LATVIJAS UNIVERSITĀTE
FIZIKAS UN MATEMĀTIKAS FAKULTĀTE
MATEMĀTIKAS NODAĻA

NEPĀRTRAUKTO TESTU ROC LĪKNES

KURSA DARBS

Autors: **Anastasija Tetereva**

Stud. apl. at05001

Darba vadītājs: doc. Dr.math. Jānis Valeinis

RĪGA 2009

Anotācija

Šajā kursa darbā *ROC* līknes (Receiver-Operating Characteristic Curves) tiek apskatītas kā matemātisks objekts. Izmantojot definīciju un analoģijas ar citiem matemātiskiem objektiem tiek pētītas to īpašības un raksturojošie lielumi. Īpaša uzmanība ir pievērsta līkņu, un ar tiem saistīto lielumu, interpretācijai, kā arī pielietojumam praksē. Darba otrajā daļā tiek izskatīti vairāki *ROC* līkņu novērtēšanas veidi, aprakstīti daži ticamības intervālu konstruēšanas paņēmieni. Kursa darbs satur aprakstīto metožu ilustrācijas uz simulēto datu piemēriem.

Saturs

Ievads

ROC līknes (Reciever-Operating Characteristic curves) tika ieviestas 20. gadsimta vidū un lietotas radio signālu noteikšanas teorijā ar mērķi atšķirt īstos signālus, kuri tika pavadīti ar trokšņiem. Vēlāk šī pieeja tika vispārināta un to sāka pielietot klasificēšanas testiem daudzās nozarēs, piemēram, lēmumu pieņemšanas teorijā, ekonomikā, *data mining*, kreditēšanā u.t.t. Tagad ROC analīze ir guvusi ļoti plašu pielietojumu klāstu.

Šajā darbā ar jēdzienu *tests* sapratīsim klasificēšanas problēmu, kad balstoties uz testa skaitlisko rezultātu, kurš atrodas zem vai virs pieļaujamā *sliekšņa*, objekts tiek pieskaitīts pie vienas vai otras grupas. Pie noteiktas sliekšņa vērtības nav grūti aprēķināt 1. un 2. veida kļūdas, taču tāda pieeja nedod pilnu priekšstatu par testu kopumā. ROC analīze ir metode, ar kuras palīdzību tiek mērīta testa precizitāte neatkarīgi no lēmuma pieņemšanas sliekšņa, tāpēc tā ir plaši pielietota diagnostikas metožu skaitliskā apraksta izveidošanai un testu salīdzināšanai visām pieļaujamām sliekšņa vērtībām. Apskatīsim tikai tādus klasificēšanas testus, kuru vērtības pieder reālo skaitļu kopai.

ROC līknes palīdz novērtēt no testa iegūtas informācijas ticamību, ar dažādiem paņēmieniem izvēlēties testa optimālo slieksni, skaitliski salīdzināt vairākus testus un vizualizēt rezultātus, skaitliski izmērīt testa pareizību. Tās tiek plaši pielietotas arī citās statistikas nozarēs, piemēram loģistiskajā regresijā.

Darba struktūra ir sekojoša:

1. *ROC* līknes. Šajā nodaļā *ROC* līknes tiek definētas gan intuitīvi, gan kā matemātisks objekts, tiek izskatītas tās pamatīpašības.

2. *ROC* līkni raksturojošie lielumi. Šajā nodaļā tiek ieviesti un pētīti lielumi, kuri palīdz novērtēt līkni, interpretēt testa ticamību un salīdzināt līknes, kā rezultātā arī testus savā starpā.

3. Novērtēšanas metodes. Šajā nodaļā uzmanība tiek pievērsta gan parametriskām novērtēšanas metodēm, gan neparametriskām. Metodes tiek salīdzinātas uz simulēto datu piemēriem.

4. Ticamības intervāli. Šajā nodaļā ir aprakstīti daži ticamības intervālu konstruēšanas paņēmieni, kuri tiek ilustrēti ar piemēriem.

5. Pielikumos ir kursa darba ietvaros uzrakstīto programmu kods.

1. ROC līknes

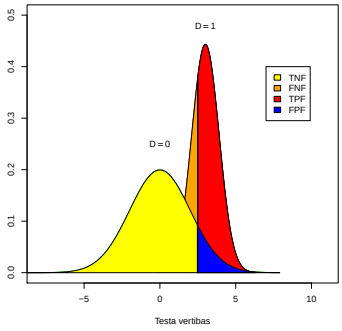
Lai labāk saprastu ROC līknes jēdzienu, nodaļas sākumā ir izskatīti piemēri no medicīnas, kreditēšanas un signālu noteikšanas teorijas. Tālāk tiek dota līkņu intuitīvā un matemātiskā definīcija. Balstoties uz matemātisko definīciju ir formulētas ROC līkņu īpašības.

Sāksim ar piemēru no medicīnas jomas, kurš bija publicēts žurnālā [1].
Intensīvās terapijas nodaļā atrodas grupa ar pacientiem, kuriem, iespējams, ir sepse (asins saindēšanās). Pieņemsim, ka eksistē tests, ar kura palīdzību tiek noteikts, vai pacients ir slims. Testa rezultāta vērtību apzīmēsim ar Y. Uzskatīsim pacientu par slimu, ja testa vērtība ir lielāka par slieksni c, par veselu, ja testa vērtība ir mazāka par c. Pieņemsim, ka eksistē arī 'zelta standarts', kurš ļauj noteikt, vai pacientam tiešām ir sepse. Salīdzinot testa un 'zelta standarta' rezultātus, iegūstam sekojošu tabulu: TPF (true positive fraction - patiesi pozitīva daļa) ir tāda slimo pacientu daļa, kuriem testa rezultāts ir pozitīvs (pacients ir slims). FPR ir pacientu daļa, kuriem testa rezultāts bija pozitīvs neskatoties uz to, ka pacients ir vesels. FNR un TNF attiecīgi ir slimi un veseli pacienti ar negatīvu testa rezultātu. Acīmredzami, ka pie fiksēta c tests ar lielākām TPF un TNF vērtībām ir uzskatāms par labāku nekā tests ar mazākām TPF un TNF. Tas nozīmē, ka var labāk atšķirt slimus pacientus no veseliem. Ilustrēsim piemēru ar attēlu 1.1. un tabulu 1.2..

ROC analīzē bieži tiek lietoti termini *sensitivity* un *specificity*. *sensitivity* (jūtīgums) rāda, cik labi ar testa palīdzību var tikt noteikts slims pacients. Par *specificity* (specifiku) tiek saukta testa spēja neuzskatīt veselus pacientus par slimiem.

1.1. tabula Piemēra 1.1.1. ilustrācija

	Testa rezultāts ir pozitīvs	Testa rezultāts ir negatīvs
Pacients ir slims	TPF	FNF
Pacients ir vesels	FPF	TNF



1.1. att. Testa vērtības sadalījuma blīvuma funkcijas slimiem un veseliem pacientiem.

Grāmatā [2] tiek apskatīts interesants *ROC* līkņu pielietošanas aspekts kreditēšanas jomā.

Pirms banka dod kredītu potenciālam klientam, tā ar kāda testa palīdzību (praksē visbiežāk pielieto 'Altmana Z' punktus un logit modeļus) novērtē klienta turpmāko maksātspēju. Klientiem, kuriem punktu skaits ir lielāks vai vienāds par noteikto slieksni, piešķir kredītu. Tādā veidā visus klientus sadala potenciāli maksātspējīgos un maksātnespējīgos. Pieņemsim, ka tie klienti, kuriem banka atteica, paņēma kredītu citā organizācijā. Pēc 'pēdējā' līguma beigu termiņa paliek skaidrs, kuri no klientiem realitātē varēja nokārtot savas kredītsaistības. Tādā veidā visi klienti sadalās uz TP (banka deva kredītu, klients to atmaksāja), TN (banka nedeva kredītu, klients nevarēja atmaksāt kredītu citai organizācijai), FP (banka deva kredītu, klients to neatmaksāja), FN (banka nedeva kredītu, klients atmaksāja kredītu citai organizācijai). Šajā gadījumā ar ROC līknes palīdzību var ne tikai noteikt optimālo slieksni, lai noteiktu klienta kredītspēju, bet arī saprast, kurš no vairākiem testiem dod ticamākus rezultātus.

Piemērs, kurš apraksta *ROC* līkņu pirmssākumus ir atrodams [3].

Otrā pasaules kara laikā bija ļoti svarīgi ar radaru palīdzību noteikt momentu, kad kādam objektam tuvojas ienaidnieka lidmašīna. Tātad, radaram bija jāatšķir pretinieku lidmašīna no putnu čivināšanas un pārējiem trokšņiem. Radara jūtīgumu pret ārējiem trokšņiem var regulēt. Jo labāk radars uztver trokšņus, jo lielāka varbūtība, ka viltus trauksmju līmenis būs liels, kas radīs zaudējumus. Ja radara jūtīgums pret trokšņiem būs mazs, tad objektam draud briesmas un ienaidnieks uzbruks nepamanīts. Ir jāatbild uz jautājumu - kurš radars precīzāk atšķir ārējos trokšņus no īstajām trauksmēm.

Visu augšminēto piemēru problemātika ir ļoti līdzīga - novērtēt, cik labi ar konkrētā

1.2. tabula: Statistisko hipotēžu pārbaudes un diagnostisko testu terminoloģijas salīdzinājums.

	Statistiskās hipotēzes pārbaude	Diagnostikas tests
Mērķis	Pārbaudīt H_0 pret H_1	Pārbaudīt $D = 0$ pret $D = 1$
1. veida kļūda	$\alpha = P[\text{noraidīt } H_0 \mid H_0]$	$FPF = P[\text{ klasificēt } D = 1 \mid D = 0]$
2. veida kļūda	$1 - \beta = P[\text{noraidīt } H_0 \mid H_1]$	$TPR = P[\text{ klasificēt } D = 1 \mid D = 1]$
ticamības attiecība	$LR(W) = P[W \mid H_1]/P[W \mid H_0]$	$LR(Y) = P[Y \mid D = 1]/P[Y \mid D = 0]$

testa palīdzību var atšķirt objektus, kuriem piemīt kāda īpašība, no objektiem, kuriem tādas īpašības nav. Apzīmējot ar D īpašību raksturojošo funkciju, pieņemsim, ka D=1, ja īpašība piemīt objektam, un D=0, prētējā gadījumā.

Nosakot testa vērtības Y populācijām ar D=1 un D=0, iegūsim divas sadalījuma funkcijas. Mainot lēmuma pieņemšanas sliekšņa (c) vērtības no $-\infty$ līdz $+\infty$) iegūsim $TPR(c)$ un $FNR(c)$ katram c. Attēlojot $FPR(c)$ uz x ass un $TPR(c)$ uz y ass, iegūsim ROC līkni. Lai labāk izprastu augstāk minētos jēdzienus, salīdzināsim statistisko hipotēžu pārbaudes terminoloģiju ar diagnostikas testu terminoloģiju 1.2.. Mēģināsim formalizēt augšminētos jēdzienus un definīcijas.

Definīcija 1. [3]

$$ROC(\cdot) = \{(FPF(c), TPF(c)), c \in (-\infty, +\infty)\}, \tag{1.0.1}$$

kur Y ir iespējamās testa rezultāta vērtības, c ir lēmuma pieņemšanas slieksnis, $TPF(c) = P[Y > c \mid D = 1]$, $FPF(c) = P[Y > c \mid D = 0]$ un D ir klasificēšanas pazīmes indikatorfunkcija.

No ROC līknes definīcijas 1.0.1 seko, ka līkne vienmēr atradīsies apgabala $[0,1] \times [0,1]$ iekšpusē, jo gan $TPR(c)$, gan $FPR(c)$ ir varbūtības, tāpēc pēc definīcijas var pieņemt vērtības no 0 līdz 1. Turklāt, $\lim_{c\to+\infty} TPF(c) = 0$, $\lim_{c\to+\infty} FPF(c) = 1$, $\lim_{c\to-\infty} TPF(c) = 1$, $\lim_{c\to-\infty} FPF(c) = 0$. ROC līkne var tikt pierakstīta vēl vienā veidā.

$$ROC(\cdot) = \{(t, ROC(t)), t \in (0,1)\}$$

Lai precizētu ROC līknes definīciju statistikas valodā, izskatīsim *ROC* līknes matemātiskās īpašības.

Apgalvojums 1. [4] *ROC līkne ir invarianta attiecībā pret stingri augošām Y transformācijām.*

Pierādījums. Pieņemsim, ka $(FTP(c), TPF(c))$ ir *ROC* līknes punkts, kurš atbilst testa vērtībai Y. Lai h ir monotoni augoša transformācija un $W = h(Y)$, $d = h(c)$. Tad $P[W \geq d|D = 0] = P[Y \geq c|D = 0]$. Tātad *ROC* līkne priekš Y sakritīs ar ROC līkni priekš W. Līdzīgi var pierādīt, ka katrs no W atkarīgās *ROC* līknes punkts ir arī no Y atkarīgās ROC līknes punkts. □

Apgalvojums 2. [4] *Ja $F_{D=1}(y)$ un $F_{D=0}(y)$ ir attiecīgi objektu ar pazīmi raksturojošas funkcijas vērtībām $D = 1$ un $D = 0$ sadalījuma funkcijas, $F_{D=1}(y) = P[Y < y|D = 1]$ un $F_{D=0}(y) = P[Y \leq y|D = 0]$, tad ROC līkne ir izskatā*

$$ROC(t) = 1 - F_{D=1}(F^{-1}_{D=0}(1 - t)), t \in (-\infty, +\infty). \tag{1.0.2}$$

Pierādījums. Šis apgalvojums seko no *ROC* līknes un sadalījuma funkcijas definīcijas 1.0.1. Pieņemsim, ka $c = (F^{-1}_{D=0}(1 - t))$, t.i. c ir slieksnis, kurš apmierina vienādojumu $FPF(c) = t$.

$$P[Y > c|D = 0] = t] \Rightarrow P[Y > c|D = 0] = 1 - P[Y \leq c|D = 0] = t] = t$$

Tātad,

$$1 - F_{D=0}(c) = t \Rightarrow c = F^{-1}_{D=0}(1 - t)$$

Atbilstošu *ROC* līkni meklēsim pēc definīcijas.

$$ROC(t) = P[Y > c|D = 1] = P[Y > F^{-1}_{D=0}(1 - t)|D = 1] = 1 - F_{D=1}({F_{D=0}}^{-1}(1 - t))$$

□

Sekas 1. [4] [5] *ROC līkne var būt pierakstīta izskatā*

$$ROC(t) = S_{D=1}(S^{-1}_{D=0}(t)),$$

kur $S(y) = P[Y > y]$.

Apgalvojums 3.[4] *Ja $S(y) = P[Y > y]$ ir izdzīvošanas funkcija, tad ROC līknes atvasinājums punktā t ir izskatā*

$$\frac{\partial ROC(t)}{\partial t} = \frac{f_{D=1}(S^{-1}_{D=0}(t))}{f_{D=0}(S^{-1}_{D=0}(t))}. \tag{1.0.3}$$

Pierādījums. Mums ir

$$\frac{\partial S_{D=1}(S^{-1}_{D=0}(t))}{\partial t} = \frac{\partial S_{D=1}(S^{-1}_{D=0}(t))}{\partial S^{-1}_{D=0}(t)} \cdot \frac{\partial S^{-1}_{D=0}(t)}{\partial t} = -f_{d=1}(S^{-1}_{D=0}(t)) \cdot \frac{\partial S^{-1}_{D=0}(t)}{\partial t}.$$

Rezultāts ir spēkā, jo

$$\frac{\partial S^{-1}_{D=0}(t)}{\partial t} = \frac{1}{\frac{\partial S_{D=0}(w)}{\partial w}} = \frac{1}{-f_{D=0}(w)},$$

kad $w = S^{-1}_{D=0}(t)$. □

Pievērsīsim uzmanību faktam, ka ROC līknes atvasinājums [**?**] var būt interpretēts kā ticamības attiecība $LR(c) = \frac{P[Y=c|D=1]}{P[Y=c|D=0]}$punktam $(t, ROC(t))$, kur c apmierina vienādojumu $c = S^{-1}_{D=0}(t)$. Daudzos gadījumos tiek sagaidīts, ka palielinoties c pieaugs arī ticamības attiecības $LR(c) = \frac{f_{D=1}(c)}{f_{D=0}(c)}$ vērtība. Dotajai funkcijai ir ļoti svarīga loma medicīnas testu pētīšanā. Mēs pieminēsim tikai svarīgāko signālu noteikšanas teorijas rezultātu.[4]

Apgalvojums 4.[4] *Uz testa vērtībām Y balstīts optimālais priekšmetu klasificēšanas kritērijs ir*

$$LR(Y) > c,$$

t.i. TPF, izpildoties dotajai nevienādībai, sasniedz maksimālo vērtību starp visām iespējamām. Izsakot t, iegūstam $t = P[LR(Y) > c|D = 0]$.

Pierādījums. Pierādījums seko no Neimaņa-Pīrsona lemmas □

Turklāt,

(i) ja testa vērtības ir tādas, ka $LR(\cdot)$ ir monotoni augoša, tad lēmuma pieņemšanas kritērijs, kurš ir balstīts uz lielākajām Y vērtībām, būs optimāls, jo tas ir līdzvērtīgs kritērijam $LR(Y) > c$;

(ii) $ROC_{W=LR(Y)}(t)$ līknei priekš $W = LR(Y)$ ir spēkā

$$ROC_{W=LR(Y)}(t) > ROC(t), \forall t \in (0, 1),$$

tāpēc tā ir optimāla ROC līkne;

(iii) optimāla ROC līkne ir ieliekta;

Pierādījums. Ja mēs definējam funkciju $L = LR(Y)$, tās atvasinājums pie sliekšņa vērtības x būs vienāds ar $\frac{P[L=x|D=1]}{P[L=x|D=0]} = x$, kas pēc definīcijas ir monotoni augošs pēc x. Tāpēc no $t = P[L > x|D = 0]$ atkarīgs ROC līknes atvasinājums ir monotoni dilstošs. □

No pēdējās īpašības var saprast, ka, ja $LR(Y)$ būtu zināma, tad Y bez grūtībām varētu būt transformēts uz tā optimālajām vērtībām. Taču praksē $LR(Y)$ nav zināma. Tā var būt novērtēta no datiem, lai iegūtu optimālo transformāciju. Gadījumos, kad Y ir viendimensionāls lielums, transformāciju pielietošanai nav lielas nozīmes.

Apgalvojums 5.[4][3] *Lai testa izmaksu funkcija ir dota veidā*

$$Cost(T) = C + C^{+}_{D=1} ROC(t)\rho + C^{-}_{D=1}(1 - ROC(t))\rho + C^{+}_{D=0} t(1-\rho), \tag{1.0.4}$$

kur, runājot pieņemtajā ROC analīzē medicīnas terminos

(i) C *ir testa (diagnostikas) izmaksas;*

(ii) $C^{+}_{D=1}$ *un* $C^{-}_{D=1}$ *ir attiecīgi slimo pacientu ārstēšanas un smagākas saslimšanas izmaksas, ja diagnostikas rezultāti būs pozitīvi vai negatīvi. Parasti praksē* $C^{-}_{D=1} >> C^{+}_{D=1}$*;*

(iii) $C^{+}_{D=0}$ *ir nevajadzīgas ārstēšanās izmaksas un morālu zaudējumu kompensācija, kas var rasties pacientam, neprecīzas diagnostikas rezultātā;*

(iv) $\rho = P[D = 1]$.

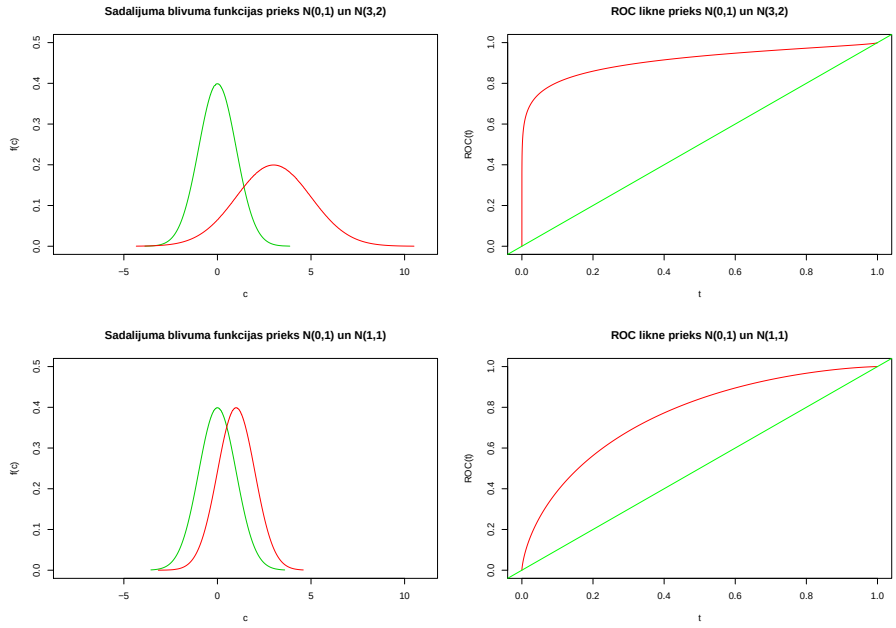
Tad slieksnim, kurš minimizē izmaksas, ir jāapmierina vienādība

$$\frac{\partial ROC(t)}{\partial t} = \frac{1-\rho}{\rho} \cdot \frac{C^{+}_{D=0}}{C^{-}_{D=1} - C^{+}_{D=1}}. \tag{1.0.5}$$

Lai saprastu, vai testa veikšanai ir ekonomiska jēga, praksē bieži vien salīdzina testēšanas izmaksas ar izmaksām, kuras var rasties, ja testu neveic, t.i.

$$Cost(\mathrm{NT}) = \rho C^{-}_{D=1} \tag{1.0.6}$$

2. ROC līkni raksturojošie lielumi



2.1. att. Divu ROC līkņu salīdzinājums

Šajā nodaļā tiks apskatīti *ROC* līkni raksturojošie lielumi un paskaidrots, kā ar to palīdzību var salīdzināt vairākas līknes.

Lai saprastu, kā ar *ROC* līkņu palīdzību var izvēlēties starp vairākiem testiem labāko, apskatīsim tās raksturojošos lielumus. Testa kvalitāte ir atkarīga no tā, cik tuvu *ROC* līkne atrodas punktam (0,1).

2.1. Laukums zem *ROC* līknes (*AUC*)

Aplūkosim piemēru, kurš ir atspoguļots attēlā 2.1.. Ir acīmredzams, ka pirmais tests, $\mathbb{N}(0,1)$ pret $\mathbb{N}(3,2)$, ir labāks par otro, $exp(5)$ pret $exp(40)$, jo sadalījuma blīvuma funkcijas atrodas tālāk viena no otras un mēs varam vieglāk atšķirt atbilstošās sadalījuma blīvuma funkcijas. Attiecīgā *ROC* līkne atrodas ļoti tuvu punktam (0,1) un tālu no 1

kvadranta bisektrises. Otrajā gadījumā atšķirt sadalījuma blīvuma funkcijas ir grūti, jo tās gandrīz sakrīt. Otrā ROC līkne atrodas diezgan tuvu diagonālei. Acīmredzami, ka vienādu salījuma blīvuma funkciju gadījumā ROC līkne sakritīs ar diagonāli un testam nebūs nekādas jēgas. Var pamanīt, ka 'labākai' līknei TPF pret FPF attiecība ir lielāka, tas nozīmē, ka vairāk ojektu, kuri ir klasificēti kā pozitīvi, tiešām ir pozitīvi.

Visvairāk izplatīts ROC līkņu raksturojošs lielums ir laukums zem ROC līknes, kuru turpmāk apzīmēsim ar AUC (Area under curve), t.i.

$$AUC = \int_0^1 ROC(t)\mathrm{dt} \tag{2.1.1}$$

Perfektam testam AUC=1, bet testam, kurš nevar dot nekādu informaciju, $ROC(t) = t$ un AUC=0.5. Skaidrs, ka tests A ir labāks par testu B, ja $ROC_A(t) \geq ROC_B(t)$, $\forall t \in (0,1)$, no tā seko $AUC_A \geq AUC_B$. Apgrieztais apgalvojums vispārīgajā gadījumā nav spēkā.

Apgalvojums 1.[6]

$$AUC = P[Y_{D=1} > Y_{D=0}].$$

Pierādījums.

$$AUC = \int_0^1 ROC(t)\mathrm{dt} = \int_0^1 S_{D=1}(S_{D=0}^{-1}(t))\mathrm{dP}[\mathrm{Y_{D=1}} > \mathrm{Y_{D=0}}] =$$

$$\int_{-\infty}^{+\infty} S_{D=1}(y)\mathrm{d}S_{D=0}^{-1}(y) = \int_{-\infty}^{+\infty} P[Y_{D=1} > y] f_{D=0}(y)\mathrm{dy}$$

Ņemot vērā, ka pieņēmums par $Y_{D=1}$ un $Y_{D=0}$ ir spēkā, iegūstam

$$AUC = \int_{-\infty}^{+\infty} P[Y_{D=1} > y | Y_{D=0} = y]\mathrm{dy} = P[Y_{D=1} > Y_{D=0}].$$

□

Lai formulētu nākamo apgalvojumu, definēsim empīrisko ROC līkni.

Apgalvojums 2. [6] *Ja*

$$\hat{F}_{D=0}(p) = \frac{1}{n_0} \sum_{i=1}^{n_0} I(y_{D=0} \leq p),$$

$$\hat{F}_{D=1}(p) = \frac{1}{n_1} \sum_{i=1}^{n_1} I(y_{D=1} \leq p),$$

tad empīriskā ROC līkne ir izskatā

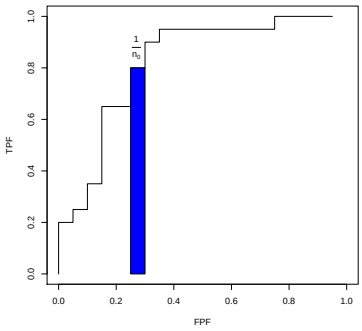
$$\hat{ROC} = 1 - \hat{F}_{D=1}(\hat{F}^{-1}_{D=0}(1-t)), \forall 0 \leq t \leq 1. \tag{2.1.2}$$

Apgalvojums 3. [6] *Laukums zem empīriskās ROC līknes ir vienāds ar ρ statistiku, kura ir vienāda ar Mann-Whitney normētu U-statistiku, t.i. $\frac{U}{n_0 n_1}$.*

$$\hat{AUC} = 1 - \sum_{j=1}^{n_0}\sum_{i=1}^{n_1} I[y_{i,D=1} > y_{j,D=0}] + \frac{1}{2} I[y_{i,D=1} = y_{j,D=0}] \tag{2.1.3}$$

Mann-Whitney U-statistikas definīcija var tikt atrasta [5].

Pierādījums. Pierādījums kļūst acīmredzams no 2.2. attēla. Sākumā vienkāršības dēļ pieņemsim, ka $y_{d=1}$ un $y_{d=0}$ visi ir dažādi, tad katrs līknes horizontāls solis, kurš atbilst testa vērtībai $y_{d=0}$ pievieno laukumam zem ROC līknes taisnstūri ar laukumu $\frac{1}{n_{D=0}} \times \hat{TPF}(y_{j,D=0}) = \frac{1}{n_{D=0}} \times \frac{\sum_i [y_{i,D=1} > y_{j,D=0}]}{n_{D=1}}$. Rezultāts ir spēkā, jo empīriskā ROC līkne sastāv no tādiem četrstūriem. Kad $y_{d=1}$ un $y_{d=0}$ vērtības atkārtojas, laukumam pieliekam klāt vēl taisnstūra trīsstūri ar katetēm $n_{D=0}$ un $n_{D=1}$, kura laukums ir $\frac{1}{2}(1/n_{D=0} \times 1/n_{D=1} y_{i,D=1} = y_{j,D=0}$. □



2.2. att. Empīriskā ROC līkne un tās laukums

2.2. *ROC* līkne pie fiksēta *FPR* un *ROC* līknes parciāls laukums

Dažreiz, piemēram, kad testa vērtības nevar pieņemt vērtības no -∞ līdz +∞, vai tādas vērtības ir sastopamas ļoti reti, vai arī pie noteiktām testa vērtībām var panākt bezkļūdainu objektu klasificēšanu, *ROC* līkne visā garumā nesastāda īpašu interesi. Šādos

gadījumos ir pieņemts apskatīt ROC līkni pie fiksēta FPR, t.i. $ROC(t_0)$. Testu A un B salīdzināšana notiek, fiksējot t_0 un apskatot $ROC(t_0)_A$ un $ROC(t_0)_B$. $ROC(t_0)$ vērtībai ir acīmredzama interpretācija - tā ir vienāda ar objektu ar $D = 1$ proporciju, ar testa rezultātu $(1 - t_0)$ kvantīli, priekš objektiem ar $D = 0$. Pie maziem t, t.i. 0.05 vai 0.01 šī kvantīle tiek interpretēta kā pieļaujams testa rezultāta augšējais slieksnis. Tāpēc $ROC(t_0)$ ir objektu, ar īpašību $D = 1$ daļa , kuru testa vērtība pārsniedz pieļaujamo. Šis raksturlielums ir nepilnīgs un nedod informāciju par visu līkni. Kompromisa variants starp AUC un $ROC(t_0)$ ir ROC līknes parciāls laukums, kas ir vienāds ar

Definīcija 1. [6][4]

$$pAUC(t_0) = \int_0^{t_0} ROC(t)\mathrm{dt}.$$

Šis raksturlielums var pieņemt vērtības no ${t_0}^2/2$ līdz t_0. ${t_0}^2/2$ - priekš neinformatīva testa, t_0 - priekš ideāla testa. Dažreiz apskata normētu ROC līknes parciālo laukumu, kurš ir interpretējams kā

$$pAUC(t_0)/t_0 = P[Y_{D=1} > Y_{D=0} | Y_{D=0} > S^{-1}_{D=0}(t_0)]. \tag{2.2.1}$$

Nepieciešamības gadījumā var apskatīt arī

$$pAUC(t_0, t_1) = \int_{t_0}^{t_1} ROC(t)\mathrm{dt}. \tag{2.2.2}$$

2.3. ROC līknes maksimālais attālums līdz $ROC(t)$=t

ROC līknes maksimālais attālums līdz taisnei $ROC(t) = t$ parāda, cik tālu līkne atrodas no pilnīgi neinformatīvas līknes un var pieņemt vērtības no 0 līdz 1. Interesanti ir tas, ka šis raksturlielums atrodas ciešā saistībā ar Kolmogorova-Smirnova statistiku, kura mēra attālumu starp diviem sadalījumiem. Ir spēkā

Apgalvojums 4. [6]

$$KS = \max_t |ROC(t) - t| = \max_t |ROC(t) - t| = \max_t |S_{D=1}(S^{-1}_{D=0}(t)) - t| =$$

$$\sup_{c \in (-\infty, +\infty)} |S_{D=1}(c) - S_{D=0}(c)|. \tag{2.3.1}$$

Tas nozīmē, ka šī statistika raksturo attālumu starp $Y_{D=0}$ un $Y_{D=1}$

2.4. Simetrijas punkts

Simetrijas punktā Sym ir spēkā apgalvojums

$$TPF = 1 - FPF,$$

citiem vārdiem

$$ROC(Sym) = 1 - Sym.$$

Runājot statistikas valodā, šis punkts izvēlas 1. veida un 2. veida kļūdu sabalansēto vērtību.

Apkoposim šīs nodaļas rezultātus tabulā 2.1..

2.1. tabula ROC līkni raksturojošie lielumi

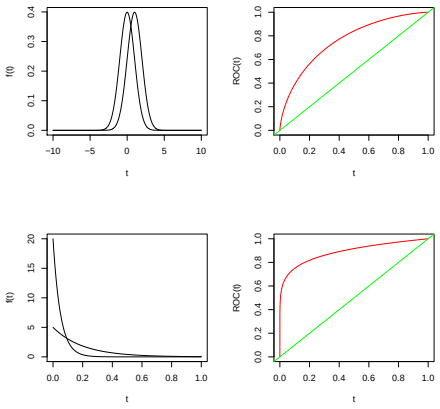
Apzīmējums	Definīcija	Interpretācija
AUC	$\int_0^1 ROC(t)\mathrm{dt}$	$P[Y_{D=1} > Y_{D=0}]$
$ROC(t_0)$	$ROC(t_0)$	$P[Y_{D=0} > q]$
$pAUC(t_0)$	$\int_0^{t_0} ROC(t)\mathrm{dt}$	$t_0 P[Y_{D=1} > Y_{D=0} \vert Y_{D=0} > q]$, $q = 1 - t_0$ $Y_{D=0}$ kvantīle
Sym	$ROC(Sym) = 1 - Sym$	Sensitivity=Specificity
KS	$\max_t \vert S_{D=1}(S_{D=0}^{-1}(t)) - t\vert$	$\sup_{c\in(-\infty,+\infty)} \vert S_{D=1}(c) - S_{D=0}(c)\vert$

3. Novērtēšanas metodes

Praksē bieži vien sanāk sastapties ar problēmu, ka ROC līkne nav zināma un, lai varētu salīdzināt vairākus testus un izmantot ROC līknes raksturlielumus statistisko lēmumu pieņemšanai, tā ir jānovērtē no datiem. Pēdējos gados ir izstrādāta vesela virkne gan parametrisko, gan neparametrisko metožu ROC līknes novērtēšanai. Apskatīsim tikai dažas no tām. Lai aptuveni saprastu, kā strādā viena vai otra metode, salīdzināsim novērtēto ROC līkni ar īsto $ROC(t) = 1 - F_{D=1}(F^{-1}_{D=0}(1-t))$. Tādam nolūkam ģenerēsim datus, $f_{D=0}$ un $f_{D=1}$ vietā ņemot attiecīgi

(i) $N(0,1)$ un $N(1,1)$,

(ii) $exp(40)$ un $exp(5)$.

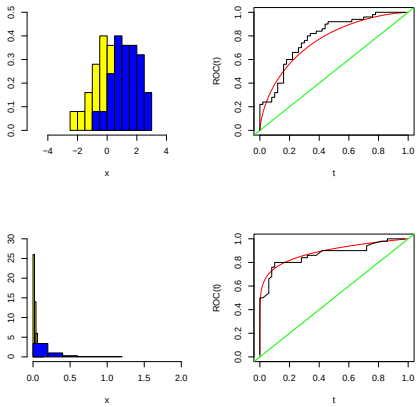


3.1. att. $N(0,1)$, N, $exp(40)$, $exp(5)$ un atbilstošās ROC līknes

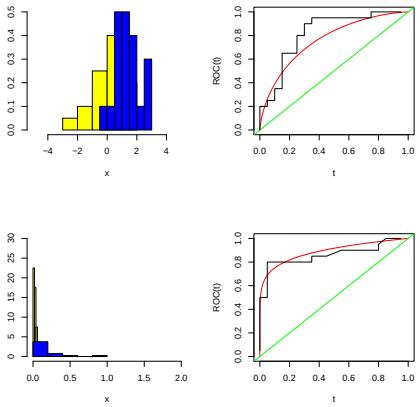
Uzzīmēsim dotās sadalījuma blīvuma funkcijas un tām atbilstošās ROC līknes 3.1., kā arī aprēķināsim galveno ROC līkni raksturojošo lielumu - $A\hat{U}C$

3.1. Empīriskā ROC līkne

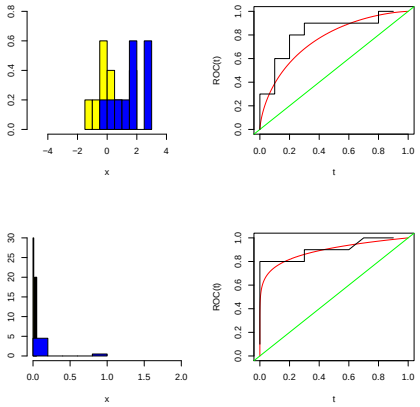
Atcerēsimies empīriskās ROC līknes definīciju 2.1.2. Jāatzīmē, ka ņemt $t \in (0,1)$ praksē nav jēgas, tas nedod rezultāta uzlabojumu, jo pakāpienveida funkcijas lēcieni notiek tikai punktos, kur ir dati, tāpēc vektora t vietā ņem $sort(y_{D=0}, y_{D=1})$. Pieņemsim, ka izlašu apjomi sakrīt un ir vienādi ar n. Apskatīsim, kāda atkarība pastāv starp ROC līknes novērtēšanas, ar $R\hat{O}C$ precizitāti, no izlašu apjomiem.



3.2. att. Izlases $N(0,1)$, $N(1,1)$, $exp(40)$, $exp(5)$ un atbilstošās $R\hat{O}C$ līknes, $n = 50$



3.3. att. Izlases $N(0,1)$, $N(1,1)$, $exp(40)$, $exp(5)$ un atbilstošās $R\hat{O}C$ līknes, $n = 20$



3.4. att. Izlases $N(0,1)$, $N)$, $exp(40)$, $exp(5)$ un atbilstošās $R\hat{O}C$ līknes, $n = 10$

Izmantojot kursa darba ietvaros uzrakstīto **R** kodu, aprēķināsim laukumus un tos salīdzināsim 3.1.. Atcerēsimies, ka laukumu aprēķināšana var tikt aizvietota ar normētu Mann-Whitney U-statistiku. Laukumu novērtējumi mainās diezgan būtiski atkarībā no n. Redzam, ka, samazinoties izlases apjomam, $A\hat{U}C$ aug un tiek pārvērtēts, kas varētu novest pie neprecīziem secinājumiem par testa kvalitāti. Var iedomāties, kas būs gadījumā, kad izlašu apjoms būs mazāks par 10. Taču praksē, īpaši medicīnā, tādi gadījumi ir sastopami.

3.1. tabula AUC un $A\hat{U}C$ dažādiem izlases apjomiem

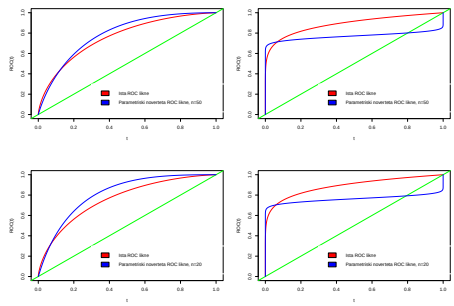
	AUC	$A\hat{U}C, n = 50$	$A\hat{U}C, n = 20$	$A\hat{U}C, n = 10$
$N(0,1)$, $N(1,1)$	0.7603	0.7896	0.820	0.820
$exp(40)$, $exp(5)$	0.889	0.868	0.860	0.910

3.2. tabula Parametrisko ROC līkņu parametru novērtējumu salīdzinājums

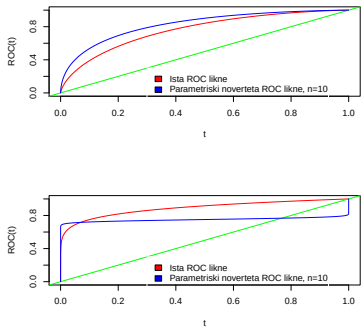
	Statistika	$N(0,1)$	$N(1,1)$	$exp(40)$	$exp(5)$
Īstās vērtības	μ	0	1	-	-
	$\sqrt{\sigma^2}$	1	1	-	-
	AUC	0.7603		0.8890	
n=50	μ	0.6914	1.2348	0.0279	0.1923
	$\sqrt{\sigma^2}$	1.1297	0.9182	0.0267	0.2206
	AUC	0.7884		0.8680	
n=20	μ	0.1955	1.3763	0.0292	0.1759
	$\sqrt{\sigma^2}$	1.0586	0.8013	0.0240	0.2044
	AUC	0.8131		0.7609	
n=10	μ	0.2112	0.5665	0.0224	0.1871
	$\sqrt{\sigma^2}$	0.9850	1.0474	0.0172	0.2431
	AUC	0.8271		0.7504	

3.2. Parametriskā ROC līkne

Parametriskās metodes būtība ir sekojoša - mēs pieņemam, ka dati ir sadalīti jau pēc iepriekš zināma sadalījuma, novērtējam sadalījuma funkcijas parametrus, piemēram, izmantojot maksimālās ticamības novērtējumus un būvējam $ROC(t)$ kā nepārtraukto funkciju no t.[3] Ilustrēsim šo metodi ar piemēru. Izmantosim datus, ģenerētus no jau iepriekš minētajiem sadalījumiem, t.i. $N(0,1)$, $N(1,1)$ un $exp(40)$, $exp(5)$. Būvējot parametrisko novērtējumu, pieņemsim, ka dati ir neatkarīgi un sadalīti normāli. Izmantosim maksimālās ticamības paramertu novērtējumus, t.i., ja mums ir dati $x_1, x_2, ..., x_n$, tad $\hat{\overline{x}} = \frac{\sum_{i=1}^{n} x_i}{n}$ un $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_i - \hat{\overline{x}})^2}{n}$. Parametru novērtēšanas rezultātus apkoposim tabulā 3.2., ROC līknes novērtēšanu ar parametrisko metodi ilustrēsim ar zīmējumiem 3.5. ,3.6..



3.5. att.: ROC līknes no $N(0,1)$, $N(1,1)$ un $exp(40)$, $exp(5)$ un atbilstošās parametriski novērtētās $R\hat{O}C$ līknes



3.6. att.: ROC līknes no $N(0,1)$, $N(1,1)$ un $exp(40)$, $exp(5)$ un atbilstošās parametriski novērtētās $R\hat{O}C$ līknes

Var redzēt, ka eksponenciālā sadalījuma metode strādā ne pārāk labi. Šādos gadījumos var mēģināt transformēt datus, lai tie atgādinātu datus no normālā sadalījuma. Parasti izmanto Box-Cox transformāciju [7], kura ir izskatā

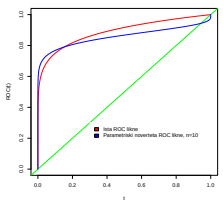
$$x(\lambda) = \begin{cases} \frac{x^{\lambda}-1}{\lambda} & ,\lambda \neq 0 \\ \ln x & ,\lambda = 0 \end{cases} \tag{3.2.1}$$

λ izvēlas tādu, lai maksimizētu maksimālās ticamības funkcijas logaritmu

$$f(x,\lambda) = -\frac{n}{2}\ln[\sum_{i=1}^{n}\frac{(x_i(\lambda)-\overline{x}(\lambda))^2}{n}] + (\lambda-1)\sum_{i=1}^{n}\ln(x_i),$$

$$\overline{x}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}x_i(\lambda).$$

Praksē bieži vien izmanto $\lambda = 0.5$ vai $\lambda = 0$. Eksponenciāli sadalītiem datiem pielietosim transformāciju ar $\lambda = 0.5$ un atkal uzzīmēsim ROC līkni. Ir sasniegti acīmredzami rezultātu uzlabojumi 3.7..



3.7. att.: ROC līkne no $exp(40)$, $exp(5)$ un atbilstošā parametriski novērtētā ROC līkne, $n = 10$. Gadījums, kad tiek lietota Box-Cox transformācija

3.3. Neparametriskā *ROC* līkne

Viens no veidiem, kā novērtēt *ROC* līkni ir izmantot kodolu gludināšanu. Šī metode tika aprakstīta vairākās zinātniskās publikācijās un grāmatās [8] [9] [10] [11]. Definēsim gludinātas sadalījuma un sadalījuma blīvuma funkcijas, ar kurām tiek aizstātas īstās funkcijas.[12].

Definīcija 1. Pieņemsim, ka $k(x)$ ir sadalījuma blīvuma funkcija, $K(x)$ ir sadalījuma funkcija, tad gludinātas sadalījuma blīvuma un sadalījuma funkcijas būs izskatā

$$\tilde{f}_m(x) = \frac{1}{nh}\sum_{i=1}^{n} k(\frac{x - X_i}{h}),$$

$$\tilde{F}_m(x) = \frac{1}{n}\sum_{i=1}^{n} K(\frac{x - X_i}{h}).$$

, , No teorijas ir zināms, ka novērtēšanas rezultāts nav tik ļoti atkarīgs no funkcijas $k(x)$ izvēles. Svarīgākais ir pareizi izvēlēties h. Tāpēc funkcijas $k(x)$ vietā ņemsim [3]

$$k(\frac{x - y_i}{h}) = \begin{cases} \frac{15}{16}[1 - (\frac{x-y_i}{h})^2]^2 & ,\forall x \in (y_i - h, y_i + h) \\ 0 & ,\text{pretējā gadījumā} \end{cases}$$

ko var pārrakstīt kā

$$k(x) = \begin{cases} \frac{15}{16}[1 - x^2]^2 & ,\forall x \in (-1, 1) \\ 0 & ,\text{pretējā gadījumā} \end{cases}$$

Joslas platumu novērtēsim izmantojot vairākus paņēmienus : iebūvētās **R** komandas *hcv* un *bw.nrd* ('rule-of-thumb') [5], pašrakstīto kodu priekš 'cross-validation', pēc 'Normālā sadalījuma likuma' [12], kā arī pēc formulas [4]

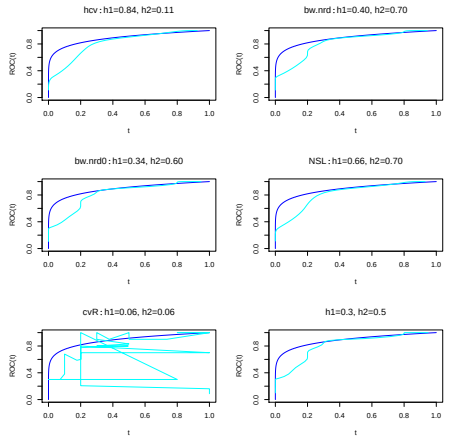
$$h = 0.9\min(SD, IQR/1.34)/n^{1/5},$$

kur SD ir standartnovirzes novērtējums no datiem un IQR ir 3. un 1. empīrisko kvartiļu starpība. Formula dod tādu pašu novērtējumu kā iebūvētā procedūra *bw.nrd*0. Atcerēsimies, ka 'cross-validation' metodes būtība ir atrast tādu h, pie kura funkcija

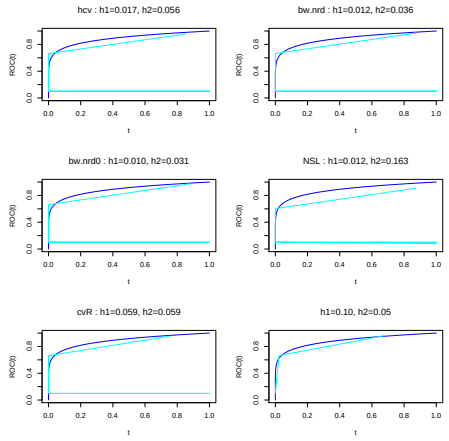
$$\frac{1}{n}\sum_{i=1}^{n}\int \hat{f}_{-i}^2(x)dx - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(y_i)$$

sasniedz savu minimālo vērtību [12]. Abos gadījumos, ģenerēsim izlases ar apjomu $n = 10$. Rezultāti ir redzami attēlos 3.8. 3.9.. Normālajam sadalījumam sanāk apmierinoši

rezultāti, ko nevar teikt par eksponenciālo. Neviena no augšminētajām joslas platuma novērtēšanas metodēm nestrādā. Vienīgā metode, ar kuru izdodas piemeklēt 'labu' h - ir 'uz aci'. Tas nozīmē, ka šāda veida sadalījumiem ir jāpiemēro vai jāizstrādā citas novērtēšanas metodes. Iespējams, problēma ir apstāklī, ka visas izmantotās metodes novērtē $F_{n_1}(\cdot)$ un $F_{n_2}(\cdot)$ atsevišķi katrai izlasei.



3.8. att.: ROC līknes no $N(0,1)$, $N(1,1)$ un atbilstošās neparametriski novērtētās $\hat{ROC}$ līknes



3.9. att.: ROC līknes no $exp(40)$, $exp(5)$ un atbilstošās neparametriski novērtētās $\hat{ROC}$ līknes

4. Ticamības intervāli

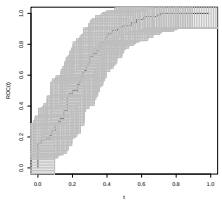
4.1. Vienlaicīgi, apvienotie ticamības intervāli

Vienlaicīgi, apvienotie ticamības intervāli izmanto Kolmogorova-Smirnova testa statistiku D, lai neatkarīgi novērtētu ticamības intervālus priekš TPF un FPF. Tādā veidā ap katru empīriskās ROC līknes lēcienpunktu tiek būvēts četrstūris, kura malu garums ir $2d$ un $2e$. $d = D/\sqrt{(n_1)}$ - ticamības intervāla platums priekš TPF un $e = D/\sqrt{(n_2)}$ - ticamības intervāla platums priekš FPF. Tālāk, lai iegūtu augšējo ticamības intervālu priekš ROC līknes, tiek savienoti četrstūru augšējās kreisās virsotnes. Priekš apakšējā ticamības intervāla iegūšanas, savieno četrstūru apakšējās labās virsotnes 4.1.. Jāņem vērā, ka šī metode var tikt izmantota priekš izlasēm ar apjomu 35 un vairāk. Jāpatur prātā arī tas, ka, ja priekš TPF un FPF tiks izvēlēts $(1-\alpha)$ ticamības rezultāts, tad ROC līknei tas būs $(1-\alpha)^2$ ticamības intervāls [13]. Ilustrēsim šo metodi ar piemēru. Ģenerēsim datus no $N(0,1)$ un $N(1,1)$. Ņemsim $n = 50$ un $n = 100$. Lai precīzāk novērtētu ticamības intervālu platumu, simulēsim Kolmogorova-Smirnova statistikas vērtības priekš $n = 50$ un $n = 100$ ar $\alpha = 0.25$ un $\alpha = 0.005$, lai $(1-\alpha)^2$ attiecīgi būtu vienāds ar 0.95 un 0.99. Simulāciju rezultātus attēlosim tabulā 4.1.. Ticamības intervālus attēlosim

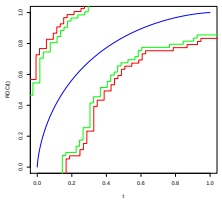
4.1. tabula Simulēta Kolmogorova-Smirnova statistika

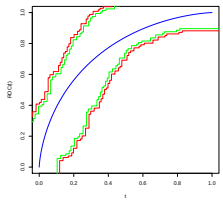
	$\alpha = 0.25$	$\alpha = 0.005$
$n = 50$	1.45	1.67
$n = 100$	1.47	1.69

zīmējumos 4.2.



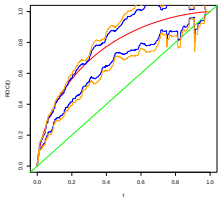
4.1. att. Vienlaicīgu, apvienoto ticamības intervālu konstrukcija

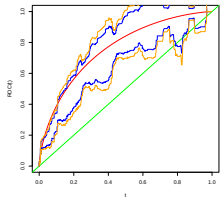




4.2. att. Vienlaicīgi, apvienotie ticamības intervāli ($n = 50, n = 100$)

4.2. Uz normālo sadalījumu balstītie ticamības intervāli





4.3. att. Uz normālo sadalījumo balstītie ticamības intervāli ($n = 50, n = 100$)

Priekš lielām n vērtībām, ROC līknes dispersija var tikt aprēķināta kā

$$var(\hat{ROC}(t) = \frac{ROC(t)(1 - ROC(t))}{n_{D=1}} + (\frac{f_{D=1}(c)}{f_{D=0}(c)})^2 \cdot \frac{t(1-t)}{n_{D=0}},$$

kur $c = S^{-1}_{D=0}(t)$, $f_{D=1}$ un $f_{D=0}$ ir sadalījuma blīvuma funkcijas.

Dispersija katrā empīriskās ROC līknes punktā var tikt novērtēta, aizstājot $f_{D=1}$ un $f_{D=0}$ ar gludinātām sadalījuma blīvuma funkcijām, ņemot t vietā vektoru FPF un priekš katra FPF aprēķinot $c = \hat{S}^{-1}_{D=0}(t)$.

ROC līknes ticamības intervāli var būt konstruēti, kā ir aprakstīts grāmatā [4].

$$\hat{ROC}(t) = \pm\Phi^{-1}(1 - \frac{\alpha}{2})\sqrt{\hat{var}(\hat{ROC})(t)}$$

Ģenerēsim datus no $N(0, 1)$ un $N(1, 1)$. Ņemsim $n = 500$ un $n = 300$. Rezultātus attēlosim 4.3.

4.3. Butstrapotie ticamības intervāli

Kā pēdējie ir konstruēti butstrapa ticamības intervāli, kuri ir aprakstīti sekojošās publikācijās [14] [15]. Šajā darbā Butstrapa metode tiek modificēta, t.i. $ROC(\hat{FPF})$ aizvietota ar $ROC(FPF)$. Pie lieliem n metodei ir jāstrādā pietiekoši labi, jo $ROC(\hat{FPF})$ ir *tuva* $ROC(FPF)$. Lai konstruētu ticamības intervālus, tiek ņemtas butstrapa izlases no datiem, katru reizi konstruējot empīrisko $\hat{ROC}_b$ līkni un aprēķinot

$$\hat{ROC}_b(FPF) - ROC(FPF), \forall FPF.$$

Procedūru atkārto 10000 reizes. Tālāk paņem 0.95 un 0.99 kvantīles no butstrapotām $\hat{ROC}(FPF) - ROC(FPF)$ vērtībām katrā FPF punktā. Rezultāti ir attēloti zīmējumos 4.4.. Lai konstruētu vienlaicīgus ticamības intervālus, katru (no 10000) aprēķina

$$sup_{FPF}(\hat{ROC}_b(FPF) - ROC(FPF)),$$

izskaitļo 0.95 un 0.99 kvantīles no iegūtām vērtībām. Rezultāti ir attēloti 4.4..

4.4. att. Punktveida Butstrapa ticamības intervāli ($n = 100, n = 50$)

4.5. att. Vienlaicīgie Butstrapa ticamības intervāli ($n = 100, n = 50$)

Secinājumi

Šajā kursa darbā tika apskatīts viens no svarīgākajiem klasificēšanas testu objektiem - *ROC* līknes. Balstoties uz matemātisko definīciju bija izpētītas tās īpašības. Bija aprakstītas sakarības starp *ROC* līknes raksturojošiem lielumiem un citiem statistiskiem objektiem. No simulēto datu piemēru pētīšanas tika izvirzīta hipotēze par to, ka apskatītas novērtēšanas un ticamības intervālu konstruēšanas metodes strādā pietiekoši labi priekš normāli sadalītiem datiem un var nedot vajadzīgus rezultātus priekš datiem no citiem sadalījumiem, piemēram, eksponenciāla. Tātad, ir plašāk jāpēta jau eksistējošās metodes un jāizstrādā citi paņēmieni, kuri var palīdzēt strādāt ar jebkura sadalījuma *ROC* līknēm.

Izmantotā literatūra un avoti

[1] Gregory Campbell Mark H.Zweig. Receiver-operating characteristic (*roc*) plots: A fundamental tool in clinical medicine. *Clinical Chemistry*, 39/4, 1993.

[2] Stephen Satcell George Christodoulakis. *The Analytics of Risk model Validation*, volume 216. Academic Press, New York, 2007.

[3] Donna K.McClish Xiao-Hua Zhou, Nancy A.Obuchowski. *Statistical Methods in Diagnostic Medicine*. A John Wiley & Sons,INC., PUBLICATIONS, New York, 2002.

[4] Margaret Sullivan Pepe. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, New York, 2004.

[5] Alan T.Arnholt Maria Dolores Ugarte, Ana F.Militino. *Probability and Statistics with R*. Taylor & Francis Group, A CHAPMAN & HALL BOOK, USA, 2008.

[6] Mehryar Mohri Gorinna Cortes. Confidence intervals for area under the *roc* curve.

[7]

[8] Zsuzsanna Horvath. Confidence bands for *roc* curves, 2000.

[9] Chris J.Lloyd. Using smoothed reciever operating characteristic curves to summarize and compare diagnostic systems. *The Annals of Statistics*, 24/1, 1998.

[10] Anindya Roy Jiezhun Gu, Subhashis Ghosal. Non-parametric estimation of *roc* curve.

[11] Zhou Yong Chris J.Lloyd. Kernel estimators of the *roc* curve are better than empirical, 1999.

[12] Larry Wasserman. *All of Nonparametric Statistics*. Springer, New York, 2006.

[13] Foster Provost Sofus A.Macskassy. *roc* confidence bands: An empirical evaluation, 2005.

[14] Nicolas Vayatis Patrice Bertail, Stephan Clemencon. On bootstrapping the roc curve.

[15] Nicolas Vayatis Patrice Bertail, Stephan Clemencon. On constructing accurate confidence bands for *roc* curves through smooth resampling.

A Izveidoto programmu kods

```
##########################################################
#####Definejam parametrus binormalajam sadalijumam#####
##########################################################
n1<-50;
n2<-50;
mi1<-0;
mi2<-1;
sd1<-1;
sd2<-1;

##############################
#####Teoretiska ROC likne#####
##############################

t=seq(0,1,by=0.0001);
ROC<-function(t)
{ROC<-1-pnorm(qnorm((1-t),mi1,sd1),mi2,sd2)
}
plot(t,ROC(t),xlim=c(0,1),ylim=c(0,1),type='l',col="red")
abline(0,1,col="green")

#########################
#####Teoretiskais AUC#####
#########################
```

```
integrate(ROC,0,1)

##############################################
#####Generejam divus normalus sadalijumus#####
##############################################
set.seed(2);
x<-rnorm(n1,mi1,sd1)
set.seed(4);
y<-rnorm(n2,mi2,sd2)
#############################
#####Empiriska ROC likne#####
#############################
minim<-min(min(x),min(y))
maxim<-max(max(x),max(y))
TPR=c();
FPR=c();
c<-sort(c(x,y));
for (i in 1:length(c))
{
TPR[i]<-length(y[y>c[i]])/length(y)
FPR[i]<-length(x[x>c[i]])/length(x)
}
points(sort(FPR),c(sort(TPR)),xlim=c(0,1),ylim=c(0,1),type="l")
#####################################################################
#####Ticamibas intervali (Kolmogorova-Smirnova) (1-alpha)^2 #####
#####################################################################
uTPR<-sort(TPR)
uFPR<-sort(FPR)
d1<-1.36/sqrt(length(uTPR))#pie alfa=0.05
d2<-1.63/sqrt(length(uTPR))#pie alfa=0.01
e1<-1.36/sqrt(length(uFPR))#pie alfa=0.05
e2<-1.63/sqrt(length(uFPR))#pie alfa=0.05
```

```
for (i in 1:length(uFPR))
{
polygon(c(uFPR[i]-d1,uFPR[i]-d1,uFPR[i]+d1,uFPR[i]+d1),c(uTPR[i]-e1,
uTPR[i]+e1,uTPR[i]+e1,uTPR[i]-e1),xlim=c(0,1),
ylim=c(0,1),border = "grey")
}

KSuFPR<-FPR-d1;
KSuTPR<-TPR+e1;
KSdFPR<-FPR+d1;
KSdTPR<-TPR-e1;
plot(sort(FPR),c(sort(TPR)),xlim=c(0,1),ylim=c(0,1),type="l")
lines(KSuFPR,KSuTPR,xlim=c(0,1),ylim=c(0,1),type="l",col="green",add=T)
lines(KSdFPR,KSdTPR,xlim=c(0,1),ylim=c(0,1),type="l",col="green")

KSuFPR<-FPR-d2;
KSuTPR<-TPR+e2;
KSdFPR<-FPR+d2;
KSdTPR<-TPR-e2;

points(KSuFPR,KSuTPR,xlim=c(0,1),ylim=c(0,1),type="l",col="red")
points(KSdFPR,KSdTPR,xlim=c(0,1),ylim=c(0,1),type="l",col="red")
lines(t,ROC(t),xlim=c(0,1),ylim=c(0,1),type='l',col="blue")
#########################
#####Empiriskais AUC#####
#########################
sTPR<-c(sort(TPR),1)
sFPR<-c(sort(FPR),1)
auc<-0;
for (i in 2:length(sTPR))
{
```

```
auc<-auc+(sFPR[i]-sFPR[i-1])*sTPR[i]
}
auc
############################
#####Mann-Whitney U test#####
############################
r<-rank(c(x,y))
u1<-sum(r[seq(along=x)])-length(x)*(length(x)+1)/2
u2<-length(x)*length(y)-u1
MWU<-u2/(length(x)*length(y))
MWU

wilcox.test(y,x)$statistic/(n1*n2)
################################################
#####Parametriska ar novertetiem parametriem#####
################################################

#####Maximum likelihood
mi1ML<-mean(x);
mi2ML<-mean(y);
sd1ML<-sd(x);
sd2ML<-sd(y);

t=seq(0,1,by=0.0001);
ROCML=1-pnorm(qnorm((1-t),mi1ML,sd1ML),mi2ML,sd2ML)
points(t,ROCML,xlim=c(0,1),ylim=c(0,1),type='l',col="blue",add=T,tpe="l")
#########################
#####Kernel smoothing#####
#########################
h1<-0.39
h2<-0.5
c<-seq(min(min(x),min(y)),max(max(x),max(y)),by=0.01);
```

```
k<-function(u){1/sqrt(2*pi)*exp((-u)^2/2)}

kod<-function(d)
{
y<-15/16*(1-d^2)^2
y[d<(-1)]<-0
y[d>1]<-0
y
}
dati1<-sort(x)
dati2<-sort(y)
fkodA<-function(x){1/(n1*h1)*sum(kod((x-dati1)/h1))}
fkodB<-function(x){1/(n2*h2)*sum(kod((x-dati2)/h2))}


#max(dati1)
hist(dati1,prob=T)
points(c,sapply(c,fkodA),type="l")
hist(dati2,prob=T)
points(c,sapply(c,fkodB),type="l")


funA1<-Vectorize(fkodA)
funA2<-function(x) integrate(funA1, -Inf, x)$value
funA3<-Vectorize(funA2)

funB1<-Vectorize(fkodB)
funB2<-function(x) integrate(funB1, -Inf, x)$value
funB3<-Vectorize(funB2)

plot(dati1,dnorm(dati1,mi1,sd1),type="l",col=3,xlim=c(-3,9))
```

```
points(dati2,dnorm(dati2,mi2,sd2),type="l",col=2)

xx<-seq(min(min(dati1),min(dati2)),max(max(dati2),max(dati1)),by=0.1)

plot(dati1,pnorm(dati1,mi1,sd1),type="l",col=2,xlim=c(min(xx),max(xx)) )
points(xx,funA3(xx),type="l",col=3)
points(dati2,pnorm(dati2,mi2,sd2),type="l",col=4)
points(xx,funB3(xx),type="l",col=5)
plot(t,ROC(t),type="l",co=4)
points(1-funA3(xx),1-funB3(xx),type="l",col=5)
####################################
########Butstrapa punktveida########
####################################
n1<-100;
n2<-100;
mi1<-0;
mi2<-1;
sd1<-1;
sd2<-1;
set.seed(2);
dati1<-rnorm(n1,mi1,sd1)
set.seed(4);
dati2<-rnorm(n2,mi2,sd2)
maxi<-Vectorize(max)
ROC<-function(t)
{ROC<-1-pnorm(qnorm((1-t),mi1,sd1),mi2,sd2)
}
m<-10000
S1<-c();
S2<-c();
for (j in 1:m)
{
```

```
x<-sample(dati1,n1,replace=TRUE)
y<-sample(dati2,n2,replace=TRUE)

minim<-min(min(x),min(y))
maxim<-max(max(x),max(y))
TPR=c();
FPR=c();
c<-sort(c(x,y));
for (i in 1:length(c))
{
TPR[i]<-length(y[y>c[i]])/length(y)
FPR[i]<-length(x[x>c[i]])/length(x)
}
roc_hat<-stepfun(sort(FPR),sort(c(TPR,1)),right=TRUE)
#plot(stepfun(sort(FPR),sort(c(TPR,1)),right=TRUE),add=T, type="l")
ROCV<-Vectorize(ROC);
ROC_TRUE<-ROCV(sort(FPR));

Starpiba1<-abs(ROCV(sort(FPR))-roc_hat(sort(FPR)))
Starpiba2<-abs(roc_hat(sort(FPR))-ROCV(sort(FPR)))
S1<-c(S1,Starpiba1);
S2<-c(S2,Starpiba2);

}
M<-c()
for (i in 1:m)
{
M[i]<-max(S1[i],S2[i])
}
M<-matrix(S,nrow=100)

TB<-c();
```

```
for (i in 1:(n1+n2))
{
TB[i]<-quantile(sort(M[i,]),probs=0.95)
}
####################################
########Butstrapa punktveida########
####################################
for (j in 1:m)
{
x<-sample(dati1,n1,replace=TRUE)
y<-sample(dati2,n2,replace=TRUE)

minim<-min(min(x),min(y))
maxim<-max(max(x),max(y))
TPR=c();
FPR=c();
c<-sort(c(x,y));
for (i in 1:length(c))
{
TPR[i]<-length(y[y>c[i]])/length(y)
FPR[i]<-length(x[x>c[i]])/length(x)
}
roc_hat<-stepfun(sort(FPR),sort(c(TPR,1)),right=TRUE)
#plot(stepfun(sort(FPR),sort(c(TPR,1)),right=TRUE),add=T, type="l")
ROCV<-Vectorize(ROC);
ROC_TRUE<-ROCV(sort(FPR));

Starpiba<-max(abs(ROCV(sort(FPR))-roc_hat(sort(FPR))),
abs(roc_hat(sort(FPR))-ROCV(sort(FPR))))

S<-c(S,Starpiba);
}
```

```
TB<-quantile(sort(S),probs=0.99)
##############################
########Uz NS balstitie########
##############################
for (j in 1:m)
{
x<-sample(dati1,n1,replace=TRUE)
y<-sample(dati2,n2,replace=TRUE)

minim<-min(min(x),min(y))
maxim<-max(max(x),max(y))
TPR=c();
FPR=c();
c<-sort(c(x,y));
for (i in 1:length(c))
{
TPR[i]<-length(y[y>c[i]])/length(y)
FPR[i]<-length(x[x>c[i]])/length(x)
}
roc_hat<-stepfun(sort(FPR),sort(c(TPR,1)),right=TRUE)
#plot(stepfun(sort(FPR),sort(c(TPR,1)),right=TRUE),add=T, type="l")
ROCV<-Vectorize(ROC);
ROC_TRUE<-ROCV(sort(FPR));

Starpiba<-max(abs(ROCV(sort(FPR))-roc_hat(sort(FPR))),
abs(roc_hat(sort(FPR))-ROCV(sort(FPR))))

S<-c(S,Starpiba);
}
TB<-quantile(sort(S),probs=0.99)
h1<-0.84
h2<-0.11
```

```
kod<-function(d)
{
y<-15/16*(1-d^2)^2
y[d<(-1)]<-0
y[d>1]<-0
y
}
dati1<-sort(x)
dati2<-sort(y)
fkodA<-function(x){1/(n1*h1)*sum(kod((x-dati1)/h1))}
fkodB<-function(x){1/(n2*h2)*sum(kod((x-dati2)/h2))}
funA1<-Vectorize(fkodA)
funB1<-Vectorize(fkodB)
cc<-(1-quantile(x,probs=FPR));
ROC<-function(t)
{ROC<-1-pnorm(qnorm((1-t),mi1,sd1),mi2,sd2)
}
ROCV<-Vectorize(ROC);
varROC<-c();
varROC<-(ROC(FPR)*(1-ROC(FPR))/n2+((funB1(cc)/funA1(cc))^2)*(FPR*(1-FPR)/n1))
TI<-c();
TI<-qnorm(1-0.01/2)*sqrt(varROC)
```

Kursa darbs "Nepārtraukto testu ROC līknes" izstrādāts LU Fizikas un Matemātikas fakultātē.

Ar savu parakstu apliecinu, ka pētījums veikts patstāvīgi, izmantoti tikai tajā norādītie informācijas avoti un iesniegtā darba elektroniskā kopija atbilst izdrukai.

Autors: Anastasija Tetereva

_____ _____

(paraksts) (datums)

Rekomendēju darbu aizstāvēšanai.
Vadītājs: doc. Dr.math. Jānis Valeinis

_____ _____

(paraksts) (datums)

Recenzents:

_____ _____

(paraksts) (datums)

Darbs iesniegts Matemātikas nodaļā _____

(datums)

_____

(darbu pieņēma)

Darbs aizstāvēts kursa gala pārbaudījuma komisijas sēdē

_____ prot. Nr. _____, vērtējums_____

(datums)

Komisijas sekretārs/-e: _____ _____

(Vārds, Uzvārds) (paraksts)