STK2100: Solutions Week 14

Lars H. B. Olsen

14.04.2021

ISLR

Section 8.3.1

This you have to do on your own. Just read the section and understand the well described R-commands in said section.

Exercise 8.11

a)

```r
library(ISLR)
head(Caravan)
```

```
##   MOSTYPE MAANTHUI MGEMOMV MGEMLEEF MOSHOOFD MGODRK MGODPR MGODOV MGODGE MRELGE
## 1      33        1       3        2        8      0      5      1      3      7
## 2      37        1       2        2        8      1      4      1      4      6
## 3      37        1       2        2        8      0      4      2      4      3
## 4       9        1       3        3        3      2      3      2      4      5
## 5      40        1       4        2       10      1      4      1      4      7
## 6      23        1       2        1        5      0      5      0      5      0
##   MRELSA MRELOV MFALLEEN MFGEKIND MFWEKIND MOPLHOOG MOPLMIDD MOPLLAAG MBERHOOG
## 1      0      2        1        2        6        1        2        7        1
## 2      2      2        0        4        5        0        5        4        0
## 3      2      4        4        4        2        0        5        4        0
## 4      2      2        2        3        4        3        4        2        4
## 5      1      2        2        4        4        5        4        0        0
## 6      6      3        3        5        2        0        5        4        2
##   MBERZELF MBERBOER MBERMIDD MBERARBG MBERARBO MSKA MSKB1 MSKB2 MSKC MSKD
## 1        0        1        2        5        2    1     1     2    6    1
## 2        0        0        5        0        4    0     2     3    5    0
## 3        0        0        7        0        2    0     5     0    4    0
## 4        0        0        3        1        2    3     2     1    4    0
## 5        5        4        0        0        0    9     0     0    0    0
## 6        0        0        4        2        2    2     2     2    4    2
##   MHHUUR MHKOOP MAUT1 MAUT2 MAUT0 MZFONDS MZPART MINKM30 MINK3045 MINK4575
## 1      1      8     8     0     1       8      1       0        4        5
## 2      2      7     7     1     2       6      3       2        0        5
## 3      7      2     7     0     2       9      0       4        5        0
## 4      5      4     9     0     0       7      2       1        5        3
## 5      4      5     6     2     1       5      4       0        0        9
## 6      9      0     5     3     3       9      0       5        2        3
##   MINK7512 MINK123M MINKGEM MKOOPKLA PWAPART PWABEDR PWALAND PPERSAUT PBESAUT
```

```
## 1        0        0       4        3       0       0       0        6       0
## 2        2        0       5        4       2       0       0        0       0
## 3        0        0       3        4       2       0       0        6       0
## 4        0        0       4        4       0       0       0        6       0
## 5        0        0       6        3       0       0       0        0       0
## 6        0        0       3        3       0       0       0        6       0
##   PMOTSCO PVRAAUT PAANHANG PTRACTOR PWERKT PBROM PLEVEN PPERSONG PGEZONG
## 1       0       0        0        0      0     0      0        0       0
## 2       0       0        0        0      0     0      0        0       0
## 3       0       0        0        0      0     0      0        0       0
## 4       0       0        0        0      0     0      0        0       0
## 5       0       0        0        0      0     0      0        0       0
## 6       0       0        0        0      0     0      0        0       0
##   PWAOREG PBRAND PZEILPL PPLEZIER PFIETS PINBOED PBYSTAND AWAPART AWABEDR
## 1       0      5       0        0      0       0        0       0       0
## 2       0      2       0        0      0       0        0       2       0
## 3       0      2       0        0      0       0        0       1       0
## 4       0      2       0        0      0       0        0       0       0
## 5       0      6       0        0      0       0        0       0       0
## 6       0      0       0        0      0       0        0       0       0
##   AWALAND APERSAUT ABESAUT AMOTSCO AVRAAUT AAANHANG ATRACTOR AWERKT ABROM
## 1       0        1       0       0       0        0        0      0     0
## 2       0        0       0       0       0        0        0      0     0
## 3       0        1       0       0       0        0        0      0     0
## 4       0        1       0       0       0        0        0      0     0
## 5       0        0       0       0       0        0        0      0     0
## 6       0        1       0       0       0        0        0      0     0
##   ALEVEN APERSONG AGEZONG AWAOREG ABRAND AZEILPL APLEZIER AFIETS AINBOED
## 1      0        0       0       0      1       0        0      0       0
## 2      0        0       0       0      1       0        0      0       0
## 3      0        0       0       0      1       0        0      0       0
## 4      0        0       0       0      1       0        0      0       0
## 5      0        0       0       0      1       0        0      0       0
## 6      0        0       0       0      0       0        0      0       0
##   ABYSTAND Purchase
## 1        0       No
## 2        0       No
## 3        0       No
## 4        0       No
## 5        0       No
## 6        0       No
```

```r
train = 1:1000
Caravan$Purchase = ifelse(Caravan$Purchase == "Yes", 1, 0)
Caravan.train = Caravan[train, ]
Caravan.test = Caravan[-train, ]
```

b)

```r
library(gbm)
```

```
## Loaded gbm 2.1.8
```

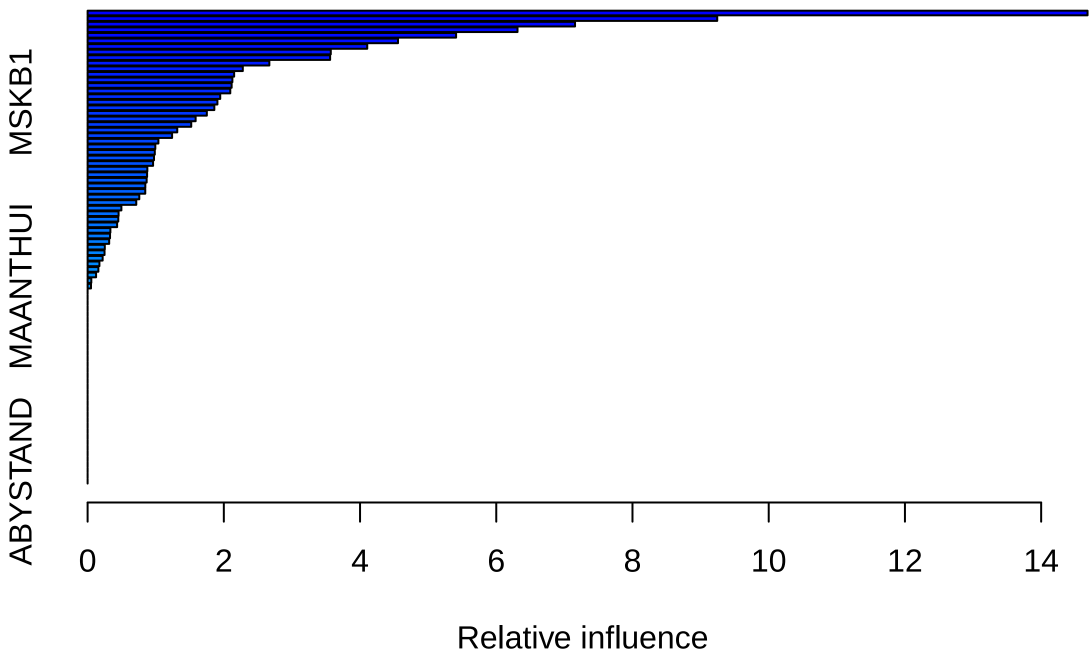
```r
set.seed(123)
Caravan.gbm=gbm(Purchase~.,data=Caravan.train,n.trees = 1000,
```

```r
                 shrinkage = 0.01,distribution = "bernoulli")
```

```
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :
## variable 50: PVRAAUT has no variation.

## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :
## variable 71: AVRAAUT has no variation.
```

```r
summary(Caravan.gbm)
```




```
##                var     rel.inf
## PPERSAUT PPERSAUT 14.68137994
## MKOOPKLA MKOOPKLA  9.24410213
## MOPLHOOG MOPLHOOG  7.15657002
## MBERMIDD MBERMIDD  6.31195378
## PBRAND     PBRAND  5.41026779
## MGODGE     MGODGE  4.55751022
## MINK3045 MINK3045  4.10560259
## ABRAND     ABRAND  3.57059504
## MOSTYPE   MOSTYPE  3.56070215
## MGODPR     MGODPR  2.66819294
## MSKC         MSKC  2.27879862
## PWAPART   PWAPART  2.15162241
## MAUT1       MAUT1  2.12566124
## MBERARBG MBERARBG  2.11473767
## MSKA         MSKA  2.09466053
## MAUT2       MAUT2  1.94714666
## MSKB1       MSKB1  1.90610028
## PBYSTAND PBYSTAND  1.86069156
## MINKGEM   MINKGEM  1.75040846
## MRELGE     MRELGE  1.58600540
## MGODOV     MGODOV  1.52158243
## MRELOV     MRELOV  1.31632452
## MFWEKIND MFWEKIND  1.24056776
```

```
## MOPLMIDD MOPLMIDD  1.03962223
## MBERHOOG MBERHOOG  0.99408658
## MBERBOER MBERBOER  0.98653159
## MAUTO        MAUTO  0.97489594
## MFGEKIND MFGEKIND  0.96135899
## APERSAUT APERSAUT  0.87687578
## MGODRK      MGODRK  0.87337866
## MINKM30    MINKM30  0.86704169
## MHHUUR      MHHUUR  0.84714186
## MINK4575 MINK4575  0.84653461
## MBERARBO MBERARBO  0.75878372
## MINK7512 MINK7512  0.71473591
## MSKD          MSKD  0.49622356
## MFALLEEN MFALLEEN  0.45489686
## MGEMOMV    MGEMOMV  0.45067730
## PMOTSCO    PMOTSCO  0.43357481
## PLEVEN      PLEVEN  0.33438953
## MOSHOOFD MOSHOOFD  0.32924947
## MSKB2        MSKB2  0.31682618
## MHKOOP      MHKOOP  0.25261322
## MZFONDS    MZFONDS  0.24844483
## MINK123M MINK123M  0.22095401
## MGEMLEEF MGEMLEEF  0.17326653
## MOPLLAAG MOPLLAAG  0.15831615
## MRELSA      MRELSA  0.12436372
## MZPART      MZPART  0.05445179
## MAANTHUI MAANTHUI  0.04958036
## MBERZELF MBERZELF  0.00000000
## PWABEDR    PWABEDR  0.00000000
## PWALAND    PWALAND  0.00000000
## PBESAUT    PBESAUT  0.00000000
## PVRAAUT    PVRAAUT  0.00000000
## PAANHANG PAANHANG  0.00000000
## PTRACTOR PTRACTOR  0.00000000
## PWERKT      PWERKT  0.00000000
## PBROM        PBROM  0.00000000
## PPERSONG PPERSONG  0.00000000
## PGEZONG    PGEZONG  0.00000000
## PWAOREG    PWAOREG  0.00000000
## PZEILPL    PZEILPL  0.00000000
## PPLEZIER PPLEZIER  0.00000000
## PFIETS      PFIETS  0.00000000
## PINBOED    PINBOED  0.00000000
## AWAPART    AWAPART  0.00000000
## AWABEDR    AWABEDR  0.00000000
## AWALAND    AWALAND  0.00000000
## ABESAUT    ABESAUT  0.00000000
## AMOTSCO    AMOTSCO  0.00000000
## AVRAAUT    AVRAAUT  0.00000000
## AAANHANG AAANHANG  0.00000000
## ATRACTOR ATRACTOR  0.00000000
## AWERKT      AWERKT  0.00000000
## ABROM        ABROM  0.00000000
## ALEVEN      ALEVEN  0.00000000
```

```
## APERSONG APERSONG  0.00000000
## AGEZONG   AGEZONG  0.00000000
## AWAOREG   AWAOREG  0.00000000
## AZEILPL   AZEILPL  0.00000000
## APLEZIER APLEZIER  0.00000000
## AFIETS     AFIETS  0.00000000
## AINBOED   AINBOED  0.00000000
## ABYSTAND ABYSTAND  0.00000000
```

PPERSAUT is the most important variable.

c)

```
Caravan.pred=predict(Caravan.gbm,Caravan.test,n.trees = 1000,type='response')
Caravan.pred=ifelse(Caravan.pred>0.2,1,0)
table(Caravan.test$Purchase,Caravan.pred)
```

```
##    Caravan.pred
##        0    1
##   0 4420  113
##   1  256   33
```

```
31/(113 + 33)
```

```
## [1] 0.2123288
```

About 21% of the people predicted to make a purchase do make one.

```
Caravan.glm=glm(Purchase~.,family='binomial',data = Caravan.train)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
Caravan.pred=predict(Caravan.glm,Caravan.test,type='response')
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```
Caravan.pred=ifelse(Caravan.pred>0.2,1,0)
table(Caravan.test$Purchase,Caravan.pred)
```

```
##    Caravan.pred
##        0    1
##   0 4183  350
##   1  231   58
```

```
58/(350 + 58)
```

```
## [1] 0.1421569
```

About 14% of people predicted to make purchase using logistic regression actually end up making one. This is lower than boosting.

Exercise 8.12

Logistic regression:

```
set.seed(1)
Caravan.glm=glm(Purchase~.,family='binomial',data = Caravan.train)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
Caravan.pred=predict(Caravan.glm,Caravan.test,type='response')
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```r
Caravan.pred=ifelse(Caravan.pred>0.5,1,0)
table(Caravan.test$Purchase, Caravan.pred)
```

```
##    Caravan.pred
##        0    1
##   0 4446   87
##   1  274   15
```

```r
mean(Caravan.pred == Caravan.test$Purchase)
```

```
## [1] 0.9251348
```

Boosting:

```r
library(gbm)
Caravan.gbm=gbm(Purchase~.,data=Caravan.train,n.trees = 1000,
                shrinkage = 0.01,distribution = "bernoulli")
```

```
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :
## variable 50: PVRAAUT has no variation.
```

```
## Warning in gbm.fit(x = x, y = y, offset = offset, distribution = distribution, :
## variable 71: AVRAAUT has no variation.
```

```r
Caravan.pred=predict(Caravan.gbm,Caravan.test,n.trees = 1000,type='response')
Caravan.pred=ifelse(Caravan.pred>0.5,1,0)
table(Caravan.test$Purchase,Caravan.pred)
```

```
##    Caravan.pred
##        0    1
##   0 4531    2
##   1  289    0
```

```r
mean(Caravan.pred == Caravan.test$Purchase)
```

```
## [1] 0.9396516
```

Bagging:

```r
library(randomForest)
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```r
c.train = Caravan.train
c.train$Purchase = factor(c.train$Purchase)

c.test = Caravan.test
c.test$Purchase = factor(c.test$Purchase)

bag = randomForest(Purchase~.,data=c.train, mtry = ncol(c.train)-1)
yhat.bag = predict(bag, newdata = c.test)
table(yhat.bag, c.test$Purchase)
```

```
##
```

```
## yhat.bag    0    1
##        0 4474  277
##        1   59   12
```

```r
mean(yhat.bag == c.test$Purchase)
```

```
## [1] 0.9303194
```

Random forests:

```r
rf = randomForest(Purchase~.,data=c.train, distribution = "bernoulli")
yhat.rf = predict(rf, newdata = c.test)
table(yhat.rf, c.test$Purchase)
```

```
##
## yhat.rf    0    1
##       0 4505  282
##       1   28    7
```

```r
mean(yhat.rf == c.test$Purchase)
```

```
## [1] 0.9357113
```

All methods give comparable performance, but note that by always classifying all instances as 0, we get an accuracy of 0.94.

```r
table(c.test$Purchase)
```

```
##
##    0    1
## 4533  289
```

```r
4533 / (4533 + 289)
```

```
## [1] 0.9400664
```

Exam STK2100 Spring 2019

Solutions are directly copied from https://www.uio.no/studier/emner/matnat/math/STK2100/v19/eksamen/solution_stk2100_2019.pdf.

Exercises 1

a)

There is no protecting factor, as all regression coefficients are larger than 0. The intercept includes the baseline effects, in this case the log-odds for a female, non-smoker who does not drink and has age = 0. The latter is the reason why the intercept does not have a "physical" meaning in this case. A solution is to center age, so the baseline is a female with average age. Since the regression coefficient is not significant, it may also be excluded from the model (although, strictly speaking, in this way the problem is removed, more than solved…).

b)

The correct answer is model 2, as it has the best performance on the test set in terms of the ROC-curve (valid explanations: largest area under the curve, closest curve to the top-left corner, farthest curve from the intercept). The ROC curve is a graphical tool to visualize the performance of a classification model, and it displays the sensitivity and (1 minus) specificity when moving the threshold used to discriminate between the two response classes. Sensitivity = true positive / (true positive + false negative), where true positive are the observations correctly identified as positive by the model, and false negative the observations incorrectly classified as negative by the model. Specificity = true negative / (false positive + true negative), where true negative are the observations correctly identified as negative by the model, and false positive the observations incorrectly classified as positive by the model.

c)

The plot shows how many times a classification based on the selected model is better than a random choice when classifying a specific amount of observations. In particular, for the highlighted point, this means that if we want to classify 35% of the observations, if we select those to which the model gives the highest probability to be, let us say, positive, we identify 1.6 times more positive than when repeating the procedure at random.

d)

The values in the table can be computed as:

- $a_{21} = 47 * \exp(-0.96141)/(1 + \exp(-0.96141)) \approx 13$
- $a_{22} = 152 * \exp(-0.96141 + 1.11964)/(1 + \exp(-0.96141 + 1.11964)) \approx 82$
- $a_{11} = 47 - 13 = 34$
- $a_{12} = 152 - 82 = 70$

e)

The predicted response for a 40-year old man who smokes and drinks, on average, 16 ounces of alcoholic drinks per week is 0 (no oral cancer). It could be found following the path: left (drink less than 21.0312 ounces per week), left (younger than 47.5 years), right (older than 34.5 years).

Cross-validation is a procedure to estimate the prediction error, often used to select the tuning parameter(s) of a model (here the number of leaves). It consists in splitting the observations in K approximately equal-sized folds and using in turn one fold as an independent test set, in which the performance of a model (or, more generally, a method) trained on the remaining $K - 1$ folds is evaluated. The K estimates of the performance measure (e.g., deviance) are then summarized (e.g., by averaging) into a single measure. If the goal is finding the best parameter, the procedure is performed for each parameter and that corresponding to the best performance is selected.

Here cross-validation is used to select the tuning parameter. In order to do this, it is important to evaluate the results obtained with a specific value of the tuning parameter in an independent set, to avoid overfitting. Having a moderately low number of observations (less than 200, see point d), it is not reasonable to split them in separated training and test sets. Cross-validation is used because it allows to take advantage of all observations to find the tuning parameter while keeping the independence between training and test set.

There are at least 4 possible ways to prune the tree while keeping the right response for the described subject. All the following choices are correct:

Comment from Lars: I think (i) is incorrect, the "0" and "1" should be flipped. Can argue that some of the other trees should have "$\{0, 1\}$" instead of a single value.

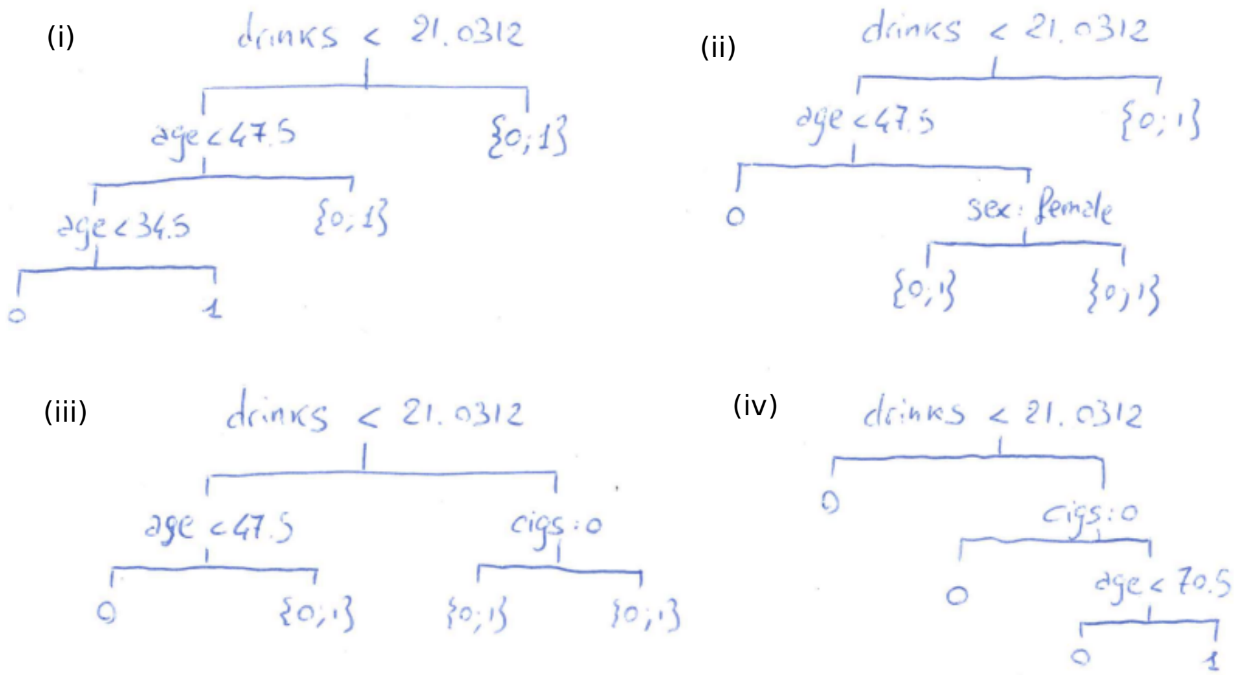


Figure 1: Different possible trees in exercise 3e).

Extra Exercise

Disclaimer, the solutions are screenshots from Vinnie Ko's solutions proposals.

Exercise 2

a)

$$\begin{aligned}
\mathrm{EPE}(f) &= \mathrm{E}[L(Y, f(\boldsymbol{X}))] = \mathrm{E}[(Y - f(\boldsymbol{X}))^2] \\
&= \int_{\boldsymbol{x}} \int_y (y - f(\boldsymbol{x}))^2 p(\boldsymbol{x}, y) dy d\boldsymbol{x} \\
&= \int_{\boldsymbol{x}} \int_y (y - f(\boldsymbol{x}))^2 p(\boldsymbol{x}) p(y|\boldsymbol{x}) dy d\boldsymbol{x} \\
&= \int_{\boldsymbol{x}} \left(\int_y (y - f(\boldsymbol{x}))^2 p(y|\boldsymbol{x}) dy \right) p(\boldsymbol{x}) d\boldsymbol{x} \\
&= \int_{\boldsymbol{x}} \left(\mathrm{E}_{Y|\boldsymbol{X}} \left[(Y - f(\boldsymbol{X}))^2 | \boldsymbol{X} = \boldsymbol{x} \right] \right) p(\boldsymbol{x}) d\boldsymbol{x} \\
&= \mathrm{E}_{\boldsymbol{X}} \left[\mathrm{E}_{Y|\boldsymbol{X}} \left[(Y - f(\boldsymbol{X}))^2 | \boldsymbol{X} = \boldsymbol{x} \right] \right]
\end{aligned}$$

We are looking for a function f that minimizes $\mathrm{EPE}(f)$ given the data (i.e. $\boldsymbol{X} = \boldsymbol{x}$). $\mathrm{EPE}(f)$ becomes

$$\mathrm{EPE}(f) = \mathrm{E}_{\boldsymbol{X}} \left[\mathrm{E}_{Y|\boldsymbol{X}=\boldsymbol{x}} \left[(Y - f(\boldsymbol{x}))^2 | \boldsymbol{X} = \boldsymbol{x} \right] \right]$$

Since all $\boldsymbol{X}$ are replaced by the given data $\boldsymbol{x}$, we can ignore $\mathrm{E}_{\boldsymbol{X}}[\cdot]$. So,

$$\mathrm{EPE}(f) = \mathrm{E}_{Y|\boldsymbol{X}=\boldsymbol{x}} \left[(Y - f(\boldsymbol{x}))^2 | \boldsymbol{X} = \boldsymbol{x} \right]$$

We are looking for a function f that minimizes this expression, which is by definition

$$f(\boldsymbol{x}) = \arg\min_c \mathrm{E}_{Y|\boldsymbol{X}=\boldsymbol{x}} \left[(Y - c)^2 | \boldsymbol{X} = \boldsymbol{x} \right].$$

b)

We want the value c that minimizes L.

$$\begin{aligned} L &= \mathrm{E}_{Y|\boldsymbol{X}=\boldsymbol{x}}\left[(Y-c)^2|\boldsymbol{X}=\boldsymbol{x}\right] \\ &= \mathrm{E}_{Y|\boldsymbol{X}=\boldsymbol{x}}\left[Y^2-2Yc+c^2|\boldsymbol{X}=\boldsymbol{x}\right] \\ &= \mathrm{E}_{Y|\boldsymbol{X}=\boldsymbol{x}}\left[Y^2|\boldsymbol{X}=\boldsymbol{x}\right] - 2c\mathrm{E}_{Y|\boldsymbol{X}=\boldsymbol{x}}\left[Y|\boldsymbol{X}=\boldsymbol{x}\right] + c^2 \end{aligned}$$

Take the first derivative.

$$\frac{\partial L}{\partial c} = -2\mathrm{E}_{Y|\boldsymbol{X}=\boldsymbol{x}}\left[Y|\boldsymbol{X}=\boldsymbol{x}\right] + 2c$$

This first derivative should equal to 0.

$$\begin{aligned} -2\mathrm{E}_{Y|\boldsymbol{X}=\boldsymbol{x}}\left[Y|\boldsymbol{X}=\boldsymbol{x}\right] + 2c = 0 \\ c = \mathrm{E}_{Y|\boldsymbol{X}=\boldsymbol{x}}\left[Y|\boldsymbol{X}=\boldsymbol{x}\right] \end{aligned}$$

Take the second derivative.

$$\frac{\partial^2 L}{\partial c^2} = 2 > 0$$

Therefore, $c = \mathrm{E}_{Y|\boldsymbol{X}=\boldsymbol{x}}\left[Y|\boldsymbol{X}=\boldsymbol{x}\right]$ is the minimizer of L.

c)

In the previous exercise, we showed that $c = \mathrm{E}_{Y|\boldsymbol{X}=\boldsymbol{x}}\left[Y|\boldsymbol{X}=\boldsymbol{x}\right]$ is the minimizer of $\mathrm{EPE}(f)$. We plug in the given expression of Y into this solution.

$$\begin{aligned} c &= \mathrm{E}_{Y|\boldsymbol{X}=\boldsymbol{x}}\left[Y|\boldsymbol{X}=\boldsymbol{x}\right] \\ &= \mathrm{E}_{Y|\boldsymbol{X}=\boldsymbol{x}}\left[g(\boldsymbol{x})+\varepsilon|\boldsymbol{X}=\boldsymbol{x}\right] \\ &= g(\boldsymbol{x}) + \mathrm{E}_{Y|\boldsymbol{X}=\boldsymbol{x}}\left[\varepsilon|\boldsymbol{X}=\boldsymbol{x}\right] \\ &= g(\boldsymbol{x}) \end{aligned}$$

So, $f(\cdot)$ is the optimal predictor when $f(\cdot) = g(\cdot)$.

d)

$$\begin{aligned}
\mathrm{EPE}(f) &= \mathrm{E}\left[(Y - f(\boldsymbol{X}))^2\right] \\
&= \mathrm{E}\left[(Y - \mathrm{E}[Y] + \mathrm{E}[Y] - f(\boldsymbol{X}))^2\right] \\
&= \mathrm{E}\left[(Y - \mathrm{E}[Y])^2 + (\mathrm{E}[Y] - f(\boldsymbol{X}))^2 + 2(Y - \mathrm{E}[Y])(\mathrm{E}[Y] - f(\boldsymbol{X}))\right] \\
&= \mathrm{E}\left[(Y - \mathrm{E}[Y])^2\right] + \mathrm{E}\left[(\mathrm{E}[Y] - f(\boldsymbol{X}))^2\right] + 2\mathrm{E}\left[(Y - \mathrm{E}[Y])(\mathrm{E}[Y] - f(\boldsymbol{X}))\right] \\
&= \mathrm{E}\left[(Y - \mathrm{E}[Y])^2\right] + \mathrm{E}\left[(\mathrm{E}[Y] - f(\boldsymbol{X}))^2\right] + 2\mathrm{E}_X\left[\mathrm{E}\left[(Y - \mathrm{E}[Y])(\mathrm{E}[Y] - f(\boldsymbol{X}))|\boldsymbol{X}\right]\right] \\
&\qquad \text{(Law of total expectation)} \\
&= \mathrm{E}\left[(Y - \mathrm{E}[Y])^2\right] + \mathrm{E}\left[(\mathrm{E}[Y] - f(\boldsymbol{X}))^2\right] + 2\mathrm{E}_X\left[(\mathrm{E}[Y] - f(\boldsymbol{X}))\mathrm{E}\left[(Y - \mathrm{E}[Y])|\boldsymbol{X}\right]\right] \\
&= \mathrm{E}\left[(Y - \mathrm{E}[Y])^2\right] + \mathrm{E}\left[(\mathrm{E}[Y] - f(\boldsymbol{X}))^2\right] \\
&= \mathrm{Var}(Y) + \mathrm{E}\left[(\mathrm{E}[Y] - f(\boldsymbol{X}))^2\right] \\
&= \mathrm{Var}(f(\boldsymbol{X}) + \varepsilon) + \mathrm{E}\left[(\mathrm{E}[Y] - f(\boldsymbol{X}))^2\right] \\
&= \mathrm{Var}(f(\boldsymbol{X})) + \mathrm{Var}(\varepsilon) + \mathrm{E}\left[(\mathrm{E}[Y] - f(\boldsymbol{X}))^2\right] \\
&= \mathrm{Var}(f(\boldsymbol{X})) + \sigma^2 + \mathrm{E}\left[(\mathrm{E}[Y] - f(\boldsymbol{X}))^2\right]
\end{aligned}$$

The last term will be 0 when $\mathrm{E}[Y] = f(\boldsymbol{X})$. So, the lower bound is $\mathrm{Var}(f(\boldsymbol{X})) + \sigma^2$.

Exercise 3

a)

This is quite straightforward.

$$\begin{aligned}
\mathrm{EPE}(f) &= \mathrm{E}[L(Y, f(\boldsymbol{X}))] = \mathrm{E}[1 - I_{\{f(\boldsymbol{x})\}}(y)] \\
&= \int_{\boldsymbol{x}} \int_y \left(1 - I_{\{f(\boldsymbol{x})\}}(y)\right) p(\boldsymbol{x}, y) dy d\boldsymbol{x} \\
&= \int_{\boldsymbol{x}} \int_y \left(1 - I_{\{f(\boldsymbol{x})\}}(y)\right) p(\boldsymbol{x}) p(y|\boldsymbol{x}) dy d\boldsymbol{x} \\
&= \int_{\boldsymbol{x}} \left(\int_y (1 - I_{\{f(\boldsymbol{x})\}}(y)) p(y|\boldsymbol{x}) dy \right) p(\boldsymbol{x}) d\boldsymbol{x} \\
&= \int_{\boldsymbol{x}} \left(1 - \Pr(Y = f(\boldsymbol{x}) | \boldsymbol{X} = \boldsymbol{x})\right) p(\boldsymbol{x}) d\boldsymbol{x}
\end{aligned}$$

b)

$$\text{EPE}(f) = \int_{\boldsymbol{x}} \{1 - \Pr(Y = f(\boldsymbol{x})|\boldsymbol{X} = \boldsymbol{x})\}\, p(\boldsymbol{x}) d\boldsymbol{x}$$

We are looking for a function f that minimizes this expression, which is by definition

$$f(\boldsymbol{x}) = \arg\min_k [1 - \Pr(Y = k|\boldsymbol{X} = \boldsymbol{x})] = \arg\max_k [\Pr(Y = k|\boldsymbol{X} = \boldsymbol{x})] \text{ where } k \in 0, 1.$$

Since $f(\boldsymbol{x})$ is a binary predictor, we have only 2 options for the value of $f(\boldsymbol{x})$: 0 and 1.
We are maximizing $\Pr(Y = k|\boldsymbol{X} = \boldsymbol{x})$. So, if $\Pr(Y = 0|\boldsymbol{X} = \boldsymbol{x}) < \Pr(Y = 1|\boldsymbol{X} = \boldsymbol{x})$, $k = 1$. And if $\Pr(Y = 0|\boldsymbol{X} = \boldsymbol{x}) > \Pr(Y = 1|\boldsymbol{X} = \boldsymbol{x})$, $k = 0$.
Notice that $\Pr(Y = 0|\boldsymbol{X} = \boldsymbol{x}) + \Pr(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = 1$. So, the decision boundary is at $\Pr(Y = 0|\boldsymbol{X} = \boldsymbol{x}) = \Pr(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = 0.5$

Therefore,

$$f(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \Pr(Y = 1|\boldsymbol{X} = \boldsymbol{x}) > 0.5 \\ 0 & \text{otherwise} \end{cases}$$

c)

Intuitively,

$$f(\boldsymbol{x}) = \begin{cases} K-1 & \text{if } K-1 = \arg\max_k [\Pr(Y = k|\boldsymbol{X} = \boldsymbol{x})] \\ K-2 & \text{if } K-2 = \arg\max_k [\Pr(Y = k|\boldsymbol{X} = \boldsymbol{x})] \\ \vdots & \vdots \\ 1 & \text{if } 1 = \arg\max_k [\Pr(Y = k|\boldsymbol{X} = \boldsymbol{x})] \\ 0 & \text{otherwise} \end{cases}$$

d)

Assume $k_{opt} : \arg\max_k [\Pr(Y = k|\boldsymbol{X} = \boldsymbol{x})]$. We get an error when $Y \neq k_{opt}$. The probability that this happens is $1 - \Pr(Y = k_{opt}|\boldsymbol{X} = \boldsymbol{x})$ which corresponds to $1 - \max_k \Pr(Y = k|\boldsymbol{x})$.