Linear discriminant analysis

Igor Adamiec

Mały background

W regresji logistycznej modelowaliśmy prawdopodobieństwo, ze dana obserwacja X należy do klasy k Pr(k | X).

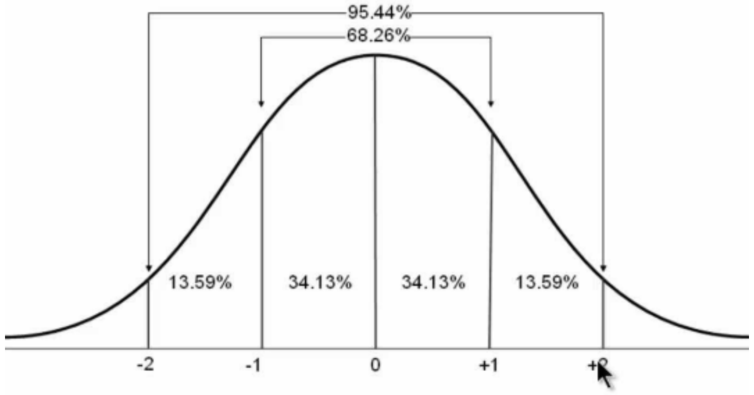
W LDA modelujemy rozkłady predyktorów X osobno dla każdej z klas.

Następnie używamy twierdzenia Bayesa żeby przekształcić to w Pr(k |X) (tak jak w regresji logistycznej).

Gdy zakłada się, że rozkłady badanych zmiennych są normalne, to model jest bardzo podobny do modelu log-reg.

[Normal distribution - clearly explained]

[What is statistical distribution?]



Czemu nie log-reg?

- Gdy klasy są od siebie dobrze odseparowane, to estymacje parametrów dla modelu regresji logistycznej są zaskakująco niestabilne (?). LDA nie ma tego problemu,
- Gdy liczba obserwacji jest mała, a rozkład predyktorów jest normalny dla każdej z klas, to LDA jest bardziej stabilne,
- LDA jest popularna gdy w zbiorze jest więcej niż 2 klasy.

Twierdzenie Bayesa dla klasyfikacji

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}.$$

1. Przyjmijmy, że obserwacje możemy zakwalifikować do K różnych klas (K≥2),
2. π_k to prawdopodobieństwo a priori, że dana obserwacja należy do k-tej klasy,
3. $f_k(x) \equiv \Pr(X = x|\, Y = k)$ – ten potrójny znak równości oznacza, że lewa strona jest tożsama z prawą czyli równanie jest zawsze spełnione. Funkcja fk(x) oznacza funkcję gęstości obserwacji X, która pochodzi z k-tej klasy,

4. Fk(x) jest duże jeżeli prawdopodobieństwo, że dana obserwacja należy do k-tej klasy jest wysokie (i na odwrót),
5. Pi k liczy się bardzo prosto – $\pi_k = \frac{liczba\ elementów\ należących\ do\ k{-}tej\ klasy}{liczba\ wszystkich\ elementów}$
6. pk(x) to nasze prawdopodobieństwo a posteriori – Pr(Y = k|X)

LDA dla p = 1

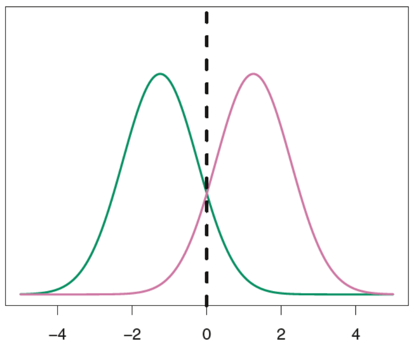
Zakładamy, że rozkład fk(x) jest normalny, więc funkcja gęstości wygląda tak:

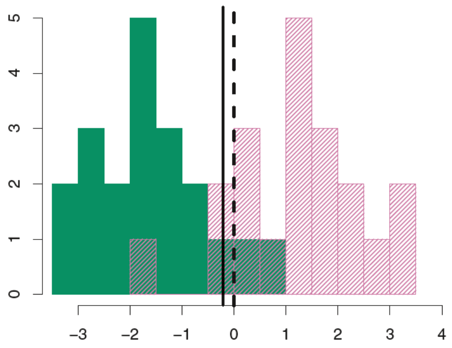
$$f_k(x) = \underbrace{\frac{1}{\sqrt{2\pi}\sigma_k}}_{\text{Odchylenie standardowe}} \exp\left(-\frac{1}{2\underbrace{\sigma_k^2}_{\text{wariancja}}}(x - \overbrace{\mu_k}^{\text{średnia}})^2\right)$$

W LDA zakładamy, że wariancja dla każdej z klas jest taka sama: $\sigma_1^2 = \cdots = \sigma_K^2$
(różne wariancje zakładamy w QDA).
Po podstawieniu funkcji gęstości do wzoru twierdzenia Bayesa z poprzedniego slajdu, wszystko wygląda tak:

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_k)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x - \mu_l)^2\right)}.$$

Funkcje gęstości i histogramy dla dwóch klas. Przerywana linia to podział na podstawie twierdzenia Bayesa, a linia ciągła na prawym wykresie to podział LDA.
Różnica wynika z tego, że LDA jest estymowane tylko na podstawie danych, które mamy – próbki z populacji.





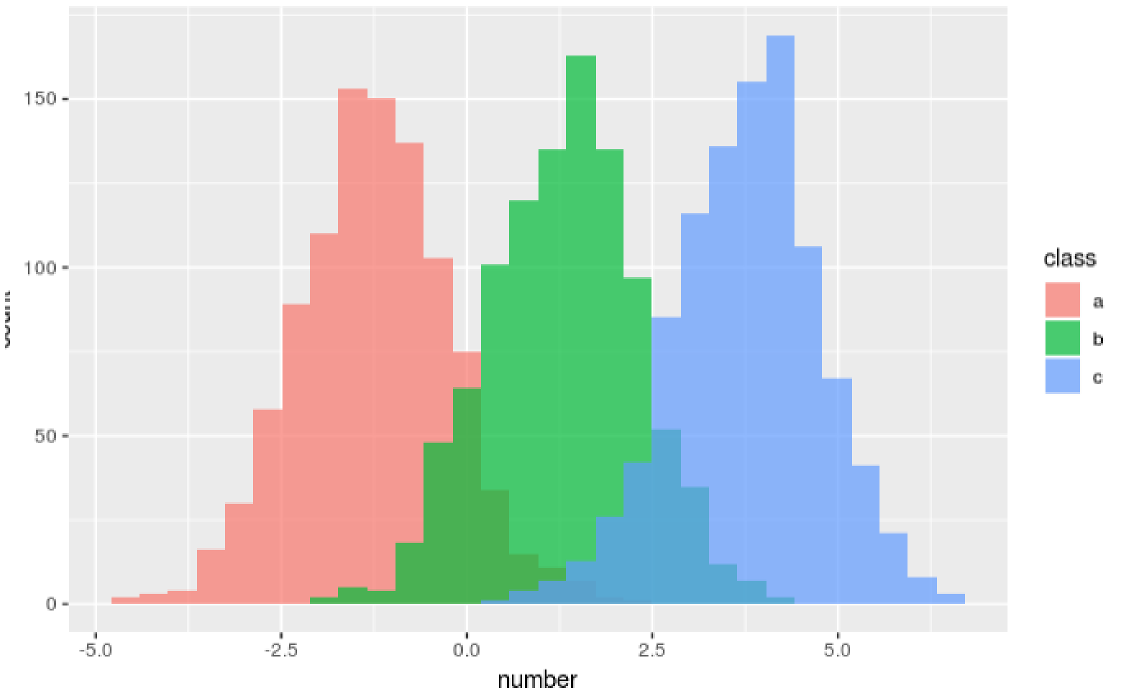
Ciąg dalszy matmy:
Po zlogarytmowaniu wzoru z poprzedniego slajdu i uproszczeniu wzoru można zauważyć, że obserwacja zostanie przydzielona do klasy, dla której poniższa wartość jest największa

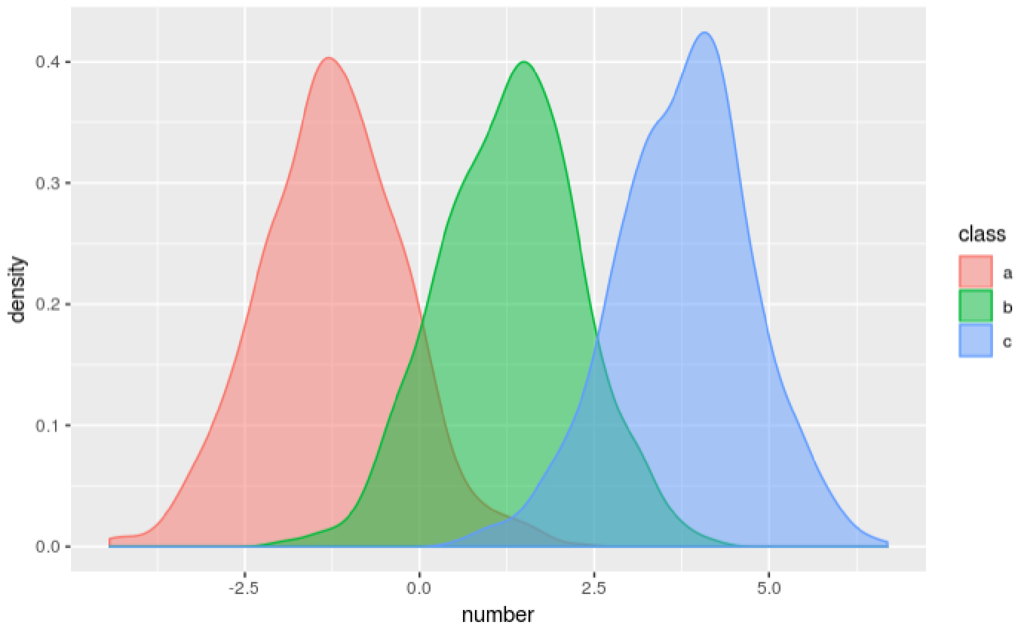
$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

W przypadku gdy prawdopodobieństwa pk są takie same dla każdej z klas, to barierą dla klasyfikatora Bayesa jest wzór”

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}.$$

Zrobiłem zbiór danych składający się z dwóch zmiennych: klasy i liczby z rozkładu normalnego.
Klasa a: średnia – 1,25,
Klasa b: średnia 1,25,
Klasa c: średnia 3,75.
Każda z klas liczy po 1000 obserwacji, a ich wariancje wynoszą 1.



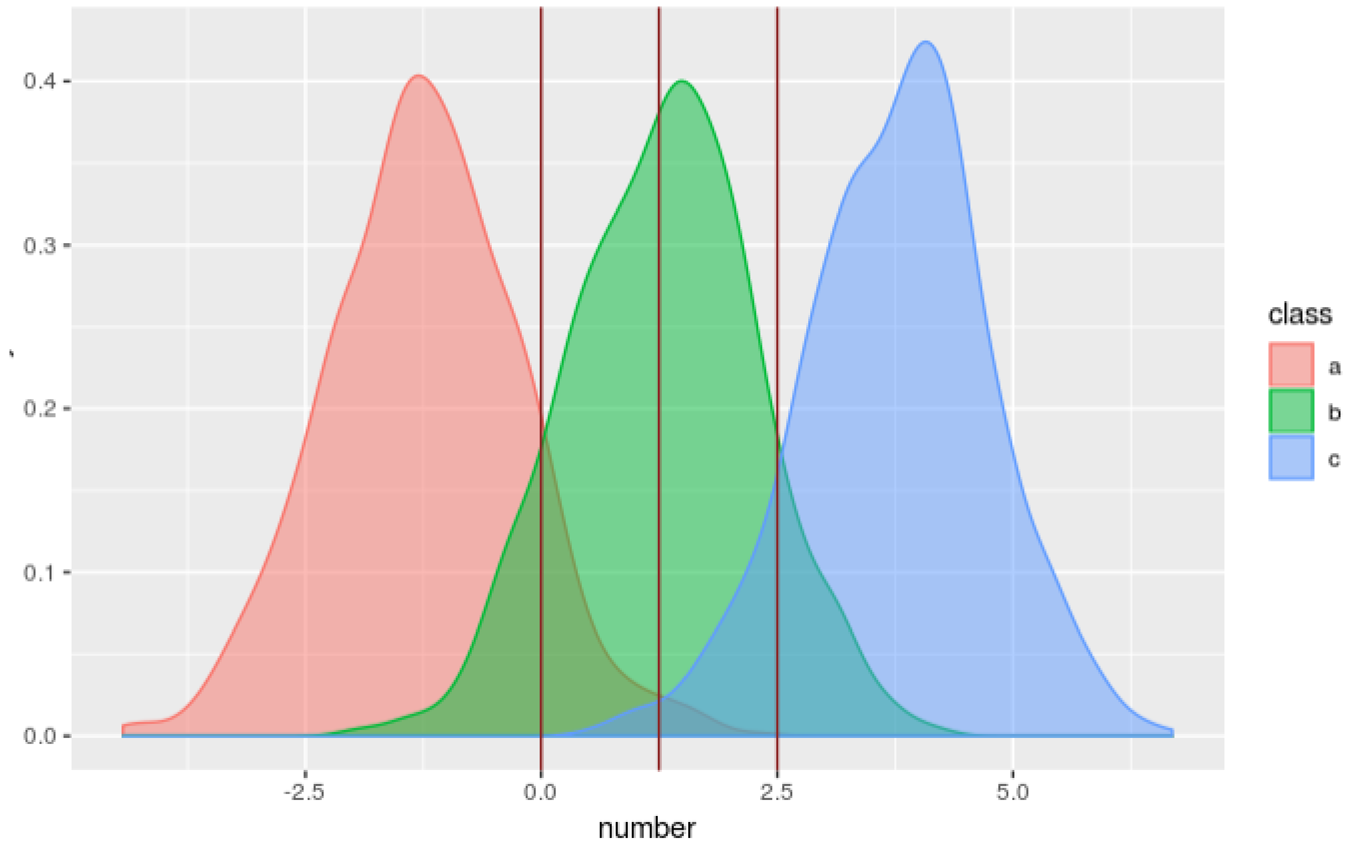


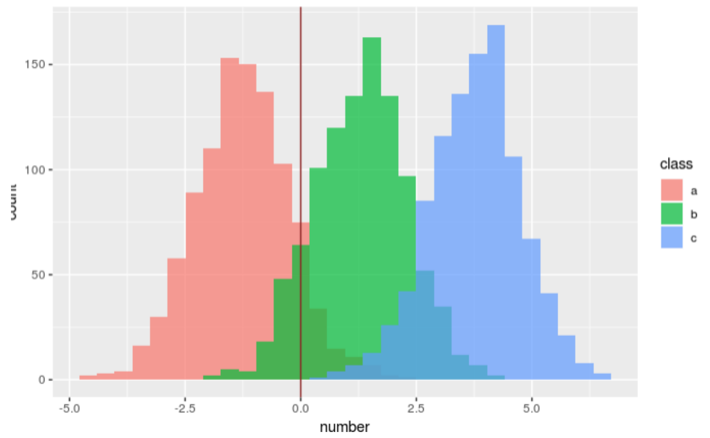
Zgodnie ze wzorem z poprzedniego slajdu, punkty podziału wynoszą:

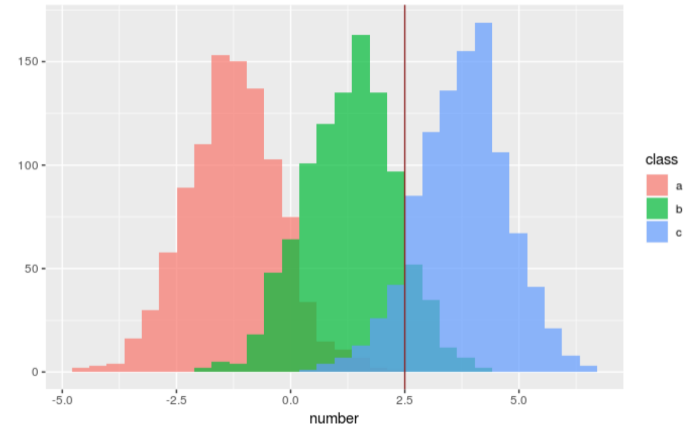
a) pomiędzy klasą a i b: (-1,25 + 1,25)/2 = 0,

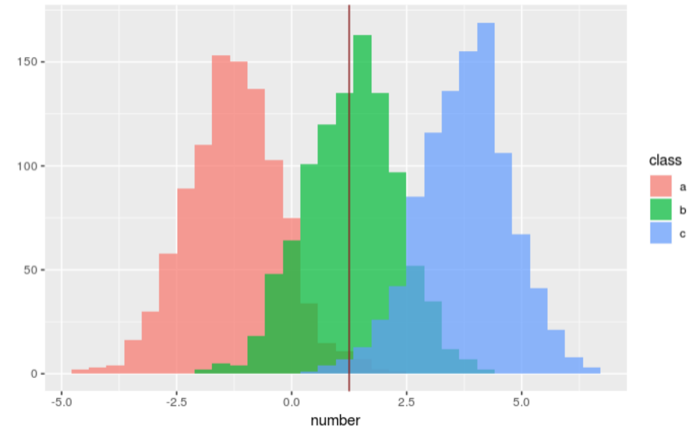
b) Pomiędzy klasą b i c: (1,25 + 3,75)/2 = 2,5,

c) Pomiędzy klasą a i c: (-1,25 + 3,75)/2 = 1,25









Estymacja parametrów

Jako, że nasze dane nigdy nie są całą populacją, a jedynie jej próbką, to wszystkie potrzebne parametry musimy wyestymować.

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

$$\hat{\pi}_k = n_k/n.$$

$$\hat{\delta}_k(x) = x \cdot \frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k)$$

```{r}
abc %>% group_by(class) %>% summarise(mean(number))
```

class <chr>	mean(number) <dbl>
a	-1.248279
b	1.292368
c	3.770710

3 rows

```
[1] 0.9963902
```

```{r}
(abc_pred <- abc %>%
  mutate(lda_a = (number *(srednia_a/var))-((srednia_a^2)/(2*var))+log(1000/3000),
         lda_b = (number *(srednia_b/var))-((srednia_b^2)/(2*var))+log(1000/3000),
         lda_c = (number *(srednia_c/var))-((srednia_c^2)/(2*var))+log(1000/3000),
         predicted = case_when(lda_a > lda_b & lda_a > lda_c ~ "a",
                               lda_b > lda_a & lda_b > lda_c ~ "b",
                               TRUE ~ "c")))
```

class <chr>	number <dbl>	lda_a <dbl>	lda_b <dbl>	lda_c <dbl>	predicted <chr>
a	-0.570526355	-1.165778911	-2.67674657	-10.3925781	a
a	-1.660927260	0.200276835	-4.09105117	-14.5190595	a
a	-0.438126578	-1.331649531	-2.50501743	-9.8915282	a
a	-1.691136453	0.238122953	-4.13023401	-14.6333822	a
a	-3.068781465	1.964038509	-5.91710861	-19.8469019	a
a	0.287367152	-2.240549068	-1.56401572	-7.1459909	b
a	-0.650445956	-1.065655524	-2.78040630	-10.6950235	a
a	-1.629817435	0.161302404	-4.05070017	-14.4013283	a
a	-3.348377878	2.314317277	-6.27975916	-20.9049984	a
a	-3.003063972	1.881707543	-5.83186972	-19.5982025	a

1-10 of 3,000 rows

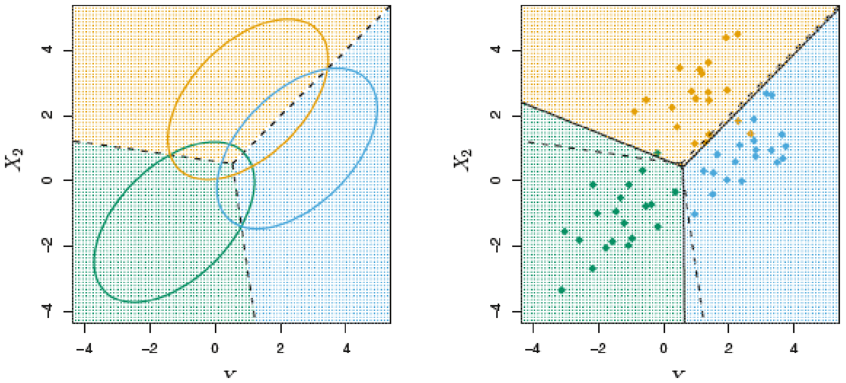
Previous 1 2 3 4 5 6

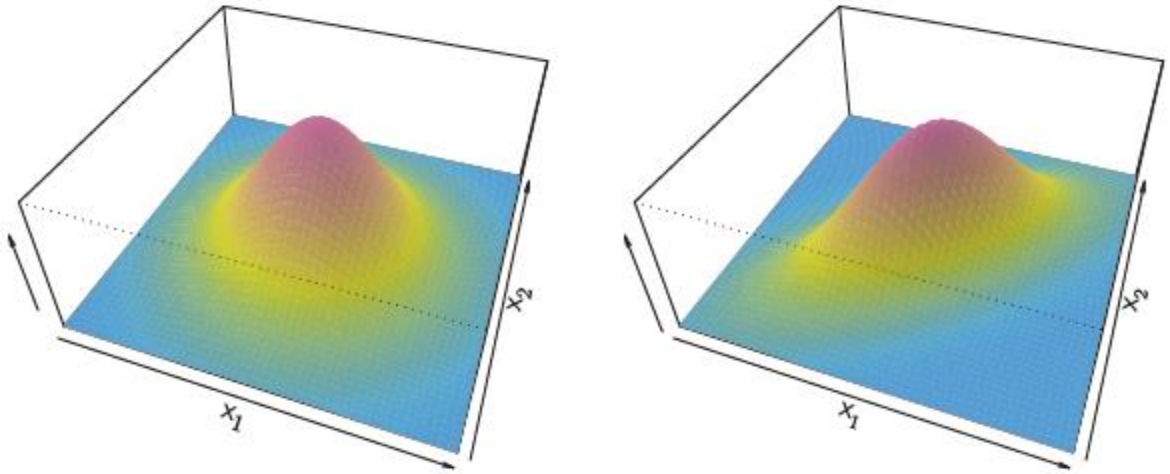
```{r}
mean(abc_pred$class !=abc_pred$predicted)
```

```
[1] 0.131
```

LDA dla P > 1

- Zakładamy, że zmienne niezależne pochodzą z rozkładu normalnego wielowymiarowego – takiego, w którym każda ze zmiennych pochodzi z rozkładu normalnego i są ze sobą jakoś skorelowane





Funkcja gęstości rozkładu wielowymiarowego normalnego

$$f(x) = \frac{1}{(2\pi)^{p/2}|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \mathbf{\Sigma}^{-1}(x-\mu)\right).$$

To nie jest suma – to jest macierz kowariancji pomiędzy zmiennymi

Średnia ze wszystkich zmiennych dla obserwacji

$$\delta_k(x) = x^T \mathbf{\Sigma}^{-1}\mu_k - \frac{1}{2}\mu_k^T \mathbf{\Sigma}^{-1}\mu_k + \log \pi_k$$

Sposoby oceniania klasyfikacji

Confussion matrix

Załóżmy, że „a” to 0, a „b” to 1

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

```
```{r}
table(actual = abc_pred$class, predicted = abc_pred$predicted)
```

      predicted
actual   a   b
     a 879 121
     b 126 874
```

sensivity – 874/(874+126) – TP/(TP +FN)

Specificity – 879/(879+126) – TN/(TN+FN)

Ustalanie granicy podziału (thresholdu)

Przyjęło się, że tak jak w klasyfikatorze Bayesowskim, naszą granicę podziału ustalamy na prawdopodobieństwo równe 0,5.
Możemy jednak dowolnie wybrać tę liczbę.

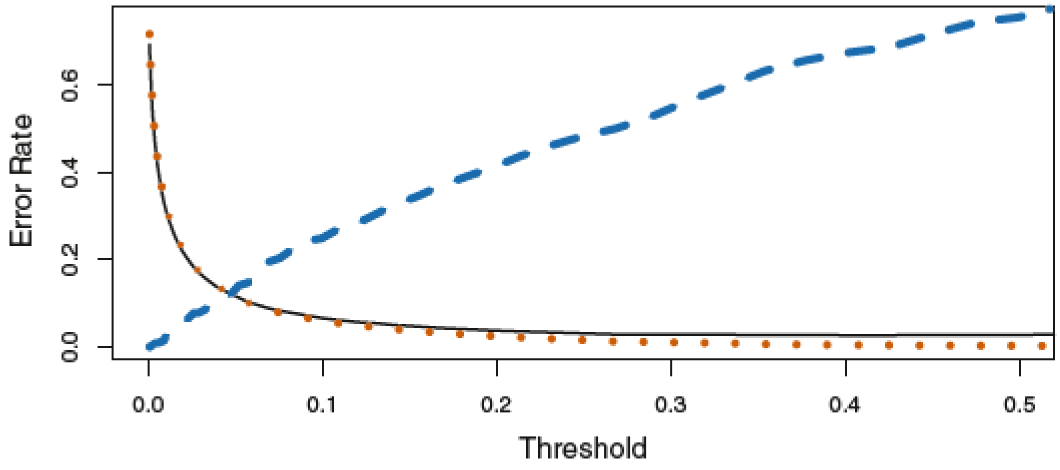
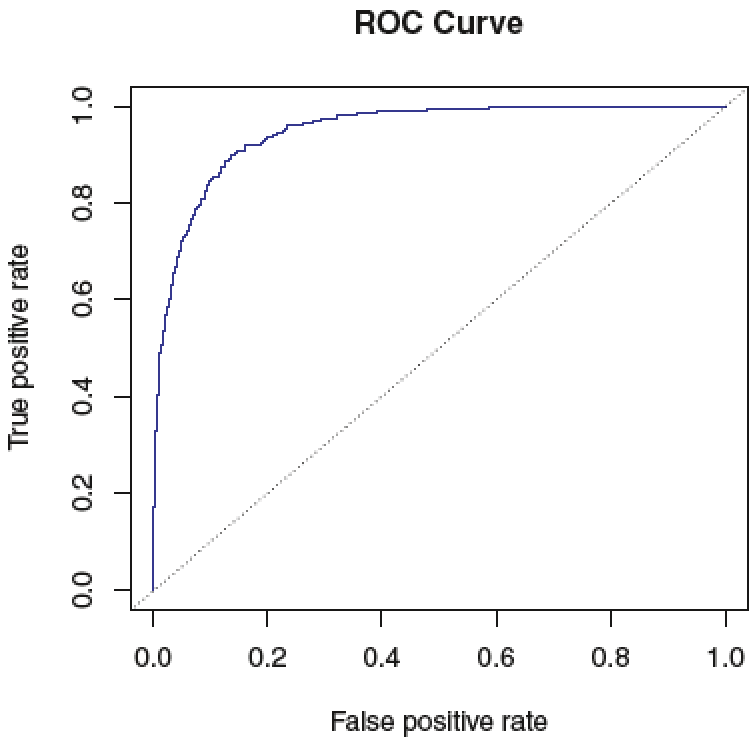


FIGURE 4.7. *For the* `Default` *data set, error rates are shown as a function of the threshold value for the posterior probability that is used to perform the assignment. The black solid line displays the overall error rate. The blue dashed line represents the fraction of defaulting customers that are incorrectly classified, and the orange dotted line indicates the fraction of errors among the non-defaulting customers.*

Krzywa ROC – receiver operating characteristic



		Predicted class		
		− or Null	+ or Non-null	Total
True class	− or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N*	P*	

Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N*	

Inne miary

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$\text{F1} = 2 \text{ x } \frac{Precision * Recall}{Precision + Recall}$$