Lecture 1 exercise 1

Statistical Learning, 2011

Niels Richard Hansen
September 19, 2011

1 Solution of lecture exercise 1

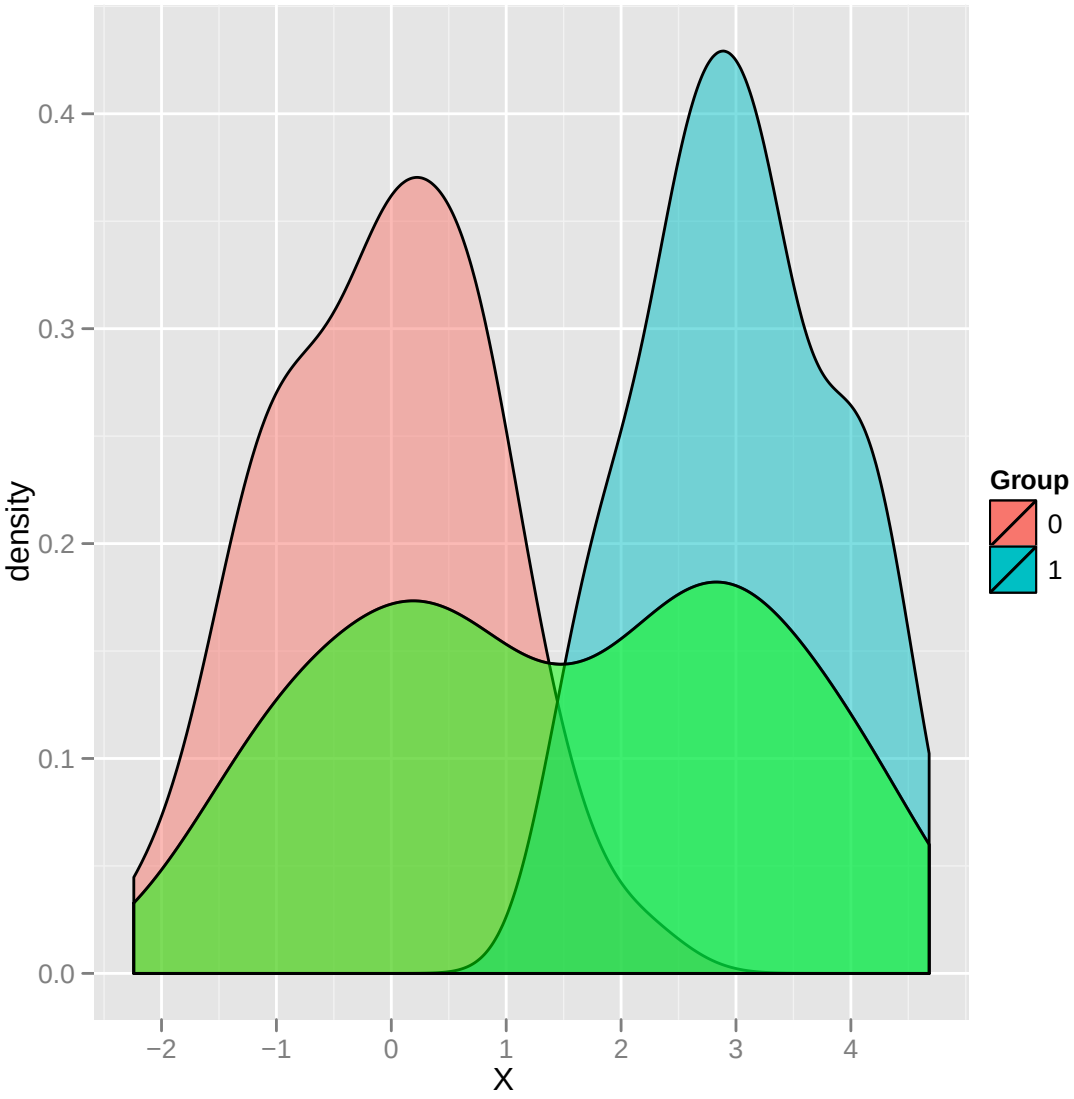
The simulation can be done as follows.

```
> N <- 100
> Y <- rbinom(N, 1, 0.5)
> mu <- c(0, 3)[Y + 1]
> X <- rnorm(N, mu, 1)
```

This is just one way, where we use that `rnorm` take a vector as mean values.

The following density plot is produced using ggplot2. An alternative is provided if ggplot2 is not installed. It shows the empirical marginal distribution of X and the empirical conditional distributions of X divided according to the two groups.

```
> if(require(ggplot2)) {
+   p <- qplot(X, geom = "density") +
+     geom_density(aes(fill = factor(Y)), alpha = 0.5) +
+       geom_density(fill = alpha("green", 0.5)) +
+         scale_fill_discrete("Group")
+   print(p)
+ } else {
+   breaks <- pretty(X, 12)
+   hist(X[Y == 1], breaks, freq = FALSE, ylim = c(0, 0.8),
+        xlim = range(breaks), col = "red", main = "Histogram",
+        xlab = "X")
+   hist(X[Y == 0], breaks, freq = FALSE, col = "blue", add = TRUE)
+   hist(X, breaks, freq = FALSE, add = TRUE, col = rgb(0, 1, 0, 0.7))
+ }
```

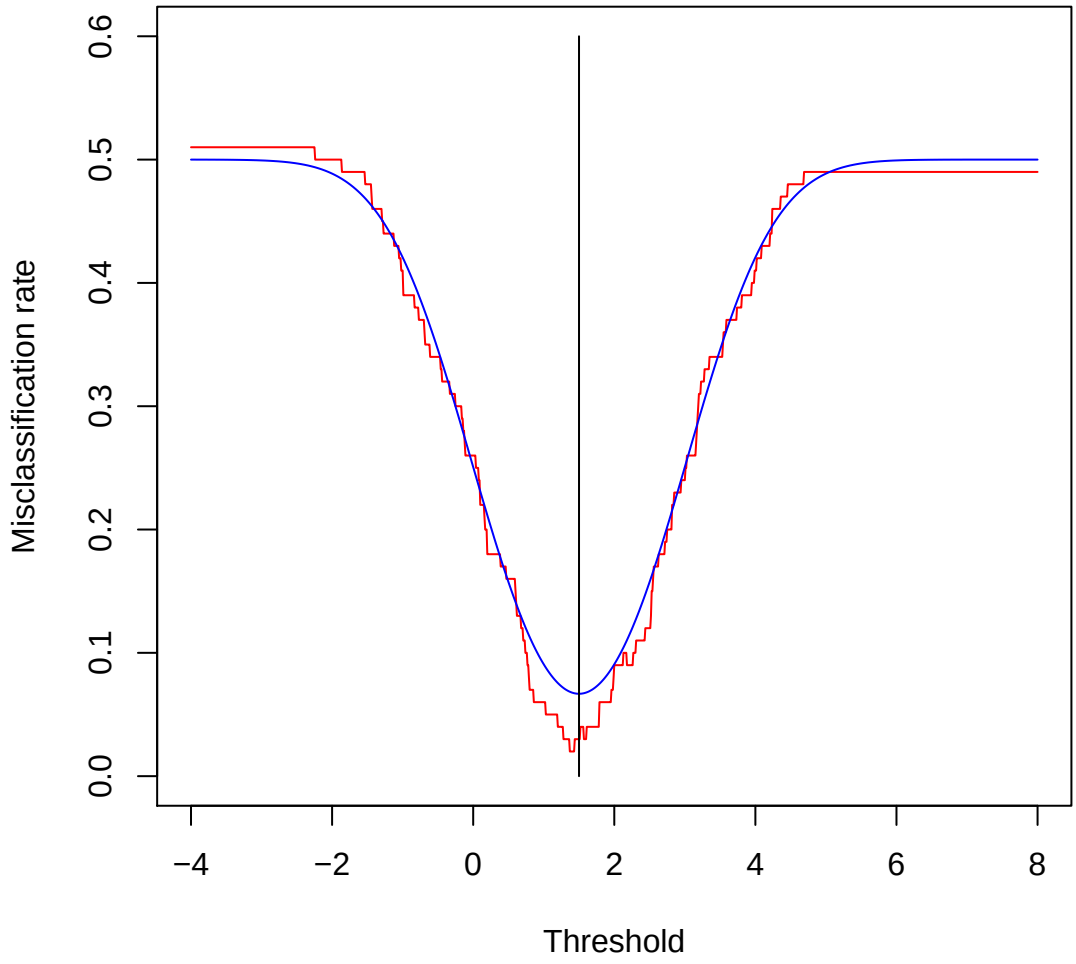


We then compute, for a grid of threshold values, the average prediction error for each of these threshold values. This is done using the `sapply` function. Regarding the theoretical computation we note that, with Φ_μ the distribution function for the normal distribution with mean μ and variance 1,

$$\begin{aligned} \mathrm{EPE}(f_t) &= P(X < t \mid Y = 1)/2 + P(X \geq t \mid Y = 0)/2 \\ &= (\Phi_3(t) + 1 - \Phi_0(t))/2. \end{aligned}$$

```
> thres <- seq(-4, 8, 0.01)
> aveMisclas <- function(t) mean((X >= t) != Y)
> misclasEmp <- sapply(thres, aveMisclas)
> theoMisclas <- function(t) (pnorm(t, 3, 1) + 1 - pnorm(t))/2
> misclasTheo <- theoMisclas(thres)
> plot(thres, misclasEmp, type = "l", col = "red", ylim = c(0, 0.6),
```

```
+      xlab = "Threshold", ylab = "Misclassification rate")
> lines(thres, misclasTheo, type = "l", col = "blue")
> lines(c(1.5, 1.5), c(0, 0.6))
```



With ϕ_μ the density for the normal distribution with mean μ and variance 1, the Bayes classifier is found by solving

$$\phi_3(t) = \phi_0(t)$$

which amounts to $t = 3/2$ (the midpoint between the two means). The resulting classifier, which is the Bayes classifier, is thus $f_{3/2}$. The Bayes rate is

$$(\Phi_3(3/2) + 1 - \Phi_0(3/2))/2 = 0.0668$$

N.B. Note that optimization of EPE over f_t can be done by equating the derivative to 0, which yields the equation above. However, finding this optimum does not guarantee that we have found the Bayes classifier in general, as the Bayes classifier need not be of the form

f_t for any t. In general, for a two-class classification problem the equation

$$\pi_0 g_0(x) = \pi_1 g_1(x)$$

characterizes the boundary for the Bayes classifier between the values of x, which are classified as ones, and the values of x, which are classified as zeroes.