mathbf{\Phi}^T = \mathbf{A}_w^T$ is what we are looking for. Hence, the transformation $\mathbf{T}$ is

$$\mathbf{y} = \mathbf{A}_w^T(\mathbf{x} - \mu). \tag{22}$$

c)

$$\mathbf{x}_w = \mathbf{A}_w^T(\mathbf{x}_0 - \mu) = (-.5, .4082, -.8018)^T.$$

d)

$$(\mathbf{x}_0 - \mu)^T\Sigma^{-1}(\mathbf{x}_0 - \mu) = \mathbf{x}_w^T\mathbf{x}_w = 1.0595$$

e) From the properties of the Gaussian we know that a linear transformation of a Gaussian random variable is Gaussian. From the linearity of the expected value it follows that, if $\mathbf{y} = \mathbf{T}^T\mathbf{x}$, then $\mu_{\mathbf{y}} = E[\mathbf{y}] = \mathbf{T}^T\mu$. Finally, if $\mathbf{y} = \mathbf{T}^T\mathbf{x}$,

$$\begin{aligned} \mathbf{\Sigma}_{\mathbf{y}} &= E[(\mathbf{y} - \mu_{\mathbf{y}})(\mathbf{y} - \mu_{\mathbf{y}}{}^T)] \qquad (23) \\ &= E[\mathbf{T}^T(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T\mathbf{T}] \qquad (24) \\ &= \mathbf{T}^T\mathbf{\Sigma}\mathbf{T}. \qquad (25) \end{aligned}$$

Hence, $P_Y(\mathbf{y})$ is a Gaussian of mean $\mathbf{T}^T\mu$ and covariance $\mathbf{T}^T\mathbf{\Sigma}\mathbf{T}$. While the Mahalanobis distances are the same, the scaling terms are not($\frac{1}{|\mathbf{T}^T\mathbf{\Sigma}\mathbf{T}|^{\frac{1}{2}}}$ for $P_Y(\mathbf{y})$, and $\frac{1}{|\mathbf{\Sigma}|^{\frac{1}{2}}}$ for $P_X(\mathbf{x})$). The two terms will be equal only when $|\mathbf{T}| = \pm 1$. For a general linear transformation, the probabilities are not the same.

f) We did the first part in **b)** and the second part in **e)**. For the whitening transform, the normalization is not preserved in general. As shown in **e)**, normalization is preserved only if $|\mathbf{T}| = \pm 1$, i.e. if $|\mathbf{A}_w| = \pm 1$, which, in general, need not hold.

4. Problem 2.9.43 in DHS

a) p_{ij} is the probability that the i^{th} feature will be active ($x_i = 1$) given that the world is in state j.

b) We have seen in class that the minimum probability of error decision rule is to pick (note that we use boldface for vectors, standard typeface for scalars)

$$\begin{aligned} k^* &= \arg\max_k P_{Y|\mathbf{X}}(k|\mathbf{x}) && (26) \\ &= \arg\max_k \log P_{\mathbf{X}|Y}(\mathbf{x}|k) + \log P_Y(k). && (27) \end{aligned}$$

From the fact that the features X_i are statistically independent it follows that

$$\begin{aligned} P_{\mathbf{X}|Y}(\mathbf{x}|k) &= \prod_{i=1}^{d} P_{X_i|Y}(x_i|k) && (28) \\ &= \prod_{i=1}^{d} [P_{X_i|Y}(0|k)]^{1-x_i}[P_{X_i|Y}(1|k)]^{x_i} && (29) \end{aligned}$$

where we have used the fact that, for a binary variable X,

$$P_X(x) = \begin{cases} P_X(0), & \text{if}\, x = 0 \\ P_X(1), & \text{if}\, x = 1 \end{cases} = [P_X(0)]^{1-x}[P_X(1)]^x. \qquad (30)$$

Hence

$$\begin{aligned} k^* &= \arg\max_k \sum_{i=1}^{d} \{(1-x_i)\log P_{X_i|Y}(0|k) + x_i \log P_{X_i|Y}(1|k)\} + \log P_Y(k) && (31) \\ &= \arg\max_k \sum_{i=1}^{d} \{(1-x_i)\log(1-p_{ik}) + x_i \log p_{ik}\} + \log P_Y(k) && (32) \\ &= \arg\max_k \sum_{i=1}^{d} x_i \log \frac{p_{ik}}{(1-p_{ik})} + \sum_{i=1}^{d} \log(1-p_{ik}) + \log P_Y(k) && (33) \end{aligned}$$

5. a) A reasonable estimate for the prior probabilities can be obtained by using the number of the training samples for each class. Since

class	samples
cheetah	250
grass	1053

the following estimates are sensible

$$
\begin{aligned}
P_Y(cheetah) &= \frac{250}{250+1053} = 0.1919 \\
P_Y(grass) &= \frac{1053}{250+1053} = 0.8081.
\end{aligned}
$$

b) The histograms for the two classes are shown below. Notice that there is a significant amount of overlap, indicating that the feature that we are using (index of the 2^{nd} largest DCT coefficient) is not very good.

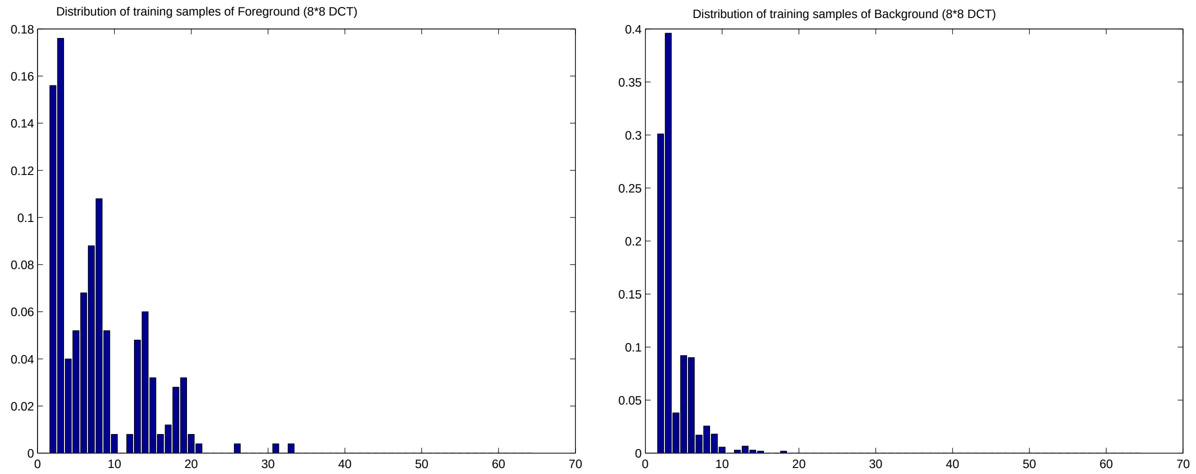
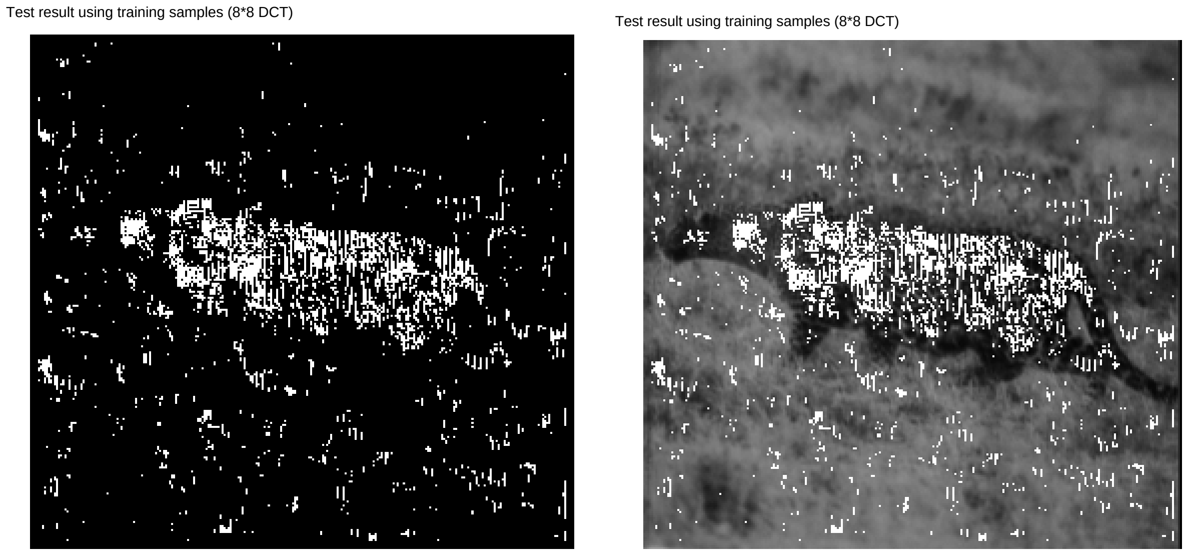


Figure 1: Class conditional histograms. Left: $P_{X|Y}(x|cheetah)$. Right: $P_{X|Y}(x|grass)$.

c) The minimum probability of error segmentation mask is shown below. It is also shown superimposed on the cheetah image. Notice that the segmentation is quite noisy, confirming what one would expect from the histograms above.



Test result using training samples (8*8 DCT)

Test result using training samples (8*8 DCT)

Figure 2: Left: segmentation mask. Right: superimposed on the `cheetah` image.

d)

We first compute the two types of error

- detection rate: $P_{X|Y}(g(x) = cheetah|cheetah) = 0.2520$,
- false alarm rate: $P_{X|Y}(g(x) = cheetah|grass) = 0.0300$

and then combine them into the probability of error

$$\begin{aligned}
P_E &= E_Y[P_{X|Y}(g(x) \neq Y|Y)] = \sum_i P_{X|Y}(g(x) \neq i|i)P_Y(i) \\
&= P_{X|Y}(g(x) = cheetah|grass)P_Y(grass) + P_{X|Y}(g(x) = grass|cheetah)P_Y(cheetah) \\
&= 0.0300 * 0.8081 + (1 - 0.2520) * 0.1919 \\
&= 0.1678
\end{aligned}$$