Ch9 - Support Vector Machine

Contents

9ASubsection

A.1 Maximal Margin Classifier . . .
A.2 Maximal Margin Hyperplane . . .
A.3 Construction of MMC . . .
A.4 Support Vector Classifier . . .
A.5 Support Vector Machine . . .
A.6 More Than Two Classes . . .
A.7 SVM vs Logistic-Reg . . .
A.8 Heart Data . . .
A.9 Scaling . . .
A.10 Breast Cancer . . .
A.11 Portgese Bank . . .

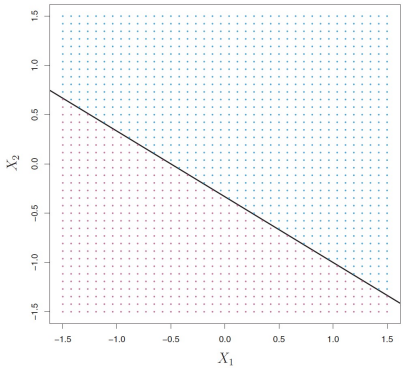
9A Subsection

[ToC]

A.1 Maximal Margin Classifier

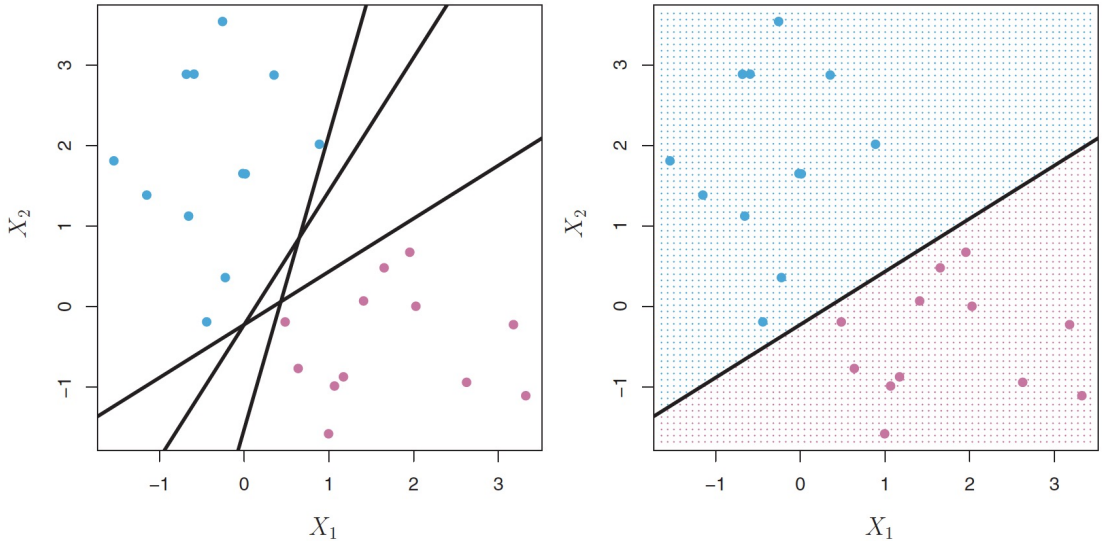
- SVM developped in 90's
- Mainly a classification method
- Maximal Margin Classifier
- Support Vector Classifier
- Support Vector Machine

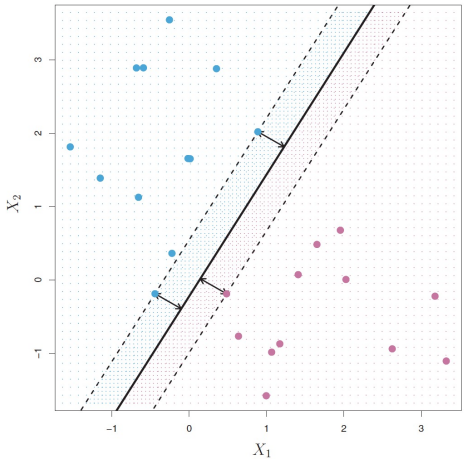
- Hyperplane (p-1 dim) in p-dimentional space.
- When p=2



A.2 Maximal Margin Hyperplane

- margin = min distance from observation to hyperplane


X_1
X_2
−1
0
1
2
3



- Three points are "Support Vectors".
- How do you construct the Maximul Margin Classifier (MMC)?

A.3 Construction of MMC

- maximize the margin M by changing β_0, β_1
- with subject to $\sum_{j=1}^{p} \beta_j^2 = 1$
- with

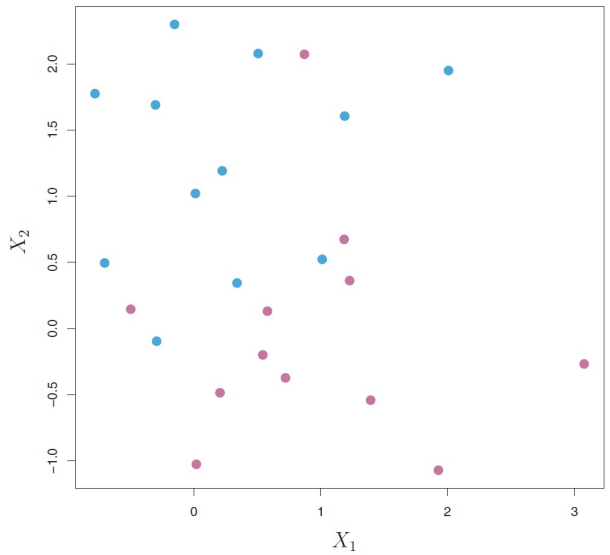
$$y_i(\beta_0 + \beta_1 x_{i1}) \geq M$$

- Second and third make sure that all obs are on the correct side of the plane, and at least M away from the hyperplane.

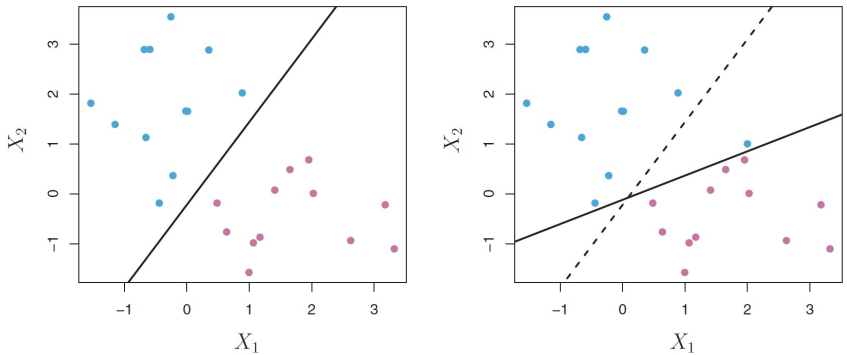
A.4 Support Vector Classifier

Non-separable case

- What if the observations were non-separable?


2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
X_2
0
1
2
3
X_1

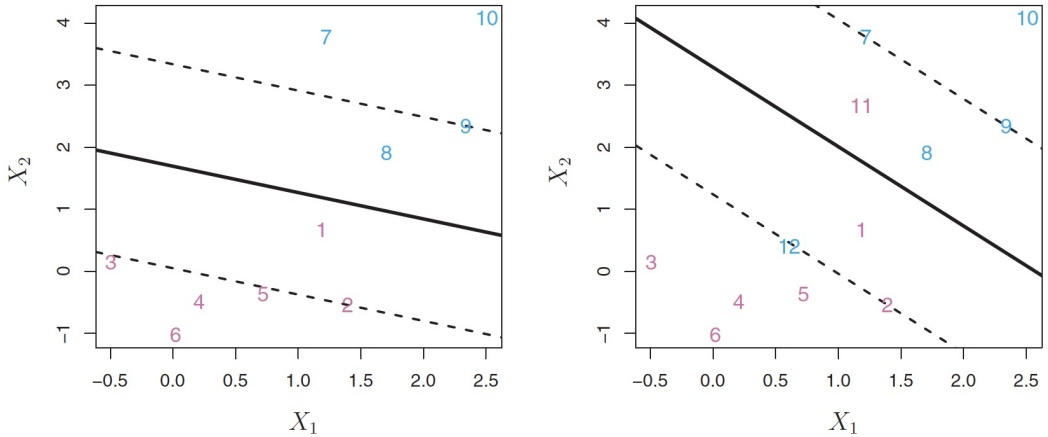
- MMC is very sensitive to a single datapoint. (Can overfit)
- No solution with $M > 0$.
- Use soft margin. (Support Vector Classifier) Miss-classifying couple of points are OK.



- maximize the margin M by changing β_0, β_1 and $\epsilon_1, \ldots, \epsilon_n$ (slack variable).
- with subject to $\sum_{j=1}^{p} \beta_j^2 = 1$
- with

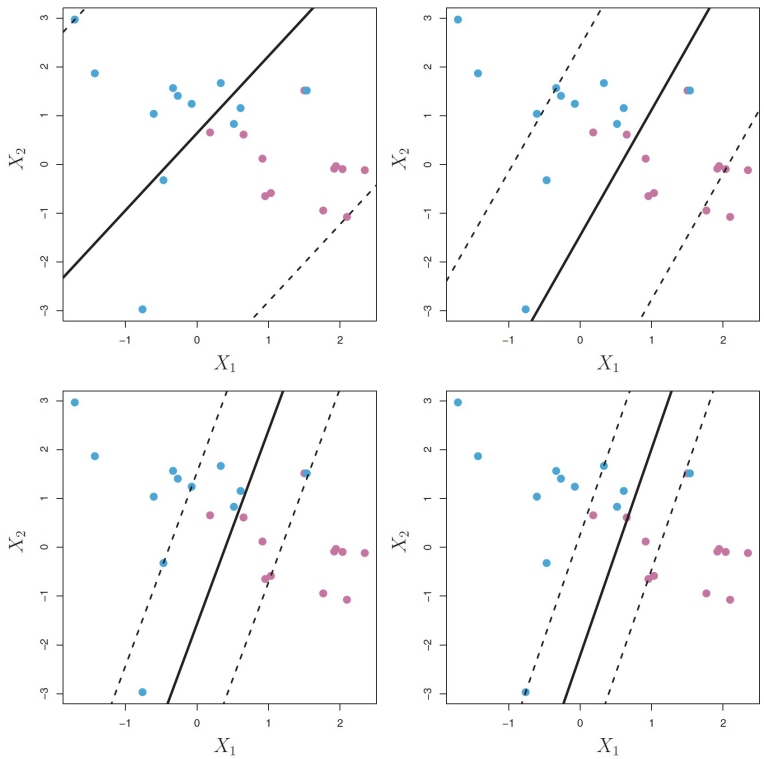
$$y_i(\beta_0 + \beta_1 x_{i1}) \geq M(1 - \epsilon_i)$$

  where $\epsilon_i \geq 0$, and $\sum_{i=1}^{n} \epsilon_i \leq C$.

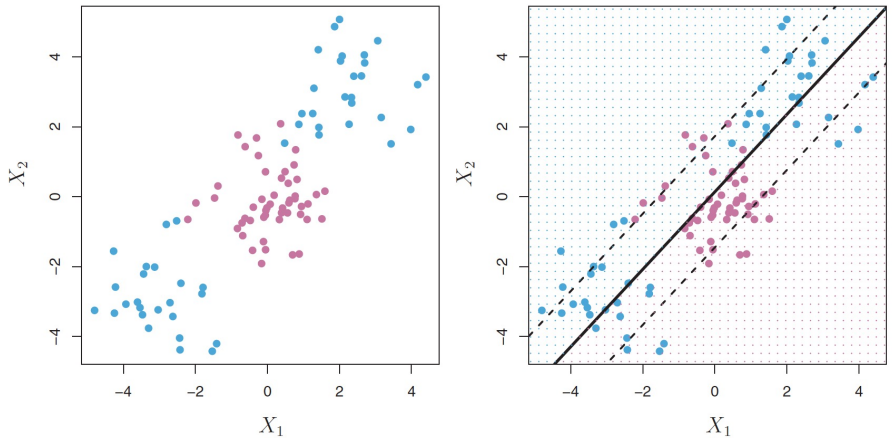


- Slack variable tells which side of hyperplane ith obs is located, relative to the margin.
- $\epsilon_i = 0$: Correct side of hyperplane, with good margin
- $\epsilon_i > 0$: Correct side of hyperplane, but within the margin (violated margin)
- $\epsilon_i > 1$: Wrong side of hyperplane

- C determines the sum of ϵ_i. It is like a budget.
- C is tuning parameter. use C-V to pick.
- If C is small, then we're looking for narrow margin that are rarely violated.
- If C is large, then we're looking for large margin that can be violated often.
- Only the obs that are on or inside of the margin affect the hyperplane.
- C controls the bias-variance trade off.


X_1
X_2

A.5 Support Vector Machine



- It turns out that when Support Vector Classifier is computed, you only need to calculate

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i \langle x, x_i \rangle$$

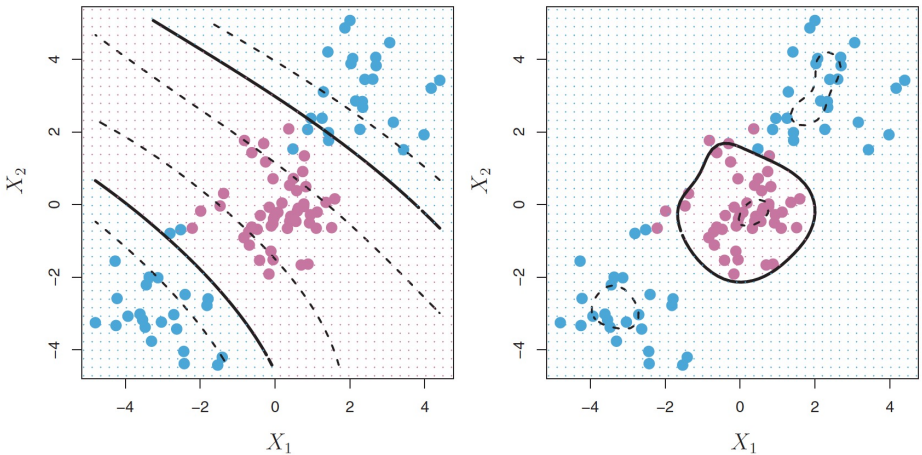
- Inner product

$$K(x_i, x_j) = \sum_{j=1}^{p} x_{ij} x_{\ell j} \quad \text{Linear Kernel}$$

- Generalize the inner product to

$$K(x_i, x_j) = \exp\Big(- \gamma \sum_{j=1}^{p} (x_{ij} - x_{\ell j})^2 \Big) \quad \text{Radial Kernel}$$

$$K(x_i, x_j) = \Big(C + \gamma \sum_{j=1}^{p} x_{ij} x_{\ell j} \Big)^d \quad \text{Polynomial Kernel}$$


X_2
X_1
−4
−2
0
2
4

A.6 More Than Two Classes

- one-versus-one approach. Construct $\binom{K}{2}$ pair-wise classifier
- Pick the one that has most frequent assignment
- one-versus-all approach. Fit K SVM. Let $\beta_{0k}, \ldots, \beta_{pk}$ denote the result. Assign observation to the class for which $\beta_{0k} + \beta_{1k}x_1 * + \cdots + \beta_{pk}x_{pk}*$ is largest.

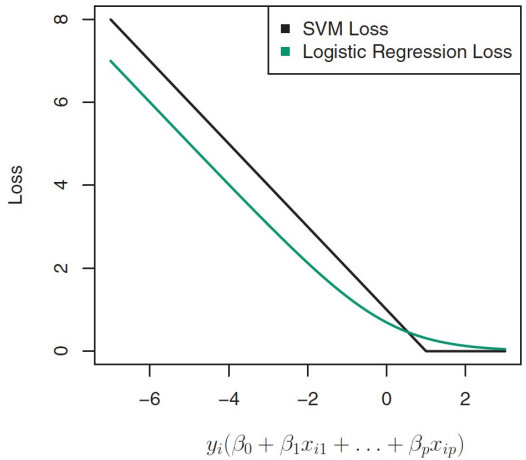
A.7 SVM vs Logistic-Reg

- It turns out SVM was actually minimizing:

$$\min_{\beta_0,\ldots,\beta_p} \left\{ \sum_{i=1}^{n} \max\left[0, 1 - y_i f(x_i)\right] + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

  were $f(x_i) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$

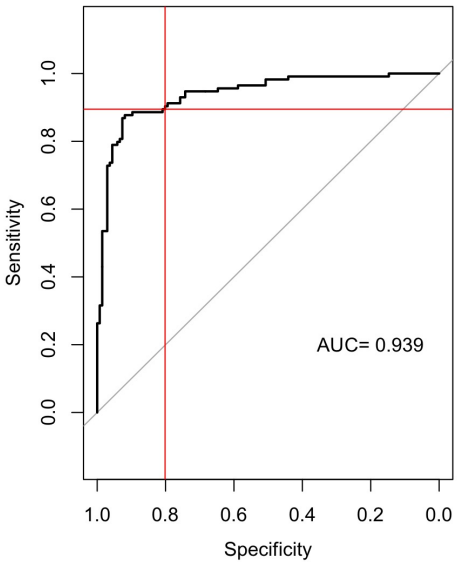
- First term is called Hinge Loss Function

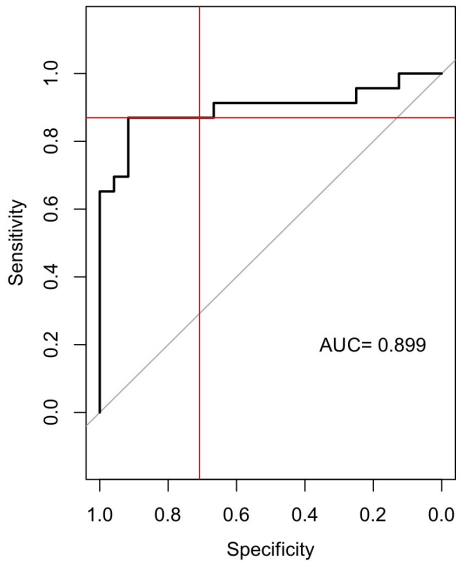

SVM Loss
Logistic Regression Loss
Loss
0
2
4
6
8
−6
−4
−2
0
2
$y_i(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})$

A.8 Heart Data

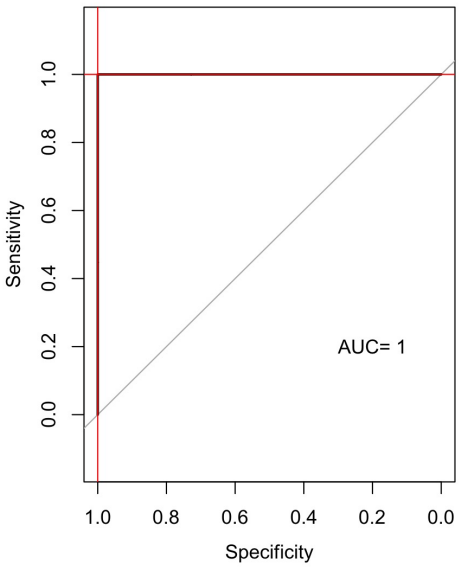
```
names(Heart)
#0  "X1"          index
#1  "Age"
#2  "Sex"
#3  "ChestPain"       Qualitative
#4  "RestBP"
#5  "Chol"
#6  "Fbs"
#7  "RestECG"
#8  "MaxHR"
#9  "ExAng"
#10 "Oldpeak"
#11 "Slope"
#12 "Ca"
#13 "Thal"     (Thalium stress test)   Qualitative
#14 "AHD"      Yes/No based on an angiographic test  Qualitative <-- Response

table(Heart2$AHD)
#  No Yes
# 160 137
```

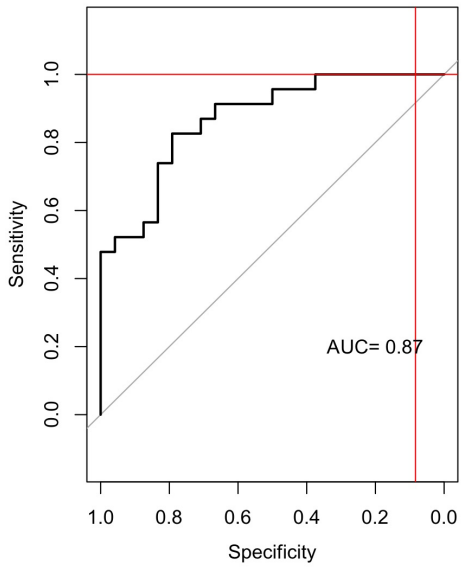
```
set.seed (my.seed)
Tuned01 = e1071::tune(svm, AHD~., data=Train.set, kernel ="linear",
                      ranges=list(
                          cost=c(0.001, 0.01, 0.1, 1, 5, 10, 100, 1000)),
              tunecontrol=tune.control(cross=5))
summary(Tuned01)

- best parameters:
 cost
 100
```


Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
AUC= 0.939
1.0
0.8
0.6
0.4
0.2
0.0
Specificity

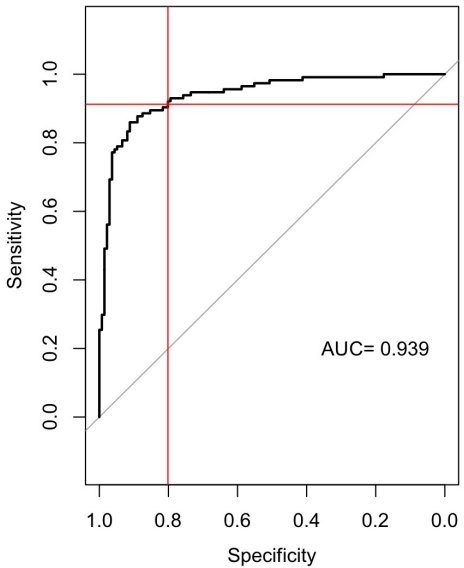

Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
AUC= 0.899
1.0
0.8
0.6
0.4
0.2
0.0
Specificity

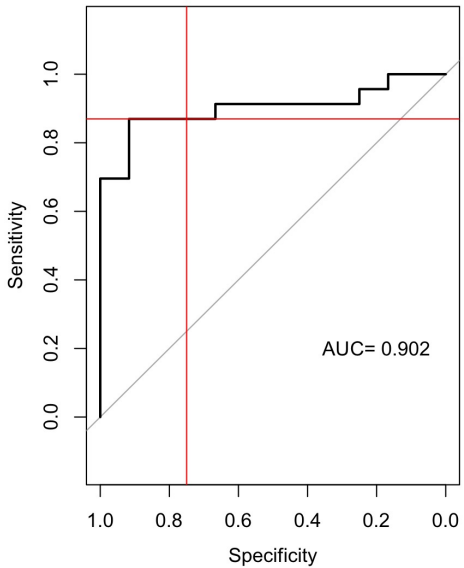
```
set.seed (my.seed)
Tuned02=e1071::tune(svm, AHD~., data=Train.set, kernel ="radial",
                    ranges=list(
                        gamma = 2^(-1:4),
                        cost=c(0.001, 0.01, 0.1, 1, 5, 10, 100, 1000)),
                    tunecontrol=tune.control(cross=5))
summary(Tuned02)


- best parameters:
 gamma cost
   0.5    5
```


Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
AUC= 1
1.0
0.8
0.6
0.4
0.2
0.0
Specificity


Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
AUC= 0.87
1.0
0.8
0.6
0.4
0.2
0.0
Specificity

```
set.seed (my.seed)
Tuned03=e1071::tune(svm, AHD~., data=Train.set, kernel ="polynomial",
                    ranges=list(
                        degree = 2^(-1:3),
                        gamma = 2^(2:5),
                        coef0=c(0, 2^(-1:2)),
                        cost=c(0.0001, 0.001, 0.01, 0.1, 1, 5, 10, 100)),
                    scale=TRUE,
                    tunecontrol=tune.control(cross=5))
summary(Tuned03)

- best parameters:
 degree gamma cost
      1    16  0.1
```


Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
AUC= 0.939
1.0
0.8
0.6
0.4
0.2
0.0
Specificity


Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
AUC= 0.902
1.0
0.8
0.6
0.4
0.2
0.0
Specificity

A.9 Scaling

A.10 Breast Cancer

A.11 Portgese Bank

Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani