**Homework 8 Solutions**

1. (10 points) *Probability of error vs MSE (10 points).*

   a. We can easily find the piecewise constant density of $Y$

$$f_Y(y) = \begin{cases} \frac{1}{4} & |y| \leq 1 \\ \frac{1}{8} & 1 < |y| \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

The conditional probabilities of $X$ given $Y$ are

$$\mathrm{P}\{X = +1 | Y = y\} = \begin{cases} 0 & -3 \leq y < -1 \\ \frac{1}{2} & -1 \leq y \leq +1 \\ 1 & +1 < y \leq +3 \end{cases}$$

$$\mathrm{P}\{X = -1 | Y = y\} = \begin{cases} 1 & -3 \leq y < -1 \\ \frac{1}{2} & -1 \leq y \leq +1 \\ 0 & +1 < y \leq +3 \end{cases}$$

Thus the best MSE estimate is

$$g_{\mathrm{MSE}}(Y) = \mathrm{E}(X|Y) = \begin{cases} -1 & -3 \leq Y < -1 \\ \phantom{+}0 & -1 \leq Y \leq +1 \\ +1 & +1 < Y \leq +3 \end{cases}$$

The minimum mean square error is

$$\begin{aligned} \mathrm{E}_Y(\mathrm{Var}(X|Y)) &= \mathrm{E}_Y(\mathrm{E}(X^2|Y) - (\mathrm{E}(X|Y))^2) = \mathrm{E}(1 - g_{\mathrm{MSE}}(Y)^2) \\ &= 1 - \mathrm{E}(g_{\mathrm{MSE}}(Y)^2) = 1 - \int_{-\infty}^{\infty} g(y)^2 f_Y(y)\, dy \\ &= 1 - \left( \int_{-3}^{-1} 1 \cdot \tfrac{1}{8}\, dy + \int_{-1}^{1} 0 \cdot \tfrac{1}{4}\, dy + \int_{+1}^{+3} 1 \cdot \tfrac{1}{8}\, dy \right) \\ &= 1 - \int_{-3}^{-1} \tfrac{1}{8}\, dy - \int_{1}^{3} \tfrac{1}{8}\, dy = 1 - \tfrac{1}{4} - \tfrac{1}{4} = \tfrac{1}{2}\,. \end{aligned}$$

   b. The probability of error of $g_{\mathrm{MSE}}$ is

$$\begin{aligned} \mathrm{P}\left(g_{\mathrm{MSE}}(Y) \neq X\right) &= P\left(-1 \leq Y \leq 1\right) && (1) \\ &= \int_{-1}^{1} f_Y(y)\, \mathrm{d}y && (2) \\ &= \int_{-1}^{1} \frac{1}{4}\, \mathrm{d}y && (3) \\ &= \frac{1}{2}. && (4) \end{aligned}$$

c. The optimal decoder is given by the MAP rule. The *a posteriori* pmf of $X$ was found in part (a). Thus the MAP rule reduces to

$$g_{\text{error}}(y) = \begin{cases} -1 & -3 \le y \le -1 \\ \pm 1 & -1 < y < +1 \\ +1 & +1 < y \le +3 \end{cases}$$

Since either value can be chosen in the center range of $Y$, a *symmetric* decoder is sufficient, i.e.,

$$g_{\text{error}}(y) = \begin{cases} -1 & y < 0 \\ +1 & y \ge 0 \end{cases}$$

The probability of decoding error is

$$\begin{aligned} \mathrm{P}\{g_{\text{error}}(Y) \ne X\} &= \mathrm{P}\{X = -1,\, Y \ge 0\} \;+\; \mathrm{P}\{X = 1,\, Y < 0\} \\ &= \mathrm{P}\{X = -1 | Y \ge 0\}\, \mathrm{P}\{Y \ge 0\} \;+\; \mathrm{P}\{X = 1 | Y < 0\} \mathrm{P}\{Y < 0\} \\ &= \tfrac{1}{4} \cdot \tfrac{1}{2} + \tfrac{1}{4} \cdot \tfrac{1}{2} = \tfrac{1}{4}\,. \end{aligned}$$

This is half the probability of error of the minimum MSE estimator.

d. The MSE of $g_{\text{error}}(y)$ is

$$\mathrm{E}\left( (g_{\text{error}}(y) - X)^2 \right) = \tfrac{3}{4} \cdot 0 + \tfrac{1}{4} \cdot 2^2 = 1\,.$$

This MSE is twice that of the minimum MSE estimator.

2. (20 points) *Halloween parade (20 points)*

a. By Bayes rule the MAP criterion consists of comparing

$$p_{R|W}(r|w) = \frac{p_{W|R}(w|r) p_R(r)}{p_W(w)}$$

for different values of $r$. Since the denominator does not depend on $r$ we can just compare the numerators.

$$\begin{aligned} p_{W|R}(0|0) p_R(0) &= 0.7\, 0.8 = 0.56, \\ p_{W|R}(0|1) p_R(1) &= 0.3\, 0.2 = 0.06, \\ p_{W|R}(1|0) p_R(0) &= 0.3\, 0.8 = 0.24, \\ p_{W|R}(1|1) p_R(1) &= 0.7\, 0.2 = 0.14. \end{aligned}$$

For $w = 0$ and $w = 1$ the MAP criterion is maximized by $r = 0$, so you always predict no rain whatever the forecast says. The probability of error is consequently the probability that it rains which is 0.2 (note that this is better than the probability of error of the website forecast, which is 0.3).

b. It is not reasonable to assume that the forecast and the humidity are independent, even if we know that the forecast does not take the humidity into account. The reason is that both variables are linked through the rain. For example, if $W = 1$ then the humidity is probably high (because it probably will rain) and if $W = 0$ the humidity is probably low (because it probably won't rain). In contrast, conditioned on whether it rains or not, it is quite reasonable to assume independence of $W$ and $H$ because $H$ is not used to produce the forecast.

c. If $H = 0.65$ then you predict rain, since under your model $f_{H|R}(0.65|0) = 0$, which implies $f_{R|H}(0|0.65) = 0$ by Bayes rule.

d. By Bayes rule and the conditional independence assumption, the MAP criterion consists of comparing

$$p_{R|W,H}(r|w,h) = \frac{f_{H|W,R}(h|w,r)p_{W|R}(w|r)p_R(r)}{p_W(w)f_{H|W}(h|w)} = \frac{f_{H|R}(h|r)p_{W|R}(w|r)p_R(r)}{p_W(w)f_{H|W}(h|w)}$$

for different values of $r$. Since the denominator does not depend on $r$ we can just compare the numerators for $h = 0.55$ and $w = 1$:

$$\begin{aligned} f_{H|R}(0.55|1)p_{W|R}(1|1)p_R(1) &= \frac{1}{0.2}\,0.7\,0.2 = 0.7, \\ f_{H|R}(0.55|0)p_{W|R}(1|0)p_R(0) &= \frac{1}{0.5}\,0.3\,0.8 = 0.48. \end{aligned}$$

As a result, you predict that it will rain.

e. If $W = 1$, the MAP estimate chooses between

$$f_{H|R}(h|1)p_{W|R}(1|1)p_R(1) = \begin{cases} \frac{1}{0.2}\,0.7\,0.2 = 0.7 & \text{if } 0.5 \le h \le 0.7 \\ 0 & \text{otherwise} \end{cases}$$

$$f_{H|R}(h|0)p_{W|R}(1|0)p_R(0) = \begin{cases} \frac{1}{0.5}\,0.3\,0.8 = 0.48 & \text{if } 0.1 \le h \le 0.6 \\ 0 & \text{otherwise.} \end{cases}$$

As a result, the MAP estimate if $W = 1$ is

$$\hat{r} = \begin{cases} 1 & \text{if } 0.5 \le h \le 0.7 \\ 0 & \text{otherwise.} \end{cases}$$

If $W = 0$, the MAP estimate chooses between

$$f_{H|R}(h|1)p_{W|R}(0|1)p_R(1) = \begin{cases} \frac{1}{0.2}\,0.3\,0.2 = 0.3 & \text{if } 0.5 \le h \le 0.7 \\ 0 & \text{otherwise} \end{cases}$$

$$f_{H|R}(h|0)p_{W|R}(1|0)p_R(0) = \begin{cases} \frac{1}{0.5}\,0.7\,0.8 = 1.02 & \text{if } 0.1 \le h \le 0.6 \\ 0 & \text{otherwise.} \end{cases}$$

As a result, the MAP estimate if $W = 1$ is

$$\hat{r} = \begin{cases} 1 & \text{if } 0.6 \le h \le 0.7 \\ 0 & \text{otherwise.} \end{cases}$$

The MAP estimate is

$$\hat{r} = \begin{cases} 1 & \text{if } 0.6 \leq h \leq 0.7 \text{ or if both } 0.5 \leq h \leq 0.6 \text{ and } W = 1 \\ 0 & \text{otherwise.} \end{cases}$$

Finally, the probability of error is

$$\begin{aligned} \mathrm{P}(\text{Error}) &= \mathrm{P}(R=0)\mathrm{P}(W=1|R=0)\mathrm{P}(0.5 \leq h \leq 0.6|R=0) \\ &\quad + \mathrm{P}(R=1)\mathrm{P}(W=0|R=1)\mathrm{P}(0.5 \leq h \leq 0.6|R=1) \\ &= 0.8\, 0.3\, \frac{0.1}{0.5} + 0.2\, 0.3\, \frac{0.1}{0.2} = 0.078. \end{aligned}$$

3. (20 points) *Heart-disease detection (20 points).*

a. By Bayes rule, the chain rule and the conditional independence assumption

$$p_{H|S,C}(h|s,c) = \frac{p_H(h)p_{S,C|H}(s,c|h)}{p_{S,C}(s,c)} = \frac{p_H(h)p_{S|H}(s|h)p_{C|H}(c|h)}{p_{S,C}(s,c)}.$$

Since the denominator does not depend on $h$, the MAP rule reduces to

$$\hat{h}(s,c) = \begin{cases} 0 & \text{if } p_H(1)p_{S|H}(s|1)p_{C|H}(c|1) < p_H(0)p_{S|H}(s|0)p_{C|H}(c|0), \\ 1 & \text{otherwise.} \end{cases}$$

b. The error rate is 0.18.

c. By Bayes rule for mixed random variables, the chain rule and the conditional independence assumption

$$p_{H|S,C,X}(h|s,c,x) = \frac{p_H(h)p_{S,C|H}(s,c|h)f_{X|H,S,C}(x|h,s,c)}{f_{X|S,C}(x|s,c)p_{S,C}(s,c)} = \frac{p_H(h)p_{S|H}(s|h)p_{C|H}(c|h)f_{X|H}(x|h)}{f_{X|S,C}(x|s,c)p_{S,C}(s,c)}.$$

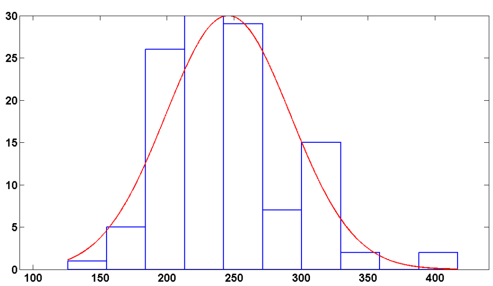
Since the denominator does not depend on $h$, the MAP rule reduces to

$$\hat{h}(s,c) = \begin{cases} 0 & \text{if } p_H(1)p_{S|H}(s|1)p_{C|H}(c|1)f_{X|H}(x|1) < p_H(0)p_{S|H}(s|0)p_{C|H}(c|0)f_{X|H}(x|0), \\ 1 & \text{otherwise.} \end{cases}$$

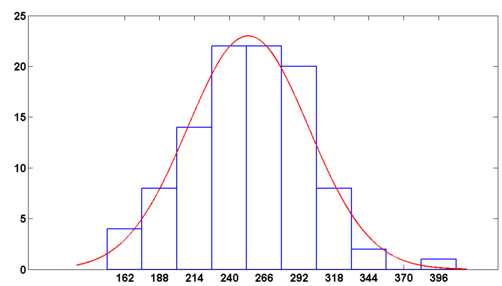
d. Figure 1 shows the histograms and the approximations to the conditional pdfs of $X$ given $H$ (scaled so that they are comparable to the histograms).

e. The error rate is 0.14. Given that there is not much difference between the cholesterol histograms in the presence and absence of heart disease and that the Gaussian fit is clearly not a very accurate representation of the empirical pdfs, we should not trust the result too much.

f. If we have enough data to estimate the joint distribution in a stable way, then this approach would probably improve the results. However, we would need more data than just 218 patients. Even without the cholesterol, the joint pmf of $C$, $S$ and $H$ has $2 \cdot 2 \cdot 4 - 1 = 15$ parameters, so we have less than 15 patients per parameter. In general, joint distributions are very difficult to approximate, since the number of parameters that are needed to characterize them are usually very large. Introducing some assumptions that are only approximately true but decrease significantly the number of parameters that we need to estimate often yield better results when there is limited data.




$H = 0$




$H = 1$

Figure 1: Histogram (blue) and approximation of the conditional pdf (red) of $X$ given the two possible values of $H$.