Assignment 8

Benjamin Jakubowski

November 16, 2015

1. Probability of Error Vs. MSE

a. Finding MSE stimate of X given Y

First recall (from Theorem 1.2 in the lecture notes) the minimum MSE estimate g_{MSE} of X given Y is

$$E(X|Y) = \arg\min_g E\left((X - g(Y))^2\right)$$

Now, if we fix $X = x$, then Y is uniformly distributed as

$$f_{Y|X}(y|x) = \begin{cases} \frac{1}{4} & \text{if } x-2 \leq y \leq x+2 \\ 0 & \text{otherwise} \end{cases}$$

Now, we find the marginal pdf of Y.

$$\begin{aligned} f_Y(y) &= \sum_{x=-1,1} f_{Y|X}(y|x)P_X(x) \\ &= f_{Y|X}(y|1)\frac{1}{2} + f_{Y|X}(y|-1)\frac{1}{2} \end{aligned}$$

When $x = 1$,

$$f_{Y|X}(y|1) = \begin{cases} \frac{1}{4} & \text{if } -1 \leq y \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

When $x = -1$,

$$f_{Y|X}(y|1) = \begin{cases} \frac{1}{4} & \text{if } -3 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

So

$$\begin{aligned} f_Y(y) &= \frac{1}{8}\mathbb{1}_{-1\leq y\leq 3} + \frac{1}{8}\mathbb{1}_{-3\leq y\leq 1} \\ &= \begin{cases} \frac{1}{8} & \text{if } -3 \leq y < -1 \\ \frac{1}{4} & \text{if } -1 \leq y \leq 1 \\ \frac{1}{8} & \text{if } 1 < y \leq 3 \end{cases} \end{aligned}$$

Now we compute the conditional PMF of X given Y:

$$P(X = x|Y = y) = \frac{P_X(x)f_{Y|X}(y, x)}{f_Y(y)}$$

If $-3 \leq y < -1$:

$$P(X = x|Y = y) = \begin{cases} \frac{1/2 \cdot 1/4}{1/8} = 1 & \text{for } x = -1 \\ \frac{1/2 \cdot 0}{1/8} = 0 & \text{for } x = 1 \end{cases}$$

If $-1 \leq y \leq 1$:

$$P(X = x|Y = y) = \begin{cases} \frac{1/2 \cdot 1/4}{1/8} = 1/2 & \text{for } x = -1 \\ \frac{1/2 \cdot 1/4}{1/8} = 1/2 & \text{for } x = 1 \end{cases}$$

If $1 < y \leq 3$:

$$P(X = x|Y = y) = \begin{cases} \frac{1/2 \cdot 0}{1/8} = 0 & \text{for } x = -1 \\ \frac{1/2 \cdot 1/4}{1/8} = 1 & \text{for } x = 1 \end{cases}$$

Finally, we find $E(X|Y = y)$:

$$g_{MSE} = E(X|Y = y) = \begin{cases} -1 & \text{if } -3 \leq y < -1 \\ -1 \cdot 1/2 + 1 \cdot 1/2 = 0 & \text{if } -1 \leq y \leq 1 \\ 1 & \text{if } 1 < y \leq 3 \end{cases}$$

Now we use iterated expectations to find the MSE:

$$E\left((X - g_{MSE}(Y))^2\right) = \int_{R_Y} E\left((X - g_{MSE}(y))^2|Y = y\right) f_Y(y)\mathrm{d}y$$

Note if $-3 \leq y < 1, 1 < y \leq 3$ then $X - g_{MSE}(y) = 0$. Morever, if $-1 \leq y \leq 1$ then $g_{MSE} = 0$. Thus, this expression simplifies to

$$\begin{aligned} \int_{y=-1}^{y=1} E(X^2|Y = y) \cdot \frac{1}{4}\mathrm{d}y &= \frac{1}{4}\int_{y=-1}^{y=1} \left[\sum_{x=-1,1} x^2 P_{X|Y}(x|y)\right] \mathrm{d}y \\ &= \frac{1}{4}\int_{y=-1}^{y=1} \left[\frac{1}{2}(-1)^2 + \frac{1}{2}(1)^2\right] \mathrm{d}y \\ &= \frac{1}{4}\int_{y=-1}^{y=1} (1)\mathrm{d}y = \frac{1}{4} \cdot 2 = \frac{1}{2} \end{aligned}$$

B. Probability of error using g_{MSE}

The probability of error of g_{MSE} is $1/2$ (since $g_{MSE}(y) \neq X$ if and only if $-1 \leq y \leq 1$, which occurs with probability $1/2$).

c. Optimal decoder g_{error} to minimize probability of error

By Theorem 2.3 in the lecture notes, the MAP estimator minimizes the probability of error. Now recall If $-1 \leq y \leq 1, P(X = -1|Y = y) = P(X = 1|Y = y) = 0.5$. Thus, we will arbitrarily set $g_{MAP}(Y) = -1$ over this interval.
Then, using the posterior distributions over found in question (a), we have:

$$g_{MAP}(y) = \begin{cases} -1 & \text{if } -3 \leq y < -1 \\ -1 & \text{if } -1 \leq y \leq 1 \\ 1 & \text{if } 1 < y \leq 3 \end{cases}$$

Now, note if $-3 \leq y < -1$ or $1 < y \leq 3$, $g_{MAP} = X$ so $P(g_{MAP} \neq X) > 0$ only if $-1 \leq y \leq 1$. Thus, the probability of error is

$$\begin{aligned} P(g_{MAP} \neq X) &= P(-1 \leq Y \leq 1, X = 1) \\ &= P(-1 \leq Y \leq 1|X = 1)P(X = 1) \\ &= 1/2 * 1/2 = 1/4 \end{aligned}$$

d. Comparing MSE of g_{error} to the minimum MSE

First, note $X - g_{MAP}(y) = 0$ if $-3 \leq y < 1$, $1 < y \leq 3$, or $-1 \leq y \leq 1$ and $X = -1$
Then using similar reasoning as in (a), the MSE of $g_{error} = g_{MAP}$

$$\begin{aligned} \int_{y=-1}^{y=1} E((X - g_{MAP}(y))^2|Y = y) \cdot \frac{1}{4} \mathrm{d}y &= \frac{1}{4} \int_{y=-1}^{y=1} \left[\sum_{x=-1,1} (X - g_{MAP}(y))^2 P_{X|Y}(x|y) \right] \mathrm{d}y \\ &= \frac{1}{4} \int_{y=-1}^{y=1} \left[\frac{1}{2}(0)^2 + \frac{1}{2}(2)^2 \right] \mathrm{d}y \\ &= \frac{1}{4} \int_{y=-1}^{y=1} (2) \mathrm{d}y = \frac{1}{4} \cdot 4 = 1 \end{aligned}$$

Thus, using g_{MAP}, you are half as likely to make an error, but the MSE is twice as large.

2. Halloween Parade

a. Predicting rain given the forecast of the website

First, recall the random variable:

- R: Whether or not it rains
- W: Whether the forecast calls for rain

The given probability distributions are:

$$P_R(r) = \begin{cases} .2 & \text{if } r = 1 \\ .8 & \text{if } r = 0 \end{cases}$$

$$P_(W = 1|R = 1) = .7$$
$$P_(W = 0|R = 0) = .7$$

We are interested in

$$\hat{R}_{MAP} = \arg \max_r P_{R|W}(r, w)$$

Well,

$$\begin{aligned} P_{R|W}(r, w) &= \frac{P_{W|R}(w, r)P_R(r)}{P_W(w)} \\ &= \frac{P_{W|R}(w, r)P_R(r)}{\sum_{r=0,1} P_{W|R}(w|r)P_R(r)} \end{aligned}$$

When $w = 0$:

$$P_{R|W}(r, w) = \begin{cases} \frac{P_{W|R}(0,0)P_R(0)}{P_{W|R}(0,0)P_R(0)+P_{W|R}(0,1)P_R(1)} = \frac{.7\cdot.8}{.7\cdot.8+.3\cdot.2} \approx .9032 & \text{for } r = 0 \\ \frac{P_{W|R}(0,1)P_R(1)}{P_{W|R}(0,0)P_R(0)+P_{W|R}(0,1)P_R(1)} = \frac{.3\cdot.2}{.7\cdot.8+.3\cdot.2} \approx .0968 & \text{for } r = 1 \end{cases}$$

When $w = 1$:

$$P_{R|W}(r, w) = \begin{cases} \frac{P_{W|R}(1,0)P_R(0)}{P_{W|R}(1,0)P_R(0)+P_{W|R}(1,1)P_R(1)} = \frac{.3\cdot.8}{.3\cdot.8+.7\cdot.2} \approx .6316 & \text{for } r = 0 \\ \frac{P_{W|R}(1,1)P_R(1)}{P_{W|R}(1,0)P_R(0)+P_{W|R}(1,1)P_R(1)} = \frac{.7\cdot.2}{.3\cdot.8+.3\cdot.2} \approx .3684 & \text{for } r = 1 \end{cases}$$

Thus, for all $W, \hat{R}_{MAP} = 0$, so using $\hat{R}_{MAP}$ we will always predict it will not rain. As such, the probability of error is simply the probability it will rain (or 0.2).

b. Adding Humidity to our model

If humidity H is known, but not used in the online weather forecast, then it is more reasonable to assume that H and W are conditionally independent given R than to assume they are independent. This is because we can reasonably assume $P_{W|H}(w|h) \neq P_W(w)$, as the humidity (can reasonably be assumed to) provide information about R (which is decidely **not** independent of W).

However, it is reasonable to assume W and H are conditionally independent, given R, since any information about R available from H is redundant (given R), and we know the online weather forecast is not using using any other information provided by H.

c. Forecast if $H = 0.65$ and the website predicts no rain

First, lets (unnecessarily for part c, but needed for subsequent questions) derive the MAP estimator of R given W and H. Again, recall we are assuming H and W are conditionally independent given R, and

$$f_{H|R}(h|1) = \begin{cases} 5 & \text{if } 0.5 \leq h \leq 0.7 \\ 0 & \text{otherwise} \end{cases}$$

$$f_{H|R}(h|0) = \begin{cases} 2 & \text{if } 0.1 \leq h \leq 0.6 \\ 0 & \text{otherwise} \end{cases}$$

Now we are interested in

$$\hat{R}_{MAP} = \arg \max_{r \in \{0,1\}} P_{R|W,H}(r|w,h)$$

Well,

$$\begin{aligned} P_{R|W,H}(r|w,h) &= \frac{P_{W|R,H}(w|r,h) f_{H|R}(h|r) P_R(r)}{P_{W,H}(w,h)} \\ &= \frac{P_{W|R}(w|r) f_{H|R}(h|r) P_R(r)}{\sum_{r=0,1} f_{W,H|R}(w,h|r) P_R(r)} \qquad \text{by conditional independence} \\ &= \frac{P_{W|R}(w|r) f_{H|R}(h|r) P_R(r)}{\sum_{r=0,1} P_{W|R}(w|r) f_{H|R}(h|r) P_R(r)} \qquad \text{again by cond. ind.} \end{aligned}$$

Now, if $w = 0$ and $H = 0.65$:

$$\begin{aligned} P_{R|W,H}(0|0,0.65) &= \frac{P_{W|R}(0|0) f_{H|R}(0.65|0) P_R(0)}{P_{W|R}(0|0) f_{H|R}(0.65|0) P_R(0) + P_{W|R}(0|1) f_{H|R}(0.65|1) P_R(1)} \\ &= \frac{0.7 \cdot 0 \cdot .8}{0.7 \cdot 0 \cdot .8 + 0.3 \cdot 5 \cdot .2} = 0 \end{aligned}$$

$$P_{R|W,H}(1|0,0.65) = \frac{0.3 \cdot 5 \cdot .2}{0.7 \cdot 0 \cdot .8 + 0.3 \cdot 5 \cdot .2} = 1$$

Thus, given $W = 0$ and $H = 0.65$, it will rain with probability 1.

d. Forecast if $H = 0.55$ and the website predicts rain

Now, if $w = 1$ and $H = 0.55$:

$$\begin{aligned} P_{R|W,H}(0|1,0.55) &= \frac{P_{W|R}(1|0) f_{H|R}(0.55|0) P_R(0)}{P_{W|R}(1|0) f_{H|R}(0.55|0) P_R(0) + P_{W|R}(1|1) f_{H|R}(0.55|1) P_R(1)} \\ &= \frac{0.3 \cdot 2 \cdot .8}{0.3 \cdot 2 \cdot .8 + 0.7 \cdot 5 \cdot .2} \approx .4068 \end{aligned}$$

$$P_{R|W,H}(1|0,0.65) = \frac{0.7 \cdot 5 \cdot .2}{0.3 \cdot 2 \cdot .8 + 0.7 \cdot 5 \cdot .2} \approx 0.5932$$

Thus, using the MAP estimator, we would forecast rain.

e. Probability of error under new model

The probability of error is

$$\begin{aligned} P(R \neq \hat{R}_{MAP}(h,w)) &= P(R \neq \hat{R}_{MAP}(h,w)|R=0)P_R(0) \\ &\quad + P(R \neq \hat{R}_{MAP}(h,w)|R=1)P_R(1) \end{aligned}$$

We'll tackle this expression in two pieces- first, consider $P(R \neq \hat{R}_{MAP}(h,w)|R=0)$. Note, given $R=0$:

$$(R \neq \hat{R}_{MAP}(h,w)) \iff \left(P_{W|R}(w|1)f_{H|R}(h|1)P_R(1) > P_{W|R}(w|0)f_{H|R}(h|0)P_R(0)\right)$$

Now, recall the probability of error is 0 unless $0.5 \leq H \leq 0.6$. Thus we can substitute into the expression above

$$\begin{aligned} P_{W|R}(w|1)f_{H|R}(h|1)P_R(1) &> P_{W|R}(w|0)f_{H|R}(h|0)P_R(0) \\ P_{W|R}(w|1) \cdot 5 \cdot 0.2 &> P_{W|R}(w|0) \cdot 2 \cdot 0.8 \\ P_{W|R}(w|1) - 1.6 \cdot P_{W|R}(w|0) &> 0 \end{aligned}$$

Now, if $w=0$:

$$P_{W|R}(0|1) - 1.6 \cdot P_{W|R}(0|0) = 0.3 - 1.6 \cdot 0.7 = -0.82 < 0$$

If $w=1$:

$$P_{W|R}(1|1) - 1.6 \cdot P_{W|R}(1|0) = 0.7 - 1.6 \cdot 0.3 = 0.22 > 0$$

Thus, given $R=0$, $R \neq \hat{R}_{MAP}(h,w) \iff H \in [0.5, 0.6]$ and $W=1$, so:

$$\begin{aligned} P(R \neq \hat{R}_{MAP}(h,w)|R=0) &= P(0.5 \leq H \leq 0.6, W=1|R=0) \\ &= P(0.5 \leq H \leq 0.6|R=0)P(W=1|R=0) \\ &= 0.2 \cdot 0.3 = 0.6 \end{aligned}$$

Now consider $P(R \neq \hat{R}_{MAP}(h,w)|R=1)$. Now, given $R=1$:

$$(R \neq \hat{R}_{MAP}(h,w)) \iff \left(P_{W|R}(w|0)f_{H|R}(h|0)P_R(0) > P_{W|R}(w|1)f_{H|R}(h|1)P_R(1)\right)$$

Recognizing this is just the reverse of the previous inequality, we conclude (given $R=1$) $R \neq \hat{R}_{MAP}(h,w) \iff H \in [0.5, 0.6]$ and $W=0$. Thus:

$$\begin{aligned} P(R \neq \hat{R}_{MAP}(h,w)|R=0) &= P(0.5 \leq H \leq 0.6, W=0|R=1) \\ &= P(0.5 \leq H \leq 0.6|R=0)P(W=0|R=1) \\ &= 0.5 \cdot 0.3 = 1.5 \end{aligned}$$

Now, let's put it all together:

$$\begin{aligned} P(R \neq \hat{R}_{MAP}(h,w)) &= P(R \neq \hat{R}_{MAP}(h,w)|R=0)P_R(0) \\ &\quad + P(R \neq \hat{R}_{MAP}(h,w)|R=1)P_R(1) \\ &= 0.06 \cdot 0.8 + 0.15 \cdot 0.2 = 0.078 \end{aligned}$$

Thus, under our new model (using $\hat{R}_{MAP}$), the probability of error is 0.078.

3. Heart Disease Detection

a. MAP estimator for H given S and C

First, recall the random variables:

- H: Whether or not the patient has heart disease
- S: Sex of the patient
- C: Type of chest pain experienced by patient

Given S and C, the MAP estimate $\hat{H}_{MAP}$ is

$$\hat{H}_{MAP}(s,c) = \arg\max_h P_{H|S,C}(h|s,c) \qquad \text{where } h \in \{0,1\}$$

Well, through repeated application of conditional independence, we find:

$$\begin{aligned} P_{H|S,C}(h|s,c) &= \frac{P(H=h,S=s,C=c)}{P(S=s,C=c)} \\ &= \frac{P(S=s|H=h,C=c)P(C=c|H=h)P(H=h)}{\sum_{h\in\{0,1\}} P(S=s,C=c|H=h)P(H=h)} \\ &= \frac{P(S=s|H=h)P(C=c|H=h)P(H=h)}{\sum_{h\in\{0,1\}} P(S=s|H=h)P(C=c|H=h)P(H=h)} \end{aligned}$$

Since the denominator is constant, we end up with:

$$\hat{H}_{MAP}(s,c) = \begin{cases} 1 & \text{if } P_{S|H}(s|0)P_{C|H}(c|0)P_H(0) < P_{S|H}(s|1)P_{C|H}(c|1)P_H(1) \\ 0 & \text{otherwise} \end{cases}$$

b. Training and testing MAP estimator

Using this MAP estimator trained on 218 observations, we misclassify 9 of the 50 patients in the test set. Thus, our error rate is 0.18.

c. New MAP estimator accounting for cholesterol

Now, let

$$\hat{H}_{MAP}(s,c,x) = \arg\max_h P_{H|S,C,X}(h|s,c,x) \qquad \text{where } h \in \{0,1\}$$

Using similar reasoning as in (a) (i.e. repeated applications of conditional independence):

$$\begin{aligned} P_{H|S,C,X}(h|s,c,x) &= \frac{f(H=h,S=s,C=c,X=x)}{k} \qquad \text{(where k is just a constant)} \\ &= \frac{P_{S|H}(s|h)\cdot P_{C|H}(c|h)\cdot f_{x|H}(x|h)\cdot P_H(h)}{k} \end{aligned}$$

Thus, we end up with:

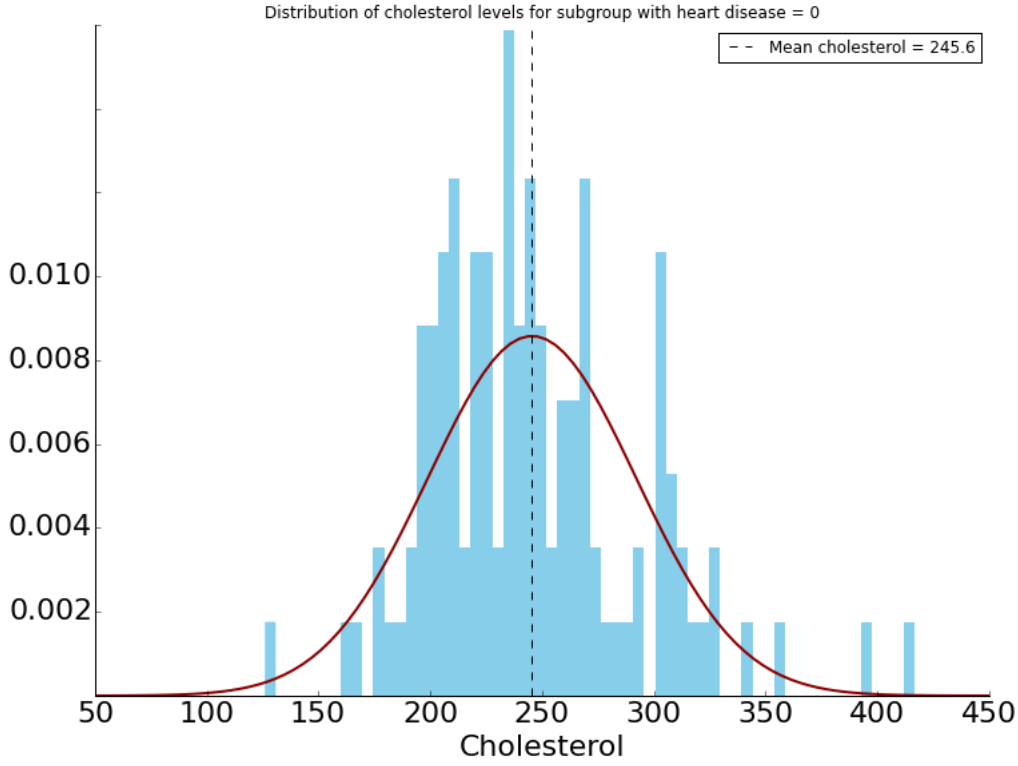
$$\hat{H}_{MAP}(s,c) = \begin{cases} 1 & \text{if } P_{S|H}(s|0)P_{C|H}(c|0)f_{x|H}(x|0)P_H(0) < P_{S|H}(s|1)P_{C|H}(c|1)f_{x|H}(x|1)P_H(1) \\ 0 & \text{otherwise} \end{cases}$$

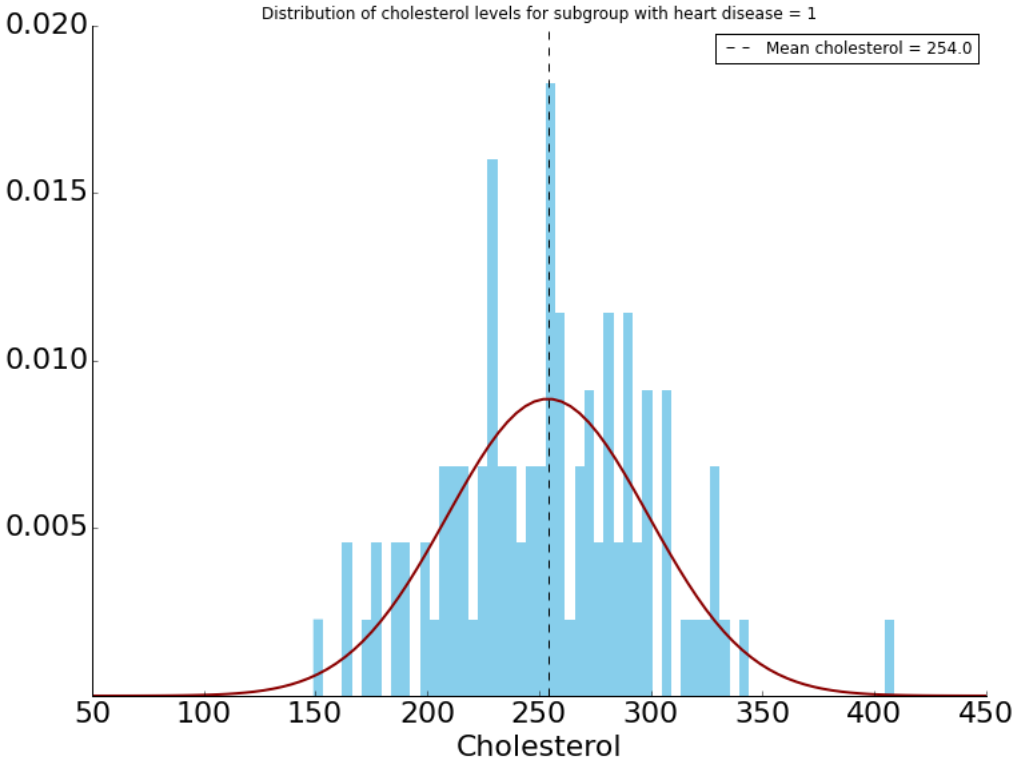
D. Estimating conditional distribution of cholesterol level

The sample mean, sample variance, and sample standard deviation cholesterol level, given heart disease status, is shown below:

Heart Disease Status	Sample Mean	Sample Variance	Sample Std. Dev.
Non-diseased	245.6	2164.2	46.5
Diseased	254.0	2026.7	45.0

In addition, the empirical conditional distributions and estimated (normal) conditional distributions are shown below. The parametric estimate appears to fit the data fairly well, except for the extreme (high) values observed in both distributions. However, the fit appears good enough to support using the parametric estimator for $f_{X|H}(x, h)$.





e. Determining the error rate for our new MAP estimate

Using our new MAP estimator trained on 218 observations, we misclassify 7 of the 50 patients in the test set. Thus, our error rate is 0.14. I trust this result, since it was determined using an out-of-sample test set.

f. Questioning conditional independence assumptions

We assumed the predictors were all conditionally independent given H. This assumption allowed us to estimate the probability distributions of predictors conditioned on only H. This is useful when the sample size is relatively small, because if the sample size is small and we don't make this assumption, then our conditional probability estimates for subgroups may become unstable.

To make this concrete, let's consider an example. In part (a), we used conditional independence to equate $P_{S|C,H}(s|c,h) = P_{S|H}(s|h)$. We then estimated this distribution by partitioning the 218 samples in the test set into four subsets ($R_S \times R_H$). If we had not made this assumption, we would have had to estimate the distribution of $P_{S|C,H}(s|c,h)$ by partitioning the 218 samples in the test set into 16 subsets ($R_S \times R_H \times R_C$). Given

that each of these subsets would contain fewer numbers of samples, we would increase the variance in our model (due to greater variance in probability estimates). Moreover, this increase in variance would be even worse in part (d), where we added yet another feature to our model.

To conclude, we can relate this all back to a key concept in modeling, namely bias-variance tradeoff. By assuming conditional independence, we effectively decreased test variance (and avoided overfitting) at the expense of greater model bias (since, if the predictors are not conditionally independent, we have a misspecified model).