11 Paskaita. Logistinė regresija.

11.1 Logistinės regresijos modelis

Išnagrinėkime regresiją tuo atveju, kai priklausomas kintamasis Y įgyja tik 2 reikšmes – 0 ir 1. Taip galima perkoduoti bet kokį dvireikšmį kintamąjį. Tokie uždaviniai iškyla, kai norime atsakyti į klausimus, pvz.,

- pagal paciento svorį ir kraujo parametrus norime sužinoti, kiek tikėtina, kad jis susirgs diabetu;
- pagal testų rezultatus norime sužinoti, kiek tikėtina, kad kompiuteriui prireiks remonto;
- žinodami rinkėjo amžių, pajamas ir išsilavinimą norime įvertinti tikimybė, kad jis balsuos už tam tikrą kandidatą į prezidentus.

Visiems šiems uždaviniams spręsti, kai priklausomas kintamasis yra dvireikšmis, yra naudojama logistinė regresija. Kategorinio kintamojo reikšmių prognozavimas tam tikra prasme yra klasifikavimo uždavinys.

Tarkime, kad stebimos nepriklausomų kintamųjų reikšmės $X_1 = x_{1i}, \ldots, X_k = x_{ki}$. Tuomet dauginės regresijos modelį galime užrašyti taip:

$$Y_i = a + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki} + \varepsilon_i, \tag{1}$$

čia Y_i yra atsitiktinis dydis, galintis įgyti reikšmes 0 arba 1 su tikimybėmis $P(Y_i = 1) = p_i, P(Y_i = 0) = 1 - p_i$, o ε_i yra atsitiktinė paklaida. Modelis (1) netinka prognozuoti dvireikšmiam kintamąjam Y dėl kelių priežasčių. Pirmiausia, duomenis netenkina tiesinės regresijos prielaidų. Iš tikrųjų, Y (tuo pačiu ir ε_i) gali įgyti tik 2 reikšmes, todėl netenkinama normalumo prielaida. Tiesinėje regresijoje prognozuojamas kintamojo Y vidurkis. Kadangi $EY_i = p_i$, tai prognozuotume tikimybę $P(Y_i = 1)$, tuo tarpu dešinioji regresijos lygties (1) puse gali įgyti reikšmes ir už intervalo $[0, 1]$ ribų.

Logistinės regresijos modelis yra:

$$p_i = \frac{e^{z_i}}{1 + e^{z_i}}, z_i = a + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki}, \tag{2}$$

Akivaizdu, kad $p_i \in [0, 1]$. Iš lygties (2) išreiškiame

$$\frac{p_i}{1 - p_i} = exp(a + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki}) = exp(z_i), \tag{3}$$

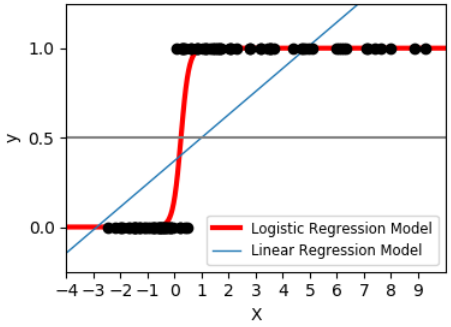
$$\ln \frac{p_i}{1 - p_i} = a + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki} = z_i. \tag{4}$$

Tikimybių santykis $\dfrac{P(Y_i = 1)}{P(Y_i = 0)} = \dfrac{p_i}{1 - p_i}$ vadinamas **galimybe** įvykti įvykiui $Y_i = 1$ (*angl.* odds). Pavyzdžiui, jei $p_i = 0,25$, tai $p_i/(1 - p_i) = 0,25/0,75 =$

1/3, t.y. galimybė Y_i įgyti reikšmę 1, o ne 0 vertinama kaip 1 ir 3 santykis. Čia mes nesakome, kad tikimybė laimėti rungtynes yra 0,7, o sakome, kad galimybė laimėti rungtynes vertinama, kaip $7:3$. Įvykio Y galimybė įvykti yra didesnė už 1 tada ir tik tada, kai

$$P(Y=1) > P(Y=0).$$

Galimybės logaritmas $z(x_i) = \ln \dfrac{p_i}{1-p_i}$ (dar vadinamas *logit* funkcija) nuo kintamųjų $x_{1i}, \ldots, x_{ki}$ reikšmių priklauso tiesiškai (žr. (4)). Taigi logistinėje regresijoje matematinis modelis sudaromas ne pačiam priklausomam kintamajam Y, o jo tikimybių santykio logaritmui (logit funkcijai). Kai $z_i \to -\infty, p_i \approx 0$. Kai $z_i \to \infty, p_i \approx 1$. Kai $z_i = 0, p_i = 0,5$. Logistinėje regresijoje nereikalaujama paklaidų normalumo, homoskedastiškumo, tačiau modelis (2) gali netikti prognozavimui dėl kintamųjų X_i multikolinearumo. Kintamieji X_i gali būti ne tik intervaliniai, bet ir kategoriniai.



1 pav.: Logistinės ir paprastosios tiesinės regresijos modeliai

Tarkime, kad stebimi duomenys $(y_i, x_{1i}, \ldots, x_{ki}), i = 1, 2, \ldots, n$. Čia y_i yra 0 arba 1, o $x_{1i}, \ldots, x_{ki}$ – intervalinių arba kategorinių kintamųjų reikšmės. Parametrų $a, b_1, \ldots, b_k$ įverčius $\hat{a}, \hat{b}_1, \ldots, \hat{b}_k$ reikia parinkti taip, kad modelis (2) būtų kuo geriau suderintas su turimais duomenimis. Šiam tikslui taikomas didžiausio tikėtinumo metodas. Kiekvienam stebėjimui pagal (2) formulę skaičiuojama tikimybė p_i. Parametrų įverčiai $\hat{a}, \hat{b}_1, \ldots, \hat{b}_k$ parenkami taip, kad tikėtinumo funkcija

$$L = \prod_{i:y_i=1} p_i \prod_{i:y_i=0} (1-p_i) \tag{5}$$

būtų maksimali. Čia

$$p_i = \frac{e^{\hat{z}_i}}{1+e^{\hat{z}_i}}, \hat{z}_i = \hat{a} + \hat{b}_1 x_{1i} + \hat{b}_2 x_{2i} + \ldots + \hat{b}_k x_{ki}. \tag{6}$$

Tuomet gauti koeficientai naudojami $\hat{z}$ taip pat ir tikimybės įverčiui $\hat{p}$ skaičiuoti esant pradinių duomenų vektoriui $\vec{x} = (x_1, \ldots, x_k)$:

$$\hat{p} = P(Y=1|\vec{x}) = \frac{e^{\hat{z}(\vec{x})}}{1+e^{\hat{z}(\vec{x})}}, \hat{z}(\vec{x}) = \hat{a} + \hat{b}_1 x_i + \hat{b}_2 x_2 + \ldots + \hat{b}_k x_k. \tag{7}$$

Žinoma, tikimybės įverčiai skaičiuojami tik tokioms x reikšmėms, kurios patenka į duomenų aibės intervalą, t.y., pvz., x_1 turi būti iš intervalo $[\min x_{1i}, \max x_{1i}]$.

Pavyzdys. Nagrinėkime 24 studentų įskaitos duomenis. Priklausomas kintamasis $Y = 1$, jei studentas gavo įskaitą, $Y = 0$, jei studentas įskaitos negavo (atitinkamas kintamasis Y pavadintas $Islaike$). Intervalinis nepriklausomas kintamasis $Prat$ reiškia, kiek praktinių užsiėmimų dirbo studentas, dvireikšmis kintamasis $Kalb$ reiškia, ar studentas iki sesijos ko nors klausė dėstytojo (1 – klausė, 0 – neklausė):

Islaike	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1
Kalb	0	0	0	0	0	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	0	0	0
Prat	19	17	13	15	19	21	17	18	23	15	13	26	30	19	22	21	24	28	30	27	21	24	20	16

Įvertinsime tikimybę, kad išlaikys įskaitą studentas, kuris pratybose dirbo 19 valandų, ir klausė dėstytojo. R pagalba įvertinsime koeficientus a, b_1, b_2:

```
> df1$Kalb <- relevel(df1$Kalb, ref = "1")
> modelis <- glm(Islaike ~ Kalb+Prat,data=df1,family="binomial")
> summary(modelis)

Call:
glm(formula = Islaike ~ Kalb + Prat, family = "binomial", data = df1)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.70639  -0.50710   0.07332   0.62691   1.83483

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.3711     5.6633  -2.184   0.0289 *
Kalb0         1.4592     1.3760   1.060   0.2889
Prat          0.5896     0.2643   2.231   0.0257 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2 pav.: Logistinės regresijos koeficientų įverčiai

Taigi galime užrašyti z formulę:

$$z = -12,371 + 1,459 \cdot Kalb + 0,590 \cdot Prat,$$

Jei $Kalb = 1$, o $Prat = 19$, tai $z = 0,289$. Pagal formulę (2) apskaičiuojame

$$P(Islaike = 1|(19,1)) = exp(0,289)/(1 + exp(0,289)) = 0,574.$$

11.2 Koeficientų interpretacija

Jei kintamasis yra intervalinis, tai atitinkamas koeficientas $\hat{b}_i$ parodo, kiek padidėja (sumažėja) $z(x)$ reikšmė (t.y. galimybės logaritmo reikšmė), jei x_i padidėja vienu vienetu, o likusieji x-ai yra fiksuoti. Pavyzdžiui, jei $Kalb = 1$, o $Prat = 20$, tai $z = 0,888, P(Islaike = 1|(20,1)) = exp(0,888)/(1 + exp(0,888)) = 0,708$. Palyginus su $Kalb = 1$, o $Prat = 19$, galimybės logaritmo reikšmė padidėjo $0,888 - 0,289 = 0,59$.

Kategorinių kintamųjų atveju skirtingoms x_i reikšmėms gaunamos skirtingos galimybių logaritmo formulės. Tuomet skaičiuojamas *galimybių santykis*

$exp(\hat{b}_i)$ (*angl.* odds ratio), kuris parodo, kaip kinta Y galimybė įgyti reikšmę 1. Studentų, kurie klausė dėstytojo, galimybė gauti įskaitą yra

$$\left(\frac{P(Y=1|Prat)}{1-P(Y=1|Prat)}\right)_{Kalb=1} = exp(-12,371+1,459\cdot 1+0,590\cdot Prat).$$

Studentų, kurie neklausė dėstytojo, galimybė gauti įskaitą yra

$$\left(\frac{P(Y=1|Prat)}{1-P(Y=1|Prat)}\right)_{Kalb=0} = exp(-12,371+1,459\cdot 0+0,590\cdot Prat).$$

Tuomet galimybių santykis yra

$$\left(\frac{P(Y=1|Prat)}{1-P(Y=1|Prat)}\right)_{Kalb=1} : \left(\frac{P(Y=1|Prat)}{1-P(Y=1|Prat)}\right)_{Kalb=0} = exp(1,459) = 4,302.$$

Nesunku įsitikinti, kad galimybių santykis didesnis už vienetą tik tada, kai atitinkamas koeficientas b teigiamas. Studentas, kuris uždavė dėstytojui klausimą, padidino galimybę gauti įskaitą $4,3$ karto. Koeficientą prie kintamojo $Prat$ galima interpretuoti taip: viena papildoma praktinių užsiėmimų valanda padidina galimybę gauti įskaitą $exp(0,59) = 1,8$ karto.

Pavyzdžiui, jei $Kalb = 1$, o $Prat = 20$, tai $P(Islaike = 1|(20,1)) = 0,708$. Tuomet galimybė gauti įskaitą yra $0,71/(1-0,71) = 2,45$, t.y. galimybė vertintina kaip $2,45 : 1$. Jei studentas dirbtų 21 praktinę valandą, jis savo galimybę padidintų $1,8$ kartų: $2,45 \cdot 1,8 = 4,4$. Iš tikrųjų, tuomet būtų $z = 1,478, P(Islaike = 1|(21,1)) = 0,814$, galimybė gauti įskaitą $0,814/(1-0,814) = 4,4$.

11.3 Klasifikavimas

Žinodami $P(Y=1|\vec{x})$ duomenis klasifikuojame pagal tokią taisyklę:

- Jei $\hat{P}(Y=1|\vec{x}) > 0,5$, tai prognozuojame, kad $Y=1$;
- Jei $\hat{P}(Y=1|\vec{x}) < 0,5$, tai prognozuojame, kad $Y=0$;
- Jei $\hat{P}(Y=1|\vec{x}) = 0,5$, tai Y reikšmė nustatoma metant monetą.

Prognozuojama labiausiai tikėtina Y reikšmė. Logistinėje regresijoje prognozuojame tikimybes ir pagal jas nustatome, kad $Y=1$ arba $Y=0$. Taip darydami mes suklystame arba nesuklystame. Vertinant logistinės regresijos modelį prasminga kalbėti ne apie liekamąsias paklaidas, bet apie teisingų prognozių procentą. R pagalba galime atspausdinti vadinamąją klasifikacinę lentelę (žr. pav. 3), kurioje pateiktos duomenų y_i reikšmės (eilutėse) ir prognozuojamos y_i reikšmės (stulpeliuose). Pagal gautus klasifikavimo rezultatus apskaičiuojami klasifikavimo jautrumas, specifiškumas ir tikslumas. Klasifikavimo *jautrumas*

nusako galimybę teisingai prognozuoti įvykį $Y_i = 1$ naudojantis sudaryta logistinės regresijos lygtimi, *specifiškumas* – įvykį $Y_i = 0$, klasifikavimo *tikslumas* nusako teisingai klasifikuotų visų įvykių nuošimtį.

Iš 11 įskaitos negavusių studentų 7 buvo klasifikuoti teisingai, tai sudaro 63,6% visų, negavusių įskaitos. Šis skaičius vadinamas modelio *specifiškumu*. Iš 13 įskaitą gavusių studentų 12 buvo klasifikuoti teisingai, tai sudaro 92,3% iš visų, gavusių įskaitą. Šis skaičius vadinamas modelio *jautrumu*. Bendras teisingai klasifikuotų studentų procentas yra 79,2. Natūralu reikalauti, kad teisingai prognozuotume ne mažiau kaip 50% visų kiekvienos kategorijos atvejų. Atkreipkime dėmesį į tai, jog logistinę regresiją galima taikyti tik jei $y_i = 0$ sudaro ne mažiau nei 20% ir ne daugiau nei 80% visų stebėjimų.

```
> predict <- predict(modelis, type = 'response')
> table(df1$Islaike, predict > 0.5)

    FALSE TRUE
  0     7    4
  1     1   12
> # Sensitivity
> 12/13
[1] 0.9230769
> # Specificity
> 7/11
[1] 0.6363636
> # Accuracy
> (7+12)/(7+4+1+12)
[1] 0.7916667
> sensitivity(Islaike, predict, threshold = 0.5)
[1] 0.9230769
> specificity(Islaike, predict, threshold = 0.5)
[1] 0.6363636
```

3 pav.: Logistinės regresijos prognozės tikslumo parametrai - jautrumas (sensitivity), specifiškumas (specificity) ir viso modelio tikslumas (accuracy).

11.4 Modelio diagnostika

Modeliui įvertinti galima taikyti χ^2 suderinamumo kriterijų, kuris tikrina, ar bent vienas iš koeficientų b_i yra reikšmingas:

$$\begin{cases} H_0 : b_1 = b_2 = \ldots = b_k = 0, \\ H_1 : \text{bent vienas } b_i \neq 0. \end{cases}$$

Tarkime, kad didžiausio tikėtinumo metodu radome parametrus $\hat{a}, \hat{b_i}, i = 1, 2, \ldots, k$, įstatėme juos į tikėtinumo funkcijos L formulę (5) ir gavome šios funkcijos maksimumą $L(\hat{a}, \hat{b})$. Tarkime dabar, kad pasirinkome logistinės regresijos modelį, kur visi $b_i = 0$, t.y. $z_i = a$. Tuomet didžiausio tikėtinumo funkcijos maksimumą pažymėkime $L(\tilde{a}, 0)$. χ^2 kriterijus remiasi tuo, kad $L(\hat{a}, \hat{b})$ mažai skiriasi nuo $L(\tilde{a}, 0)$, kai visi $b_i = 0$. Kriterijaus statistika:

$$\chi^2 = -2 \ln L(\tilde{a}, 0) + 2 \ln L(\hat{a}, \hat{b}).$$

Nulinė hipotezė atmetama (bent vienas $b_i \neq 0$), kai $\chi^2 > \chi^2_\alpha(k)$. Čia $\chi^2_\alpha(k)$ yra χ^2 skirstinio kritinė reikšmė. Jei atitinkama p-reikšmė mažesnė už pasirinktą α, nulinė hipotezė atmetama.

Jei duomenų nėra daug, vietoje χ^2 galima taikyti jo analogą – Chosmerio-Lemešou kriterijų.

Voldo kriterijus (*angl.* Wald) yra Stjudento kriterijaus tiesinėje regresijoje analogas, jis tikrina hipotezes apie atskirų koeficientų b_i reikšmingumą:

$$\begin{cases} H_0 : b_i = 0, \\ H_1 : b_i \neq 0. \end{cases}$$

Jei atitinkama p-reikšmė mažesnė už pasirinktą α, nulinė hipotezė atmetama. Modelio tinkamumui galima naudoti pseudodeterminacijos koeficientus – Makdafeno, Kokso-Snelo, Nagelkerkės. Reikėtų, kad visų šių koeficientų reikšmės būtų nemažesnės nei 0,2. Pav. 4 matome testų rezultatus. Chi kvadratu testo p-reikšmė labai maža, taigi yra bent vienas reikšmingas regresijos koeficientas. Iš antros lentelės matome, kad reikšmingas yra koeficientas prie $Prat$, koeficientas prie kintamojo $Kalb$ nėra reikšmingas. Visų pseudodeterminacijos koeficientų reikšmės nemažesnės nei 0,2.

```
> anova(modelis,
        update(modelis, ~1),  #Mūsų modelį lyginame su nuliniu modeliu
        test="Chisq")
Analysis of Deviance Table

Model 1: Islaike ~ Kalb + Prat
Model 2: Islaike ~ 1
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        21     18.601
2        23     33.104 -2  -14.503 0.000709 ***

> library(car)
> Anova(modelis, type="II", test="Wald")
Analysis of Deviance Table (Type II tests)

Response: Islaike
     Df  Chisq Pr(>Chisq)
Kalb  1 1.1246    0.28892
Prat  1 4.9780    0.02567 *


> library(rcompanion)
> nagelkerke(modelis)
$Models

Model: "glm, Islaike ~ Kalb + Prat, binomial, df1"
Null:  "glm, Islaike ~ 1, binomial, df1"

$Pseudo.R.squared.for.model.vs.null
                            Pseudo.R.squared
McFadden                           0.438109
Cox and Snell (ML)                 0.453544
Nagelkerke (Cragg and Uhler)       0.606135

$Likelihood.ratio.test
 Df.diff LogLik.diff  Chisq    p.value
      -2     -7.2516 14.503 0.00070902
```

4 pav.: Modelio adekvatumo kriterijai ir pseudodeterminacijos koeficientai.

11.5 ROC kreivės. AUC plotas.

ROC (*angl.* Receiver operating characteristic) kreivės grafiškai parodo, ar logistinės regresijos modelis yra geras, kai jo sprendimo priėmimo slenkstis kinta. ROC kreivė brėžiama atvaizduojant *true positive rate* (TPR) arba jautrumą Y ašyje ir *false positive rate* (FPR) arba 1- specifiškumą X ašyje įvairioms slenksčio reikšmėms. Pavyzdžiui, iš pav. 5 lentelės galime apskaičiuoti true positive rate.

$$TPR = \frac{TP}{\text{actual yes}} = \frac{TP}{TP + FN} = \frac{100}{105} = 0,95.$$

Dabar apskaičiuojame false positive rate.

$$FPR = \frac{FP}{\text{actual no}} = \frac{FP}{FP + TN} = \frac{10}{60} = 0,17.$$

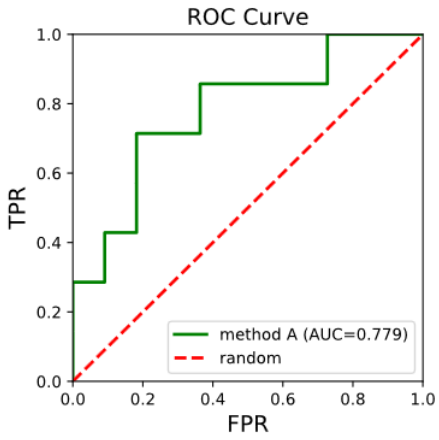
Geriausias galimas prognozavimo metodas apibūdinamas tašku, esančiu ROC erdvės viršutiniame kairiajame kampe arba koordinatėje (0,1), t.y. 100% jautrumas (nėra false negatives) ir 100% specifiškumas (nėra false positives). Atsitiktinis spėliojimas atitinka taškus išilgai įstrižainės (vadinamosios nediskriminavimo linijos) nuo apatinio kairiojo iki viršutinio dešiniojo kampo (žr. 6 pav.) Intuityvus atsitiktinio spėliojimo pavyzdys yra monetos metimas sprendimui priimti. Įstrižainė padalija ROC erdvę į dvi dalis. Taškai, esantys virš įstrižainės, yra geri klasifikavimo rezultatai (geriau nei atsitiktiniai); taškai, esantys žemiau linijos, rodo blogus rezultatus (blogiau nei atsitiktinai). Kuo aukščiau diagonalės yra kreivė, tuo geriau. Svarbus modelio "gerumo" rodiklis yra taip vadinamas plotas po kreive arba AUC (*Area under the curve*). AUC svyruoja nuo 0 iki 1, o neinformatyvaus klasifikatoriaus AUC lygus 0,5.

n=165	**Predicted: NO**	**Predicted: YES**	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

5 pav.: Klasifikacinė lentelė (*angl.* confusion matrix).

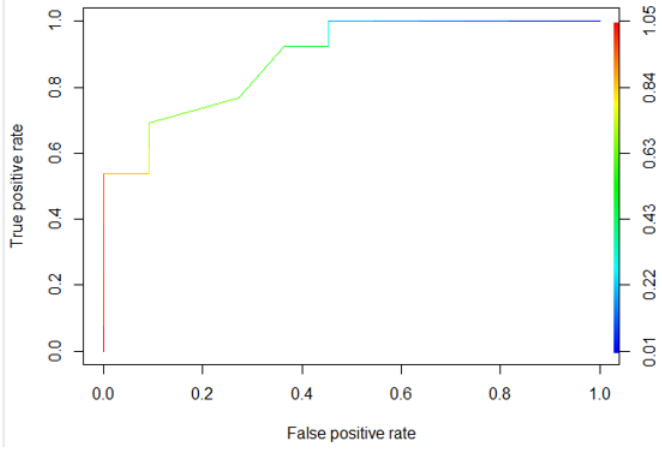
Mūsų nagrinėto pavyzdžio ROC kreivė ir AUC pavaizduoti pav. 7.

Atsakykime į klausimą: kas pasikeis, kai pakeisime klasifikavimo slenkstį? Tarkime, mes padidiname slenkstį. Tuomet kai kurie stebėjimai, kurie anksčiau buvo klasifikuojami į teigiamą klasę (arba $Y = 1$), dabar bus prognozuojami kaip neigiamos klasės nariai (arba $Y = 0$). Taigi teigiamų prognozių skaičius sumažėtų – tai reiškia, kad arba sumažės TP (true positives), arba sumažės FP (false positives) arba sumažės abu TP ir FP. Kaip tai paveiktų jautrumą (TPR)? Jautrumo skaitiklis yra TP, taigi skaitiklis gali sumažėti. Kas atsitinka su jaut-



6 pav.: ROC kreivės.

```
> predict <- predict(modelis, type = 'response')
> library(ROCR)
> ROCRpred <- prediction(predict, df1$Islaike)
> ROCRperf <- performance(ROCRpred, 'tpr','fpr')
> plot(ROCRperf, colorize = TRUE, text.adj = c(-0.2,1.7))
> auc <- performance(ROCRpred, measure = "auc")
> auc <- auc@y.values[[1]]
> auc
[1] 0.8881119
```



7 pav.: ROC kreivė).

rumo vardikliu? Vardiklis yra TP + FN, kuris iš tikrųjų yra bendras teigiamų stebėjimų skaičius, kuris nepasikeis. Taigi, didinant slenkstį, jautrumas sumažėtų arba nesikeistų. Dabar supraskime, kaip tai paveiktų FPR=1-specifiškumą. 1-specifiškumo skaitiklis yra FP, todėl skaitiklis gali sumažėti. Šios trupmenos vardiklis yra FP + TN, kuris iš tikrųjų yra neigiamų stebėjimų skaičius, kuris liktų nepakitęs. Taigi, didinant slenkstį, 1-specifiškumas sumažėtų arba nesikeistų. Atitinkamai specifiškumas padidėtų arba nepasikeistų. Taigi ROC kreivė faktiškai yra laužtė, kuri leidžia pasirinkti priimtiną mums slenkstį.