

Lecture 1 exercise 2

Statistical Learning, 2011

Niels Richard Hansen
September 19, 2011

1 Solution of lecture exercise 2

1.1 Question 1

Having determined that the Bayes classifier is of the form f_t for some t , one idea for estimating it is to estimate t by minimizing the empirical misclassification rate. We repeat the simulation here.

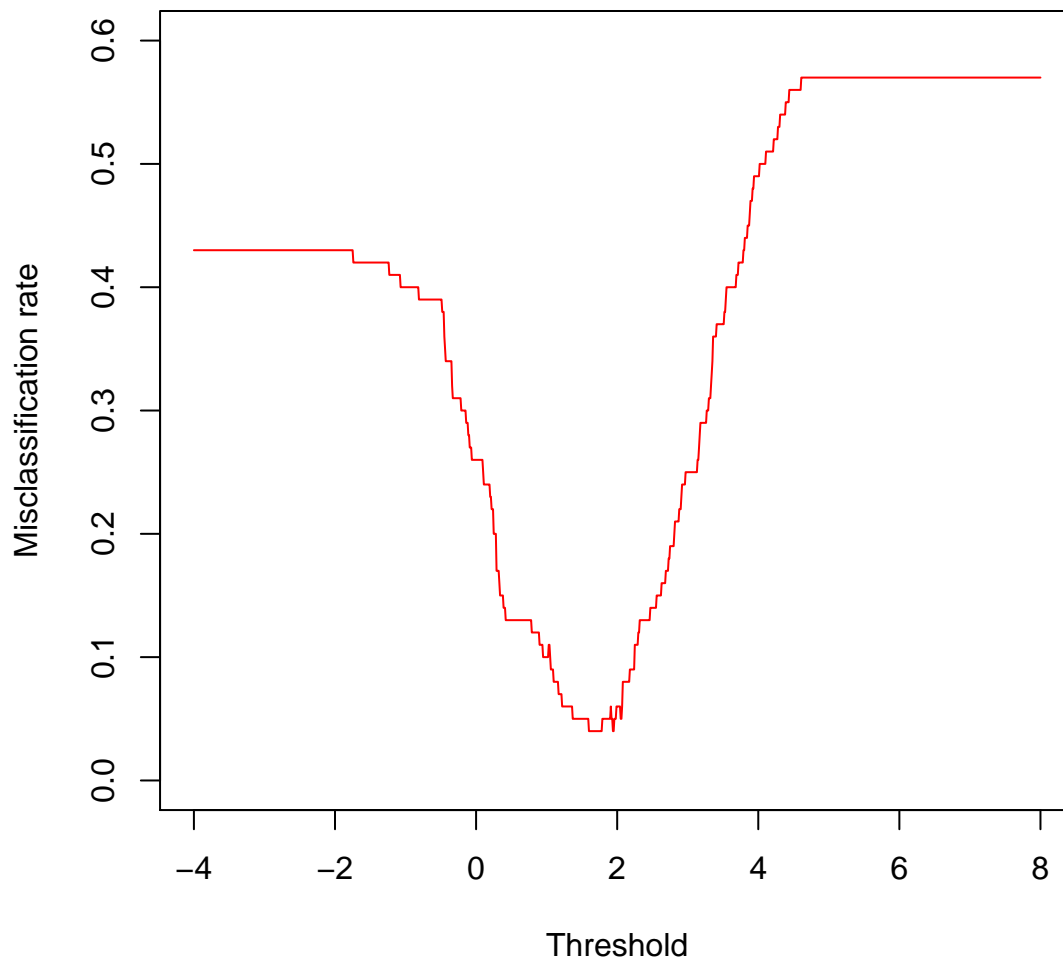
```
> N <- 100
> Y <- rbinom(N, 1, 0.5)
> mu <- c(0, 3)[Y + 1]
> X <- rnorm(N, mu, 1)
```

The minimization can be done naively in this case by evaluating the empirical misclassification rate on a grid and choose the minimizer.

```
> thres <- seq(-4, 8, 0.01)
> aveMisclas <- function(t) mean((X >= t) != Y)
> misclasEmp <- sapply(thres, aveMisclas)
> thres[which.min(misclasEmp)]
```

```
[1] 1.6
```

```
> plot(thres, misclasEmp, type = "l", col = "red", ylim = c(0, 0.6),
+       xlab = "Threshold", ylab = "Misclassification rate")
```



In the alternative, we can use a more systematic approach and use a general purpose optimizer with the empirical misclassification rate as the objective function.

```
> optimize(aveMisclas, c(-4, 8))
```

```
$minimum
[1] 1.634895
```

```
$objective
[1] 0.04
```

The latter approach may actually not work correctly (it may easily fail to find the true minimum), because the function is not smooth. In fact, it is discontinuous.

An alternative approach is to estimate the parameters in the normal distributions instead and then compute the resulting Bayes classifier for the estimated parameters.

```
> t <- (mean(X[Y == 1]) + mean(X[Y == 0]))/2
> t
```

```
[1] 1.696435
```

This solution is well behaved computationally, but it does rely on model assumptions.

1.2 Question 2

We can observe that with two classes we have the general formula for the conditional probability

$$\begin{aligned} P(Y = 1 \mid X = x) &= \frac{\pi_1 g_1(x)}{\pi_0 g_0(x) + \pi_1 g_1(x)} \\ &= \frac{h(x)}{1 + h(x)} \end{aligned}$$

where

$$h(x) = \frac{\pi_1 g_1(x)}{\pi_0 g_0(x)}.$$

If the g_k 's are densities for the normal distribution with equal variance σ^2 and mean values μ_k we get that

$$\begin{aligned} \log h(x) &= \log \frac{\pi_1}{\pi_0} + \frac{(x - \mu_0)^2 - (x - \mu_1)^2}{2\sigma^2} \\ &= \log \frac{\pi_1}{\pi_0} + \frac{\mu_0^2 - \mu_1^2 + 2(\mu_1 - \mu_0)x}{2\sigma^2} \\ &= \alpha + \beta x \end{aligned}$$

with $\alpha = \log \frac{\pi_1}{\pi_0} + \frac{\mu_0^2 - \mu_1^2}{2\sigma^2}$ and $\beta = \frac{(\mu_1 - \mu_0)}{\sigma^2}$.

This conditional model, with h as given above, is usually called the *logistic regression model*. Note that for this model the conditional probability is 0.5 for

$$x = -\frac{\alpha}{\beta},$$

which is thus the threshold for the Bayes classifier. For $\pi_0 = \pi_1 = 0.5$

$$-\frac{\alpha}{\beta} = -\frac{\mu_0^2 - \mu_1^2}{2(\mu_1 - \mu_0)} = \frac{(\mu_0 - \mu_1)(\mu_0 + \mu_1)}{2(\mu_0 - \mu_1)} = \frac{\mu_0 + \mu_1}{2}$$

as it is supposed to.

We can fit this conditional model as a logistic regression model using the `glm` function.

```
> alphabeta <- coefficients(glm(Y ~ X, family = binomial))
> t <- -alphabeta[1]/alphabeta[2]
> t
```

```
(Intercept)
1.709512
```

Bonus question: Construct a simulation study where you simulate $M = 200$ data sets and fit the threshold t for the Bayes classifier by the three different methods described above. Compare the resulting distributions.

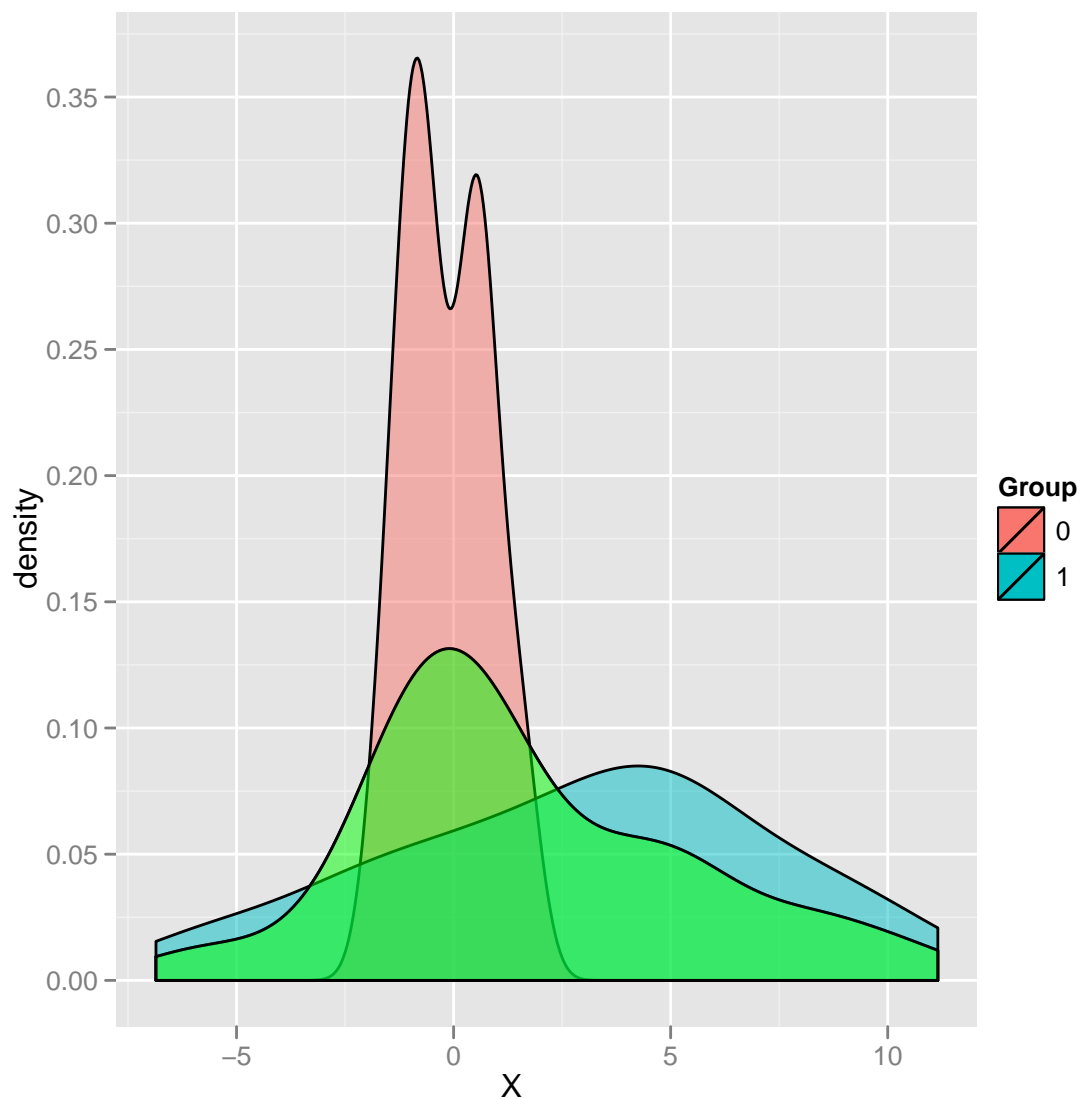
1.3 Question 3

With unequal variances the simulation can be done as follows.

```
> N <- 100
> Y <- rbinom(N, 1, 0.5)
> mu <- c(0, 3)[Y + 1]
> sd <- c(1, 5)[Y + 1]
> X <- rnorm(N, mu, sd)
```

We take a look at the density plots again. An alternative is provided if ggplot2 is not installed. It shows again 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))
+ }
```



What should be noted is that large negative values also come from group 1.

The way of computing the average prediction errors for f_t does not change, and the theoretical computation yields

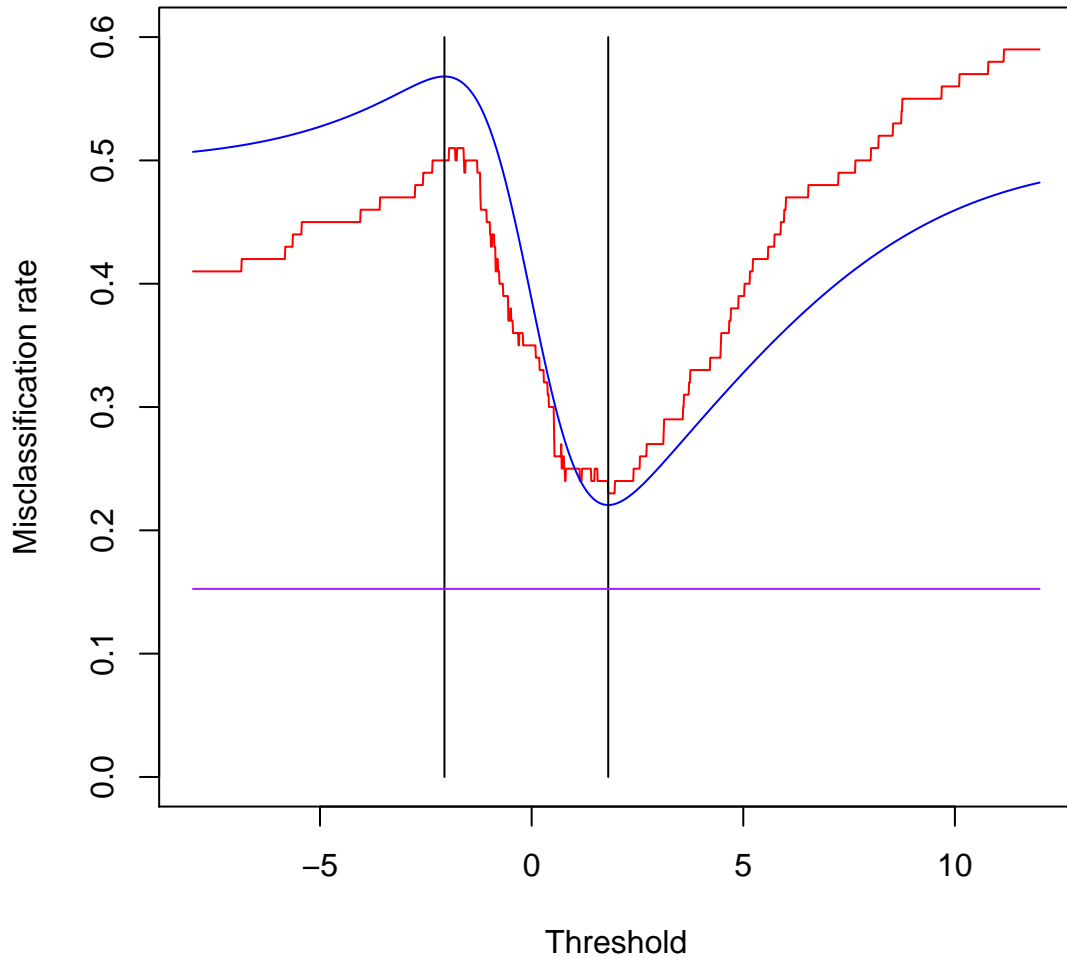
$$\text{EPE}(f_t) = (\Phi_{3,5}(t) + 1 - \Phi_{0,1}(t))/2.$$

```
> thres <- seq(-8, 12, 0.01)
> aveMisclas <- function(t) mean((X >= t) != Y)
> misclasEmp <- sapply(thres, aveMisclas)
> theoMisclas <- function(t) (pnorm(t, 3, 5) + 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")
> z <- (-6 + c(-1,1)*sqrt(36+24*4*(9+50*log(5))))/48
> lines(c(z[1], z[1]), c(0, 0.6))
> lines(c(z[2], z[2]), c(0, 0.6))
> BayesRate <- (pnorm(z[2], 3, 5) - pnorm(z[1], 3, 5) +
+               pnorm(z[1]) + 1 - pnorm(z[2]))/2
> lines(c(-8, 12), c(BayesRate, BayesRate), col = "purple")

```



Obviously there is a minimizer, which could be found by solving

$$\phi_{3,5}(t) = \phi_{0,1}(t)$$

which amounts to $(t - 3)^2 + 50 \log(5) = 25t^2$ (be careful to choose the minimizer). The resulting classifier is, however, not the Bayes classifier. The Bayes classifier is given by

$$f_B(x) = 1(t_{\text{lower}} \leq x \leq t_{\text{upper}})$$

with t_{lower} and t_{upper} the two solutions of the quadratic equation above (-2.0598 and 1.8098 , respectively). The Bayes rate is

$$(\Phi_{3,5}(t_{\text{upper}}) - \Phi_{3,5}(t_{\text{lower}}) + \Phi_{0,1}(t_{\text{lower}}) + 1 - \Phi_{0,1}(t_{\text{upper}}))/2 = 0.1525$$