LATVIJAS UNIVERSITĀTE
FIZIKAS UN MATEMĀTIKAS FAKULTĀTE
MATEMĀTIKAS NODAĻA

DIVU KORELĒTU ROC LĪKŅU SALĪDZINĀŠANA

MAĢISTRA DARBS

Autors: **Jeļena Vaļkovska**

St. apliecība: jv07040

Darba vadītājs: doc. Dr. Math. Jānis Valeinis

RĪGA 2014

Anotācija

Tika izvēlēta tēma "Divu korelētu ROC līkņu salīdzināšana", tāpēc kā ROC līknes ir viena no plaši izmantojamiem metodēm klasifikācijas uzdevumu precizitātes noteikšanai. Darba uzdevums bija aplūkot teoriju par ROC līkņu novērtēšanas un salīdzināšanas metodēm. Darba mērķis bija apskatīt un izpētīt jauno nogludinātas džeknaifa empīriskā ticamības metodi divu ROC līkņu starpībai, kas balstās uz džeknaifa pseido izlases izveidošanu. Visi aprēķini tika veikti brīvpieejas programmā R.

Atslēgas vārdi: ROC līkne, empīriskās ticamības funkcijas metode, džeknaifa pseido izlase, klasifikācijas uzdevums

Abstract

The topic of dimploma thesis was chosen "The comparison of two correlated ROC curve", because ROC curves is one of the most popular method to evaluate the accuracy of the classification problem. Main task was to look at the theory about the estimation and comparison of ROC curves. The goal was to investigate the new method for the comparison of two ROC curves - smoothed jackknife empirical likelihood, which is based on the jackknife pseudo-samle. All calculations were performed in programm R.

Keywords: ROC curve, empirical likelihood, jackknife pseudo-sample, classification problem

Saturs

Apzīmējumi **5**

Ievads **6**

1. ROC līkņu jēdziens un galvenie raksturojošie lielumi **8**

1.1. ROC līkņu novērtēšanas metodes 9

1.1.1. Parametriskā metode 10

1.1.2. Neparametriskā metode 11

1.1.3. Citas novērtēšanas metodes 11

1.1.4. Novērtējumu salīdzināšanas mērs 13

1.1.5. Simulācijas un rezultāti 14

1.2. AUC (laukums zem līknes) 15

1.2.1. AUC novērtēšanas metodes 15

1.3. Sliekšņu izvēlē 16

1.4. Ticamības intervāli ROC līknem 18

2. ROC līkņu salīdzināšana **21**

2.1. Parametriskās metodes 21

2.2. *AUC* salīdzināšana 22

2.3. Nogludinātā JEL metode ROC līkņu starpībai 24

3. Praktiskais piemērs **27**

4. Nobeigums **31**

Izmantotā literatūra un avoti **32**

Pielikumi **34**

Apzīmējumi

EL - empīriskās ticamības metode

JEL - džeknaifa empīriskās ticamības metode

TP - pareizi pozitīvi gadījumi

FN - nepareizi negatīvi gadījumi

TN - pareizi negatīvi gadījumi

FP - nepareizi pozitīvi gadījumi

s_p - specifiskums

s_n - jūtīgums

AUC - laukums zem līknes

Ievads

Apskatīsim dažas statistikas klasifikācijas problēmas, kuras var būt sastopāmas reālaja dzīvē ikdienā:

1. nepieciešamība izvērtēt studentu pieteikumus un secināt, vai persona spēs nokārtot gala pārbaudījumus;
2. konstatēt, vai personai ir kāda noteikta slimība;
3. nepieciešamība izvērtēt, vai persona spēs atdot kredītu.

Šādas un daudzās citas statistikas klasifikācijas problēmas var būt atrisinātās, izmantojot dažādas metodes, piemēram, loģistisko regresiju, klasifikācijas kokus, neironu tīklus un citas. Tomēr parasti neviena no metodēm nedod izcilu rezultātu un daži objekti tiek nepareizi klasificēti. Tāpēc ir nepieciešams novērtēt klasifikācijas modeļa kvalitāti. Viena no plaši izmantojamām metodēm ir ROC līknes.

ROC līkni (*Receiver Operating Characteristic curve*) izgudroja elektriķi un radaru inženieri Otrajā Pasaules kara laikā kā petījuma par ar trokšņiem piesarņotu radiosignālu blakusproduktu. Kopš tā laika ROC analīze tiek plaši izmantota medicīnā, radioloģija, biometrikā un citās jomās jau kādus gadus desmitus un arvien vairāk tā tiek izmantotā mašīnu mācīšanā (*maschine learning*) un datu analīzē. Šobrīd ROC līknes ir viena no populārākiem metodem klasifikācijas uzdevuma precizitātes noteikšanai.

Tika izpētītas un ievestas vairākās gan parametriskās, gan neparametriskās metodes ROC līkņu novērtēšanai un salīdzināšanai. Pedējā laikā lielu uzmanību pievērsa džeknaifa empīriskas ticamības funkcijas metode. Tas būtiski uzlabo empīriskās ticamības funkcijas metodes skaitļošanas efektivitāti, jo tiek samazināts parametru skaits. Jing et al. [8] piedāvāja džeknaifa empīsriskās ticamības funkcijas metodi U-statistikai. Gong et al. [5] izstrādājis nogludināto džeknaifa empīriskās ticamības metodi (JEL) ROC līknei, Yang un Zhao et al. [13] izstrādāja nogludināto džeknaifa empīriskās ticamības funkcijas metodi ticamības intervālu konstruēšanai divu ROC līkņu starpībai.

Šī darba uzdevumi ir:

1. iepazīties ar teoriju par ROC līknem;
2. apskatīt ROC līkņu galvenus raksturojošus lielumus un to novērtēšanas metodes;
3. apskatīt ROC līkņu salīdzināšanas metodes;
4. pielietot metodes reālai datu problēmai.

Darbs sastāv no 4 nodaļām un viena pielikuma. Pirmajā nodaļā ir definēti pamatjēdzieni, apskatītas ROC līkņu novērtēšanas metodes un galvenie raksturojošie lielumi.

Otrajā nodaļā ir apskatītas ROC līkņu salīdzināšanas metodes. Trešajā nodaļa ir veikts metožu praktiskais pielietojums un izdarīti secinājumi. Ceturtajā nodaļā ir aprakstīti galvenie darbā iegūtie rezultāti un secinājumi.

1. ROC līkņu jēdziens un galvenie raksturojošie lielumi

Pieņemsim, ka diagnostikas tests ir ar nepārtrauktu rezultātu T. Pamatojoties uz vērtību T klasificēsim objektus divās grupās. Uzskatīsim, ka tests ir pozitīvs tad un tikai tad, ja T vērtība ir vienāda vai lielāka par kādu konstantes c vērtību, ko sauksim par slieksni. Uzskatīsim, ka gadījums ir negatīvs tad un tikai tad, ja T vērtība ir mazāka par c. Diagnostikas testa precizitāte var būt apskatīta pamatojoties uz divām rādītājiem: pareizi klasificēto pozitīvo gadījumu daļa (*True Positive Rate*) TPR, un nepareizi klasificēto pozitīvo gadījumu daļa (*False Positive Rate*) FPR. Ja testa iznākums ir pozitīvs un arī īstā vērtība ir pozitīva, tad šādu gadījumu sauc par pareizi pozitīvu (*True Positive*) TP; bet ja īstā vērtība ir negatīva , tad to sauc par nepareizi negatīvo gadījumu (*False Positive*) FP. Ja gan īstā vērtība, gan testa rezultāts ir negatīvs, tad runa ir par pareizi negatīvo gadījumu (*True Negative*) TN, bet ja īstā vērtība ir pozitīva, tad tas ir nepareizi negatīvs gadījums (*False Negative*) FN (skat. 1. tabulu).

1. tabula: Diagnostikas testa iznākumi un īstās vērtības

	Īstās vērtības	
	Pozitīvi	Negatīvi
Tests Pozitīvi	TP	FP
Tests Negatīvi	FN	TN

ROC līkne ir grafiks, kurš atpoguļo pareizi klasificēto pozitīvo gadījumu daļu TPR (uz OY ases) pret nepareizi klasificēto pozitīvo gadījumu daļu FPR (uz OX ases), visiem iespējamiem sliekšņa vērtībam. Testa jūtīgums (*sensitivity*) s_n ir vērtība, kura parāda cik labi klasifikators var pareizi atšķirt pozitīvus gadījumus, t.i. pareizi pozitīvo klasificēto gadījumu daļa TPR. Citiem vārdiem sākot, jūtīgumu sastāda gadījumi, kuri ar testa palīdzību tika izvēlēti kā pozitīvie, pret visiem gadījumiem, kuri patiešām ir pozitīvie

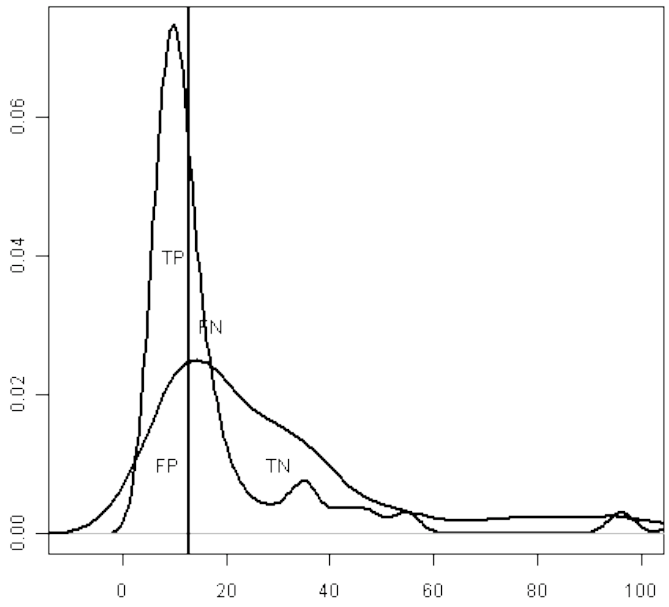
$$s_n = TP/P = TP/(TP + FN) = TPR.$$

Savukārt specifiskums (*specifisity*) s_p ir testa spēja atšķirt negatīvus gadījumus, t.i. pareizi klasificēto negatīvu gadījumu daļa

$$s_p = FP/N = TN/(FP + TN) = 1 - FPR.$$

Šie divi mēri ir cieši saistīti ar I un II veida kļūdām.

1. attēlā ir redzamas testa vērtības sadalījuma blīvuma funkcijas divam kategorijam "aizkuņģa dziedzera vēža seruma biomarkeri" [12] otrā biomarkera $CA-19-9$ datiem un atbilstošie TP, TN, FP, FN gadījumi pie noteiktās sliekšņa vērtības 12.8.



1. att.: "aizkuņģa dziedzera vēža seruma biomarkeri" [12] otrā biomarkera $CA-19-9$ sadalījuma blīvuma funkcija un optimālāka sliekšņa vērtības pēc Džoudena indeksa 12.8, atbilstošie TP, TN, FP, FN

1.1. ROC līkņu novērtēšanas metodes

Šajā nodaļā apskatīto jēdzienu definēšanai tika izmantoti literatūras avoti [4] un [7]. Pieņemsim, ka X un Y ir nepārtraukta diagnostikas testa iznākumi. Apzīmēsim ar F un G attiecīgi X un Y kumulatīvas sadalījuma funkcijas. Tad testa ROC līkne $R(p)$ ir

$$R(p) = 1 - G(F^{-1}(1-p)), \tag{1}$$

kur $p \in (0,1)$ ir $(1-s_p)$ vērtība pie noteiktās sliekšņa c vērtības.

Praksē bieži vien īstā ROC līkne nav zināma un lai varētu salīdzināt vairākus testus savā starpā, ROC līkne ir jānovērtē.

Apskatīsim dažas ROC līkņu novērtēšanas metodes.

1.1.1. Parametriskā metode

Parametriskā metode tiek izmantotā gadījumā, kad sadalījuma funkcijas F un G ir zināmas. Ja dati ir normāli sadalīti, tad tiek izmantots binormālais modelis. Pretējā gadījumā, no sākumā dati jātransformē.

Pieņemsim, ka X un Y ir neatkarīgi un normāli sadalīti mainīgie ar vidējām vērtībam μ_1 un μ_2 un dispersijam σ_1^2 un σ_2^2, attiecīgi. Tad ROC līkne ir

$$R(p) = \Phi(a + b\Phi^{-1}(p)), \tag{2}$$

kur Φ ir standarta normālā sadalījuma kumulatīva sadalījuma funkcija un $a = \frac{\mu_1 - \mu_2}{\sigma_2}$, $b = \frac{\sigma_1}{\sigma_2}$ un $0 < p < 1$.

Parametrus a un b var novērtēt tieši no testa iznākumu pirmās un otrās grupas subjektu sadalījuma vidējās vērtības un dispersijas. Tātad $\hat{a} = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}_2}$ un $\hat{b} = \frac{\hat{\sigma}_1}{\hat{\sigma}_2}$, kur $\hat{\mu}_i$ ir testa rezultātu vidējā vērtība un $\hat{\sigma}_i$ ir dispersija, $i = 0, 1$.

Ja testa rezultāti nav normāli sadalīti, tad no sākuma tiek pielietota cita metode, kas balstās uz Boksa-Koksa (Box-Cox) transformāciju:

$$X^\lambda = \begin{cases} \frac{X^\lambda - 1}{\lambda}, & \text{ja } \lambda \neq 0, \\ \log(X), & \text{ja } \lambda = 0. \end{cases}$$

Parametru λ var novērtēt, izmantojot, piemēram, maksimālās ticamīas metodi (MLE). Kad tika iegūts parametra λ novērtējums, tad jātransformē testa dati. Šādā veidā varētu būt sasniegts tas, ka datiem ir binormālais sadalījums. Pēc šīs procedūras jau transformētus datus izmanto parametru novērtēšanā.

Apskatīsim arī citu metodi parametru novērtēšanai. No sākuma pārrakstīsim formulu (2) šādā veidā

$$\Phi^{-1}(R(p)) = a + b\Phi^{-1}(p). \tag{3}$$

Tad, lai noteiktu parametrus a un b, tiek izmantota lineārā regresija starp kvantiļu vērtībām.

ROC līknes raksturīga īpašība ir izliektība, jo tad tiek garantēts, ka līkne nekad nešķērsos diagonāli. Bet lietojot binormālo modeli, var rasties problēma, ka ROC līkne nav izliekta intervālā $(0, 1)$, ja $b = 1$.

1.1.2. Neparametriskā metode

Izmantojot formulā (1) sadalījuma funkcijas F un G empīriskus novērtējumus $\hat{F}_m(x) = \frac{1}{m}\sum_{i=1}^{m} I_{(X_i \leq x)}$ un $\hat{G}_n(y) = \frac{1}{n}\sum_{j=1}^{n} I_{(Y_j \leq y)}$ attiecīgi, iegūsim ROC līknes empīrisko novērtējumu

$$\hat{R}(p) = 1 - \hat{F}_m(\hat{G}_n^{-1}(1-p)). \tag{4}$$

Empīriskā ROC līkne saglabā daudzās empīriskās sadalījuma funkcijas īpašības un tā ir vienmērīgi konverģenta ar teorētisko līkni. Šāds novērtējums ir robusts un viegli aprēķināms. Tomēr novērtējumam ir daži trūkumi, piemēram, novērtējumam var būt liels svārstīgums, it īpaši ja izlases apjoms nav liels [10]. Tas nerada īpašas problēmas nozarēs, kur ir pieejams liels datu apjoms, piemēram, sociālās vai finanšu nozarēs, bet medicīnā šāds novērtējums var būt neatbilstošs, mazu datu apjomu dēļ. Papildus, novērtētā ROC līkne nav nepārtrauktā, līdz ar to rezultātu interpretācija var būt sarežģīta.

Citas metodes tika izstrādātas, lai iegūtu nogludinātu ROC līkni, izmantojot kodolu gludināšanu, nogludinātu empīrisku sadalījuma funkciju vai log - izliektu funkciju.

1.1.3. Citas novērtēšanas metodes

Viens no citiem ROC līkņu novērtēšanas veidiem ir gludināšana ar kodoliem. Formulā (1) sadalījuma funkcijas īstās vērtības tiek aizvietotās ar nogludinātām sadalījuma blīvuma un sadalījuma funkcijas.

Definīcija 1. Pieņemsim, ka $X_1, ..., X_m \sim F$ un $Y_1, ..., Y_n \sim G$ ir savstarpēji neatkarīgi novērojumi. Tad nogludinātas sadalījuma blīvuma un sadalījuma funkcijas būs

$$\hat{f}_m(x) = \frac{1}{mh_1}\sum_{i=1}^{m} k\left(\frac{x - X_i}{h_1}\right) \text{ un } \hat{g}_n(x) = \frac{1}{nh_2}\sum_{i=1}^{n} k\left(\frac{y - Y_i}{h_2}\right),$$

$$\hat{F}_m(x) = \frac{1}{m}\sum_{i=1}^{m} K\left(\frac{x - X_i}{h_1}\right) \text{ un } \hat{G}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{y - Y_i}{h_2}\right),$$

kur $h_j,\ j = 1, 2$ ir joslas platums, kurš kontrolē gludināšanas daudzumu. Kodolu gludināšanas metodē kodolu izvēlei nav tik lielas nozīmes. Svarīgākais ir joslas platuma h izvēle.

Rufibah et al. [11] piedāvāja alternatīvu metodi, kad sadalījuma funkcijas F un G ir modelēti neparametriski, bet kodolu vietā tiek izmantots log - izliekts sadalījuma blīvuma funkcijas novērtējums.

Definīcija 2. [3] Pieņemsim, ka t ir blīvuma funkcija telpā $\mathbb{R}$. t sauc par log-izliektu, ja eksistē funkcija $\varphi : \mathbb{R} \to (-\infty, \infty)$ tāda, ka

$$t(x) = e^{\varphi(x)}$$

kādai izliektai funkcijai $\varphi : \mathbb{R} \to (-\infty, \infty)$ un $x \in \mathbb{R}$. $V_1, ..., V_n \in \mathbb{R}$ no t ir neatkarīgi un vienādi sadalīti, tad t var novērtēt, maksimizējot normalizēto log - ticamības funkcju

$$l(\varphi) = n^{-1} \sum_{i=1}^{n} \log t(V_i) = n^{-1} \sum_{i=1}^{n} \varphi(V_i)$$

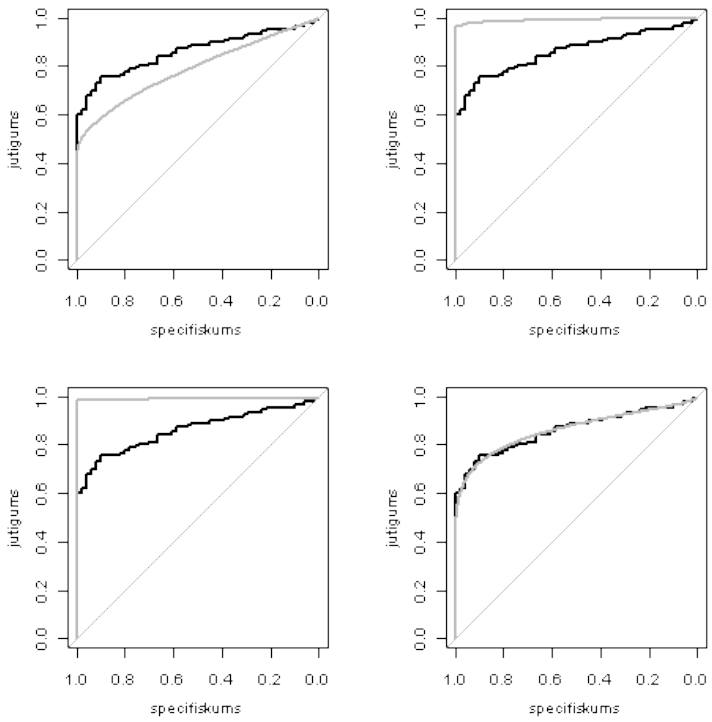
pēc visām izliektām funkcijām $\varphi : R \to (-\infty, \infty)$ tādam, ka

$$\int_{-\infty}^{\infty} e^{\varphi(x)} dx = 1.$$

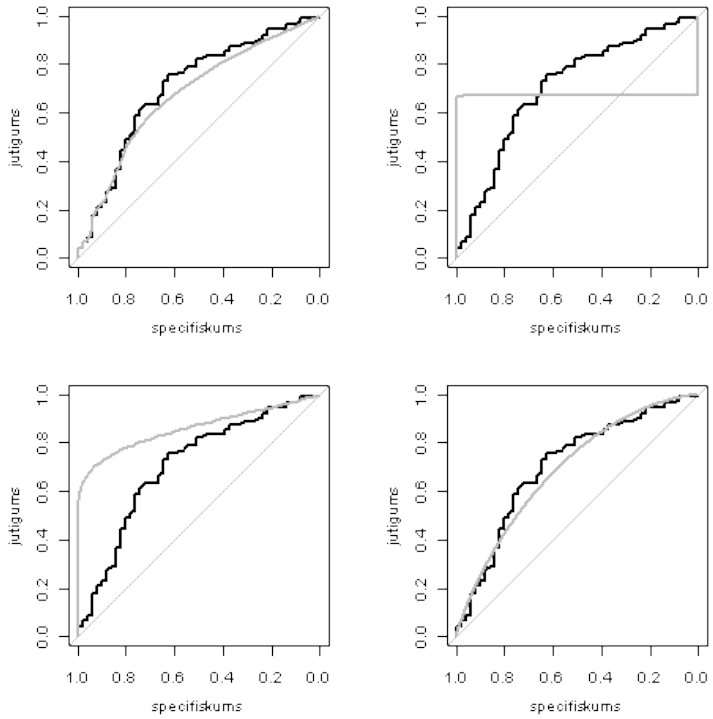
Pieņemsim, ka $X_1, ..., X_m \sim F$ un $Y_1, ..., Y_n \sim G$ ir nepārtraukti, vienādi sadalīti novērojumi. Aprēķinot log - izliektas sadalījuma blīvuma funkcijas $\hat{f}_m$ un $\hat{g}_n$ un atbilstošās log - izliektas sadalījuma funkcijas $\hat{F}_m$ un $\hat{G}_n$ un ievietojot tos izteiksmē (1), iegūsim ROC līknes log-izliektu novērtējumu:

$$\hat{R}_{m,n}(p) = 1 - \hat{F}_m(\hat{G}_n^{-1}(1-p)),$$

kur $0 < p < 1$.



2. att.: "aizkuņģa dziedzera vēža seruma biomarkeri" dati pirmā biomarkera $CA - 125$ pielietotās dažadas ROC līkņu novērtēšanas metodes: nogludināts ar kodolu; log - izliekts novērtējums; parametriskais novērtējums; binormālais novērtējums.



3. att.: "aizkuņģa dziedzera vēža seruma biomarkeri" dati otrā biomarkera $CA-19-9$ pielietotās dažadas ROC līkņu novērtēšanas metodes: nogludināts ar kodolu; log - izliekts novērtējums; parametriskais novērtējums; binormālais novērtējums.

2. un 3. attēlā var redzēt ar dažādam metodēm novērtētās ROC līknes. Varam secināt, ka ar kodolu nogludinātas ROC līknes abos gadījumos ir pārgludinātās. Pirmajam biomarkerim log - izliekts novērtējums ir līdzīgs binormālām parametriskām.

1.1.4. Novērtējumu salīdzināšanas mērs

Viens no veidiem, kā var salīdzināt cik tuvu ROC līknes novērtējums ir īstajai ROC līknei ir vidējā kvadrātiskā kļūda (*the average square error*) ASE

$$ASE(\hat{R}) = n^{-1}\sum_{i=1}^{n}(\hat{R}(u_i) - R(u_i))^2,$$

kur u_i ir punkti no intervālā $[0,1]$, kuri viena no otra atrodas vienādā attālumā. Un vidējās kvadrātiskas kļūdas attiecību definēsim šādi

$$ASER(\hat{R}) = \sqrt{\frac{ASE(\hat{R})}{ASE(\hat{R}_{m,n})}}.$$

Šī testa vērtība parāda, vai attiecīgais novērtējums ir labāks nekā empīriskās ticamības novērtējums.

Cits mērs ir integrēta absolūtā starpība (*integrated absolute difference*) IAD

$$IAD(\hat{R}) = \int_0^1 |R(p) - \hat{R}(p)|dp.$$

1.1.5. Simulācijas un rezultāti

Lai varētu salīdzināt, kāda ROC līkņu novērtēšanas metode ir labāka, tika apskatītās izlases ar dažādiem sadalījumiem un tika veiktas simulācijas. Tika aprēķinātās šādas vērtības:

1. $ASER$ vidējā vērtība katram veidam. ROC līknes novērtējums ir labāks par empīrisko novērtējumu, ja $ASER$ vērtība ir mazāka par 1.

2. Gadījumu daļa, kad novērtēšanas veids bija sliktāks par empīrisko novērtējumu, i.e. $ASER > 1$. Pieņemsim, ka ROC līknes novērtējums ir labāks par empīrisko novērtējumu, ja šī vērtība ir mazākā par 50%.

3. IAD vidējā vērtība atram veidam. Jo tūvāks ROC lķnes novērtējums ir īstajai vērtībai, jo tūvāka nullei būs IAD vērtība.

Lai veiktu simulācijas, tika izvēlēti dažādi sadalījumi: vienmērīgais $U(a, b)$ ar augšejo robežu a un apakšējo robežu b; normālais $N(\mu, \sigma)$ ar vidējo vērtību μ un standartnovirzi σ; eksponenciālais $Exp(a)$ ar koeficientu a; gamma $\Gamma(\alpha, \beta)$ ar parametriem α un β un logistiskais $Log(\mu, s)$. Visi šie sadalījumi ir log - izliekti.

Tika veidotas izlases ar apjomu 100 no katra sadalījuma 2. tabulā ir ir apkopota informācija par simulācijas stratēģiju.

2. tabula: Simulāciju scenārijs: izlases tika ģenerētas ar apjomiem m un n no sadalījumiem $U(a, b)$, $N(\mu, \sigma)$, $Exp(a)$, $\Gamma(\alpha, \beta)$

F	m	G	n
$U(0, 3)$	100	$U(-2, 1)$	100
$N(2, 1)$	100	$N(0, 1)$	100
$Exp(2)$	100	$Exp(1)$	100
$\Gamma(4, 1.5)$	100	$\Gamma(2, 0.5)$	100
$Log(2, 1)$	100	$Log(0, 1)$	100

Tika aprēķināta īstā ROC līkne un iegūti ROC līkņu novērtējumi, izmanotjot empīrisko, parametrisko, binormālo, log - izliektu metodi un gludināšanu ar kodoliem. Pēc tam tika aprēķinātas $ASER$ un IAD vērtības. Procedūra tika atkartotā 1000 reizes. Rezultāti ir apkopoti 3. tabulā.

No iegūtiem rezultātiem varam secināt, ka IAD vidējās vērtības gandrīz visiem novērtējumiem ir apmēram vienādas. Izņēmums ir nogludinātam ar kodolu novērtējumam. Tas varētu būt izskaidrojams ar to, ka ROC līkne ir pārgludināta, tātad jāpielieto cits joslas platums. Vislābākais ROC līknes novērtējums, salīdzinot pēc IAD vērtībam ir log

- izliekts novērtējums. Salīdzinot novērtējumus pēc vidējās $ASER$ vērtības, redzam, ka neviena no apskatītajiem metodem nebija labāka par empīrisko novērtējumu.

3. tabula: Simulācijas rezultāti. Izlases ajomi $m = 100$, $n = 100$, simulācijas atkārtotās 1000 reizes

Novērtējums	M($ASER$)	$ASER > 1$ (%)	M(IAD)
Empīriskais	–	–	0.1616
Parametriskais	1.9701	59.84	0.1765
Binormālais	1.9018	43.64	0.1619
Log - izliekts	1.6419	23.48	0.0739
Nogludināts ar kodolu	1.9193	32.45	0.3296

1.2. AUC (laukums zem līknes)

Skaitlisko ROC liknes interpretāciju dod vērtība AUC, t.i. laukums zem līknes (*the area under the ROC curve*), kas ir definēta šādi

$$AUC = \int_0^1 R_{m,n}(p)dp.$$

Tātad AUC ir laukums, ko ierobežo ROC līkne un nepareizi klasificēto pozitīvo gadījumu daļas FPR ase. Jo AUC vērtība ir lielākā, jo precīzāk ir klasifikators. Ekvivalentas AUC vērtības reprezentē līdzīgu testa precizitāti, bet tas neobligāti nozīme, ka ROC līknes ir identiskās, tās var būt sakrustotās vai arī tam var būt dažādas formas. Lai aprēķinātu AUC ir izstrādātas vairākās metodes. Apskatīsim dažus no tiem [6].

1.2.1. AUC novērtēšanas metodes

Parametriskā metode tiek lietota, kad testa iznākumu statistiskais sadalījums ir zināms. Ja abām grupam ir normālais sadalījums, tad parasti tiek izmantots binormālais sadalījums. Ja dati ir binormāli vai var būt pārveidoti par binormāliem, izmantojot log, kvadrātisku vai Boksa-Koksa transformāciju (Box-Cox transformation), tad attiecīgi parametri var būt viegli novērtēti ar pirmās un otrās grupas subjektu vidējo vērtību un dispersiju. Pieņemot, ka $X_1, ..., X_m$ un $Y_1, ..., Y_n$ ir neatkarīgi, vienādi sadalīti un $X \sim N(\mu_1, \sigma_1^2)$ un $Y \sim N(\mu_2, \sigma_2^2)$, tad AUC vērtība ir

$$AUC = \Phi\left(\frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}\right),$$

un to var novērtēt, ievietojot formulā izlases vidējo vērtību un standatnovirzi. Ja dati nav normāli sadalīti, tad no sākuma tiek pielietota Boksa-Koksa transformācija:

$$X^\lambda = \begin{cases} \frac{X^\lambda - 1}{\lambda}, & \text{ja } \lambda \neq 0, \\ \log(X), & \text{ja } \lambda = 0, \end{cases}$$

kur ir pieņemts, ka $X^\lambda \sim N(\mu_0, \sigma_0^2)$ un $Y^\lambda \sim N(\mu_1, \sigma_1^2)$. Jāpiemin, ka

$$AUC = P(X > Y) = P(X^\lambda > Y^\lambda).$$

Savukārt neparametriskā metode nebalstās uz zināmu sadalījumu un rezultātā iegūto AUC sauc par empīrisku. Vienā no metodēm kā neparametriski var aprēķināt AUC vērtību ir trapeces likums. Galvenā ideja ir aprēķināt laukumu figūrai, kura tiek iegūta, pievienojot katrā nepārtrauktā testa intervālā punktu $(s_n, 1 - s_p)$ un savienojot to punktu ar taisnu līniju ar OX asi. Šāda veida tiek veidotas trapeces un to laukumus var viegli aprēķināt. Sasummējot laukumus tiek iegūts AUC novērtējums.

Citas metodes lieto Manna-Vitnēja U-statistiku (Mann-Whitney U-statistic), kura ir pazīstama kā Vilkoksona (Vilcoxon) rangu-summas statistika ,un c-indeksu. Abas metodes tika atzītās par ekvivalentam. Novērtējot ROC līkni ar empīrisku metodi, atbilstošs empīriskais AUC novērtējums ir

$$AUC = \int_0^1 \hat{R}_{m,n}(p) dp.$$

AUC kodolu novērtējums, izmantojot Gaussa kodolu, ir šāds

$$AUC = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} \Phi \left(\frac{X_i - Y_i}{\sqrt{h_1^2 - h_2^2}} \right).$$

1.3. Sliekšņu izvēlē

Ideālajā gadījumā, ROC līkne ies caur punktu $(0, 1)$. Tātad eksistē tāda sliekšņa c vērtība, kurai $TPR(c) = 1$ un $FNR(c) = 0$, tas nozīmē, ka bez kļūdam var sadalīt visus gadījums divās grupās. Jo tuvāk ROC līkne ir šim ideālajam punktam, jo labākas ir testa klasificēšanas spējas. Testam, kurš nevar atšķirt gadījumus, ROC līkne būs diagonāle. Šāda situācija ir ekvivalenta gadījuma minēšanai.

Var gadīties situācijas, kad laukums zem divām ROC līknem ir apmēram vienāds, bet ROC līknem ir dažādas formas vai līknes savstarpēji krustās. Šādās situācijās pētnieks var būt ieinteresēnts aplūkot konrētās jūtīguma vai specifiskuma vērtībās. Piemēram, aplūkojot problēmu, kad nepieciešams noteikt, vai pacientam ir vēzis, ir noderīgi aplūkot lielas specifiskuma vērtības.

Lai atrast optimālo slieksni galvēnokārt tiek izmantoti trīs kritēriji. Pirmās divas metodes dod līdzīgu jūtīgumu un specifiskumu un neuzliek nekādus ierobežojumus uz

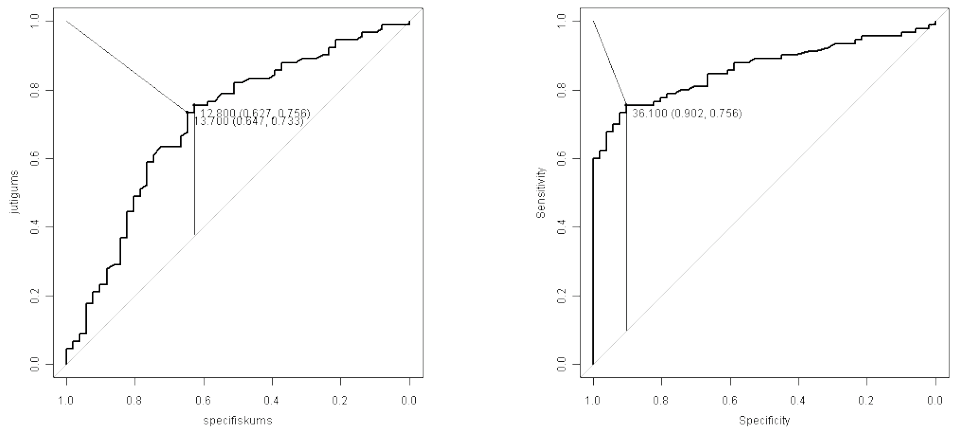
izdevumiem. Trešā kritērijā tiek ņemti vērā izdevumi, kas galenokārt ietver finansiālus izdevumus, noteicot pareizu un nepareizu diagnozi, izmaksas par personu, kura tiek arstētā, diskomfortu un pārējās izmaksas tālākai izmēklēšanai, ja tāda būs nepieciešama. Šo metodi ļoti rēti izmanto medicīnā, tāpēc kā šādas izmaksas ir diezgan grūti novērtēt.

Šie trīs kritēriji ir pazīstami kā līknes vistūvākais punkts punktam $(0, 1)$, Džoudena indekss (Youden index) un izmaksas minimizācijas kritērijs, attiecīgi. Ja s_n un s_p ir attiecīgi jūtīgums un specifiskums, tad attālums no punkta $(0, 1)$ līdz jebkuram punktam uz ROC līknes ir

$$d = \sqrt{\left[(1 - s_n)^2 + (1 - s_p)^2\right]}.$$

Lai aprēķinātu optimālo sliekšņa vērtību un atšķirt slimos subjektus no neslimiem, jāaprēķina šīs attālums katram punktam un jāizvēlās punkts, kurā attālums d ir vismazākais.

Otrais kritērijs ir Džoudena indekss, kurš maksimizē vertikālo attālumu no ekvivalences līnijas (līnija, kura atbilst gadījuma minēšanai, t.i. diagonāle) līdz punktam $[x, y]$, kur x ir $(1 - s_p)$ un y ir s_n. Citiem vārdiem sākot, Džoudena indekss J ir punkts uz ROC līknes, kurš ir vistālaiks no ekvivalences līnijas (no diagonāles). Galvēnais Džoudena indeksa mērķis ir maksimizēt starpību starp $TPR(s_n)$ un $FPR(1 - s_p)$. Tad $J = \max(s_n + s_p)$. Džoudena indekss tiek plaši lietots, tāpēc kā tās mērķis ir maksimizēt pareizi klasificēto subjektu daļu un ir vienkārši aprēķināms [9].



4. att.: Pirmā un otrā biomarkera optimālākās sliekšņa vērtības pēc Džoudena indeksa un minimālā attāluma metodes

Apskatīsim, ka strādā teorija praksē. Izmantosim "aizkuņģa dziedzera vēža seruma biomarkeri" datus. (skat. 4. attēlu). Izmantojot vimazākā attāluma metodi, iegūvam, ka pirmā biomarkera optimālāka sliekšņa vērtība ir vienāda ar 36.1, šajā punktā specifiskums ir vienāds ar 0.90196 un jūtīguma vērtība ir 0.7556. Otrā biomarkera optimālākā sliekšņa vērtība ir 13.7, šajā punktā specifiskums ir vienāda ar 0.647 un jūtīguma vērtība ir 0.733. Salīdzināsim šos rezultātus ar Džoudena indeksa metodi. Tad pirmajām biomarkerim optimālāka sliekšņa vērtība ir tāda pati 36.1, bet otrajām biomarkerim optimālāka sliekšņa

vērtība būs 12.8. Specifiskums un jūtīgums šājā punktā ir vienādi attiecīgi ar 0.6275 un 0.7556.

1.4. Ticamības intervāli ROC līknem

Pieņemsim, ka F un G ir attiecīgi slimo un neslimo subjektu populācijas sadalījuma funkcijas. Īstā ROC līkne ir

$$R(p) = 1 - F(G^{-1}(1-p)),$$

kur $0 < p < 1$. Pieņemsim, ka $X_1, ..., X_m$ ir m neatkarīgi un vienādi sadalīti ar sadalījuma funkciju F diagnostiskā testa iznākumi slimo subjektu populācijas rezultāti un attiecīgi $Y_1, ..., Y_n$ ir n neatkarīgi un vienādi sadalīti ar sadalījuma funkciju G diagnostiskā testa iznākumi slimo subjektu populācijas rezultāti. Definēsim ROC līknes empīrisko novērtējumu

$$\hat{R}_{m,n}(p) = 1 - \hat{F}_m(\hat{G}_n^{-1}(1-p)),$$

kur $\hat{F}_m$ un $\hat{G}_n$ ir attiecīgi funkciju F un G empīriskie novērtējumi. Tātad

$$\hat{F}_m(x) = \frac{1}{m}\sum_{j=1}^{m} I_{(X_j \le x)},$$

$$\hat{G}_n(x) = \frac{1}{n}\sum_{j=1}^{n} I_{(Y_j \le y)}.$$

Izmantojot faktu, ka

$$\sqrt{m+n}\{R_{m,n}(p) - R(p)\} \xrightarrow{d} N(0, \sigma^2(p)),$$

kur

$$\sigma^2(p) = \left(1 + \frac{1}{r}\right) R(p)(1-R(p)) + (1+r)p(1-p)\left\{\frac{F'(G^{-1}(1-p))}{G'(G^{-1}(1-p))}\right\}^2,$$

un

$$r := \lim_{m,n\to\infty} \frac{m}{n} \in (0, \infty)$$

ROC līknei $R(p)$ ir iespējams uzkonstruēt ticamības intervālus.

Ir dažādas metodes ticamības intervālu ROC līknei konstruēšanai. Piemēram, novērtējot funkciju F un G blīvuma funkcijas vai izmantojot bustrapa (bootstrap) metodes. Alternatīvais veids ticamības intervālu konstruēšanai, nevērtējot asimptotisko dispersiju, ieviesa Claeskens et al. ([1]), kurš piedavāja empīrisko ticamības metodi, kas balstās uz funkciju F un G gludināšanu, izmantojot dažus saistītus mainīgus. Molanes-Lopez, Van Keilegom and Veraverbeke pētīja empīriskos ticamības metodi, kas balstās uz empīriskiem novērtējumiem. Qin and Zhou izstrādāja empīriskās ticamības metodi ticamības intervalu konstruēšanai laukumama zem ROC līknes (AUC).

Diezgan nesen Jing Yuan un Zhou [8] prezentēja džeknaifa empīriskās ticamības metodi U-statistikai. Procedūra ir sekojoša. No sākuma U-statistikai tiek konstruēta pseido-izlase. Pēc tam šī pseido-izlase tiek izmantotā kā parasta izlase ar neatkarīgiem un vienādi sadalītiem gadījumiem lielumiem. Lai iegūtu empīriskā ticamības attiecību statistiku par U-statistiku, izlasei vidējai vērtībai tiek pielietots parasta empīriskās ticamības metode.

Apskatīsim šo metodi ticamības intervālu konstruēšana ROC līknei. Pieņemsim, ka w ir simetriskā blīvuma funkcija $[-1, 1]$, tad nogludināts ROC līknes $\hat{R}_{m,n}(p)$ novērtējums ir

$$\hat{R}_{m,n}(p) = 1 - \frac{1}{m}\sum_{j=1}^{m} K\left(\frac{1-p-G_n(x_j)}{h}\right),$$

kur $h = h(n) > 0$ ir joslas platums. Definēsim

$$\hat{R}_{m,n,i}(p) = \frac{1}{m-1}\sum_{j=1,\, j\neq i}^{m} K\left(\frac{1-p-G_{n,m}(x_j)}{h}\right)$$

bet ja $m+1 \leq i \leq m+n$, tad novērtējums ir

$$\hat{R}_{m,n,i}(p) = \frac{1}{m-1}\sum_{j=1}^{m} K\left(\frac{1-p-G_{n,m-i}(x_j)}{h}\right),$$

kur

$$G_{n,-i}(y) = \frac{1}{n}\sum_{j=1,\, j\neq i}^{n} I_{(y_j \leq y)},\ i = 1, \ldots, m+n$$

Džeknaifa pseido-izlase ir

$$\hat{V}_i(p) = (m+n)\hat{\Delta}_{m,n}(p) - (m+n-1)\hat{\Delta}_{m,n,i}(p),\ i = 1, \ldots, m+n. \tag{5}$$

Ar N apzīmēsim visu novērojumu skaitu. Empīriskās ticamības funkcijas punktā p, kura ir balstīta uz pseido-izlasi būs

$$L(\theta, p) = \sup\left\{\prod_{i=1}^{N} p_i : \sum_{i=1}^{N} p_i = 1, \sum_{i=1}^{N} p_i \hat{V}_i(p) = \theta,\ p_i > 0,\ i = 1, \ldots, N\right\}.$$

Izmantojot Lagranža reizinātāju metodi, iegūst, ka maksimālā vērtība tiek iegūta, ja

$$p_i = \frac{1}{(m+n)\{1 + \lambda(\hat{V}_i(p) - \theta)\}},\ \ i = 1,\ \ldots,\ m+n,$$

kur $\lambda = \lambda(p, \theta)$ apmierina vienādojumu

$$\sum_{i=1}^{N} \frac{\hat{V}_i(p) - \theta}{1 + \lambda(\hat{V}_i(p - \theta))} = 0.$$

Lai atrast λ no sākuma jānodefinē intervāls, kurā λ jāmeklē. Izmantojot nosacījumus, ka $p_i > 0$ un $p_i < 1$, $\forall i$, tad līdzīgi, ka vislielākās ticamības attiecību testa vidējai vērtībai gadījumā, tiek iegūts, ka

$$\frac{1 - n^{-1}}{\Delta - \max(\hat{V}_i)} < \lambda < \frac{1 - n^{-1}}{\Delta - \min(\hat{V}_i)} \tag{6}$$

Log-empīriskās ticamības attiecība ir

$$l_{m,n}(\theta, p) = -2 \log L(\theta, p) = 2 \sum_{i=1}^{N} \log(1 + \lambda(\hat{V}_i(p - \theta)), \tag{7}$$

Lai pierādītu, ka log - empīriskās ticamības attiecības konverģē pēc sadalījuma uz Hī-kvadrātu ar vienu brivības pakāpi, jāpierāda, ka džeknaifa dispersijas novērtējums

$$v_{m,n}(p) = \frac{1}{m+n} \sum_{i=1}^{N} \left(\hat{V}_i(p) - \frac{1}{m+n} \sum_{i=1}^{N} \hat{V}_i(p) \right)^2 \tag{8}$$

ir konsistents $(m+n)\mathrm{Var}(\hat{R}_{m,n}(p))$ novērtējums.

Teorēma 1. Pieņemsim, ka w ir blīvuma funkcija $[-1, 1]$ un pirmais w atvasinājums ir ierobežots un nepārtraukts. Otrais $R(p)$ atvasinājums ir nepārtraukts un ierobežots $p \in (0, 1)$ un $\lim_{n\to\infty} m/n = r \in (0, \infty)$. Ja $h = h(n) \to 0$, tad $nh^2/\log n \to \infty$, $nh^4 \to 0$ un $n \to \infty$, tad

$$v_{m,n}(p) \xrightarrow{p} \sigma^2(p), \text{ja } n \to \infty$$

Teorēma 2. No Teorēmas 1. pieņemumiem izriet, ka

$$l(\theta, p) \xrightarrow{d} \chi_1^2), \text{ja} n \to \infty$$

Asimptotisks $100(1-\alpha)\%$ nogludināts džeknaifa empīriskas paticamības ticamības intervāls θ:

$$I(p) = (\theta : l(\theta, p) \leq \chi_1^2(\alpha)),$$

kur $\chi_1^2(\alpha)$ ir χ_1^2 $\alpha-$ kvantīle.

2. ROC līkņu salīdzināšana

Ir trīs galvenie paņēmieni ROC līkņu salīdzināšanai:

1. Noteikt, vai divās ROC līknes ir identiskās, t.i. noteikt, vai katrs TPR punkts ir vienāds katram FPR punktam;

2. Noteikt, vai divās ROC līknes ir līdzīgās noteiktām FPR;

3. Noteikt, vai ROC līkņu rādītāji ir vienādi. Piemēram, laukumi zem ROC līknem (AUC).

Jāātzīmē, ka 1. un 3. paņēmiens nepārbaudā vienu un to pašu hipotēzi. Protams, ja divas ROC līknes ir vienādās, tad arī laukumi zem līknem ir vienādi. Bet pretējais apgalvojums nav patiess: var būt gadījumi, kad ROC līknem ir dažādas formas, bet laukumi zem līknem ir vienādi.

2.1. Parametriskās metodes

Apskatīsim pirmo paņēmienu. Pieņemsim, ka dati ir normāli sadalīti. Lai noskaidotu, vai ROC līknes ir vienādas, jānoskaidro, vai ir vienādi ROC līkņu parametri. Tātad, tiks apskatītas šādas hipotēzes

$$H_0 : a_1 = a_2 \cup b_1 = b_2$$

pret hipotēzi

$$H_1 : a_1 \neq a_2 \cup b_1 \neq b_2$$

Hipotēžu testēšanai tiks izmantotā šāda statistika

$$X^2 = \frac{\hat{a}_{12}^2 \mathrm{Var}(\hat{b}_{12}^2) + \hat{b}_{12}^2 \mathrm{Var}(\hat{a}_{12}^2) - 2\hat{a}_{12}^2 \hat{b}_{12}^2 \mathrm{Cov}(\hat{a}_{12}^2, \hat{b}_{12}^2)}{\mathrm{Var}(\hat{a}_{12}^2)\mathrm{Var}(\hat{b}_{12}^2) - \mathrm{Cov}(\hat{a}_{12}^2, \hat{b}_{12}^2)},$$

kur $a_{12} = a_1 - a_2$ un $b_{12} = b_1 - b_2$. Pie nulles hipotēzes par binormālo parametru ekvivalenci testa statistikai ir Hī-kvadrāta sadalījums ar divām brivības pakāpem, t.i. χ_2^2.

Ja dati nav normāli sadalīti, tad vispirms tiek pielietota Boksa-Koksa transformācija. Zou (2001.) pieņēma, ka pēc transformācijas divu diagnostikas testu rezultātu sadalījumam ir dažādas vidējās vērtības, bet vienādas dispersijas. Tātad $b_1 = b_2$, un lai pārbaudīt divu ROC līkņu ekvivalenci, jāparbauda vai $a_1 = a_2$. To var izadrīt, izmantojot statistiku

$$Z = \frac{\hat{a}_1 - \hat{a}_2}{\sqrt{\mathrm{Var}(\hat{a}_1) + \mathrm{Var}(\hat{a}_2) - 2\mathrm{Cov}(\hat{a}_1, \hat{a}_2)}},$$

kurai ir asimptotiskais un normālais sadalījums.

Cita metode, lai pārbaudīt, vai divas ROC līknes ir vienādas, ir salīdzinot līknes vienā noteiktā punktā uz visas līknēs. Apzīmēsim ar D pazīmi, vai subjekts ir slims vai nav. Tātad D pieņem vērtības 0 vai 1. Pie pieņēmuma par binormālitāti, salīdzinot

divas ROC līknes noteiktā $FPR = e$, ir $D(Z_e) = b_{12}Z_e - a_{12}$. Novērtējumu $\hat{D}(Z_e)$ var aprēķināt, izmantojot maksimālās ticamības metodes novērtējumus (MLE) $\hat{a}_i$ un $\hat{b}_i$, $i = 1, 2$. Hipotēžu $H_0 : D(Z_e) = 0$ pret $H_1 : D(Z_e) \neq 0$ pārbaudei var izmantot Valda (Wald) statistiku

$$Z = \hat{D}(Z_e)/\sqrt{\hat{\mathrm{Var}}(\hat{D}(Z_e))},$$

balstoties uz $\hat{D}(Z_e)$ asimptotisko normalitāti.

Divas jūtīguma vērtības s_n salīdzinājums pie nemainīgās specifiskums s_p var arī neparametrsiki. Tad testa statistika būs

$$Z = \frac{\hat{SE}_1 - \hat{SE}_2}{\sqrt{\hat{\mathrm{Var}}(\hat{SE}_1 - \hat{SE}_2)}},$$

kurai ir asimptotiski normālais sadalījums.

2.2. AUC salīdzināšana

Laukumus zem ROC līknem AUC_1 un AUC_2 salīdzināšanai, pie pieņēmuma par binormalitāti, izmanto Valda statistiku

$$Z = \frac{\hat{AUC}_1 - \hat{AUC}_2}{\sqrt{\hat{\mathrm{Var}}(\hat{AUC}_1 - \hat{AUC}_2)}}.$$

$\hat{AUC}_i$, $i = 1, 2$ var novērtēt, izmantojot formulu

$$AUC = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right)$$

un

$$AUC(e_1 \leq FPR \leq e_2) = \int_{c_1}^{c_2} \Phi(bv - a)\phi(v)dv,$$

kur $c_i = \Phi^{-1}(e_i)$ un $\phi(v)$ ir standarta normālā sadalījuma blīvuma funkcija. Dispersija $\mathrm{Var}(\hat{AUC}_1 - \hat{AUC}_2)$ ir dispersiju summa, ja dati nav korelēti. Pretējā gadījumā, jānovērtē divu laukumu novērtējumu kovariācija. ROC līknes būs korelētas tad, ja dažādi diagnostiskie testi tika veikti uz vienas un tās pašas subjektu izlases.

Līdzīgi var izmantot Z statistiku, lai iegūt AUC_1 un AUC_2 un $\mathrm{Var}(\hat{AUC}_1 - \hat{AUC}_2)$ novērtējumus. Ja dati nav korelēti, tad $\mathrm{Var}(\hat{AUC}_1 - \hat{AUC}_2)$ ir vienkārši atsevišķu laukumu dispersiju summa ([14]).

Elizabeth R. DeLong (1988.) [2] piedāvāja neparametrisku metodi divu AUC salīdzināšanai. Tika pierādīts, ka empīriskā AUC, kura ir aprēķināta ar trapecijas likumu, ir ekvivalenta Mana-Vitnēja (Mann-Whitney) statistikai.

Pieņemsim, ka izlasei no N indivīdiem tika veikti diagnostiskie testi, lai noteikt kādu interesējošo notikumu esamību. Pieņemsim, ka testa lielākās vērtības būs saistītas ar interesējošo notikumu, t.i. piemēram pozitīvo slimības statusu, tātad šiem subjektiem

$D = 1$. Pieņemsim, ka šādus subjektus ir m un $n = N - m$ indivīdiem nepiemīt šīs notikums, t.i. $D = 0$. Mana Vitnēja statistika

$$\hat{\theta} = \frac{1}{mn} \sum_{j=1}^{n} \sum_{i=1}^{m} \psi(X_i, Y_i),$$

kur ψ ir kodolu vidējais

$$\psi(X, Y) = \begin{cases} 1, & \text{ja } Y < X \\ 1/2 & \text{ja } Y = X \\ 0, & \text{ja } Y > X \end{cases}$$

Ņemot vērā, ka $E(\hat{\theta}) = \theta = P(Y < X) + 1/2P(X = Y)$. Nepārtrauktam sadalījumam $P(Y = X) = 0$. Definēsim

$$\begin{aligned} \xi_{10} &= E(\psi(X_i, Y_j)\psi(X_i, Y_k)) - \theta^2, \ j \neq k \\ \xi_{01} &= E(\psi(X_i, Y_j)\psi(X_k, Y_j)) - \theta^2, \ i \neq k \\ \xi_{11} &= E(\psi(X_i, Y_j)\psi(X_i, Y_j)) - \theta^2. \end{aligned} \tag{1}$$

Tad

$$\text{Var}(\hat{\theta}) = \frac{(n-1)\xi_{10} + (m-1)\xi_{01}}{mn} + \frac{\xi_{11}}{mn} \tag{2}$$

Pieņemsim, ka $\hat{\theta} = (\hat{\theta}^1, \hat{\theta}^2, ..., \hat{\theta}^k)$ ir statistiku vektors , kas reprezentē laukumus zem ROC līknem no $\{X_i^r\}$ un $\{Y_i^r\}$, $(i = 1, ..., m; j = 1, ..., n; 1 \leq r \leq k)$ katram no k dažādiem diagnostikas testiem. Tad, līdzīgi kā (1) definēsim

$$\begin{aligned} \xi_{10}^{rs} &= E(\psi(X_i^r, Y_j^r)\psi(X_i^s, Y_k^s)) - \theta^r\theta^s, \ j \neq k \\ \xi_{01}^{rs} &= E(\psi(X_i^r, Y_j^r)\psi(X_k^s, Y_j^s)) - \theta^r\theta^s, \ i \neq k \\ \xi_{11}^{rs} &= E(\psi(X_i^r, Y_j^r)\psi(X_i^s, Y_j^s)) - \theta^r\theta^s. \end{aligned} \tag{3}$$

r-tās un s-tās statistikas kovariācija ir

$$\text{Cov}(\hat{\theta}^r, \hat{\theta}^s) = \frac{(n-1)\xi_{10}^{rs} + (m-1)\xi_{01}^{rs}}{mn} + \frac{\xi_{11}^{rs}}{mn} \tag{4}$$

$$V_{10}^r(X_i) = \frac{1}{n} \sum_{j=1}^{n} \psi(X_i^r, Y_j^r), (i = 1, 2, ..., m)$$

un

$$V_{01}^r(Y_j) = \frac{1}{m} \sum_{i=1}^{m} \psi(X_i^r, Y_j^r), (j = 1, 2, ..., n).$$

Definēsim $k \times k$ matricu S_{01} tādu, ka (r, s)-tais elements ir

$$s_{10}^{r,s} = \frac{1}{m-1} \sum_{i=1}^{m} \left[V_{10}^r(X_i)\hat{\theta}^r \right] \left[V_{10}^s(X_i)\hat{\theta}^s \right],$$

un līdzīgi arī matricu S_{10} tādu, ka (r,s)-tais elements ir

$$s_{01}^{r,s} = \frac{1}{n-1}\sum_{j=1}^{n}\left[V_{01}^{r}(Y_j)\hat{\theta}^r\right]\left[V_{01}^{s}(Y_j)\hat{\theta}^s\right].$$

Novērtēta parametru novērtējumu kovariāciju matrica $\hat{\theta} = (\hat{\theta}^1, \hat{\theta}^2, ..., \hat{\theta}^k)$ ir

$$S = \frac{1}{m}S_{10} + \frac{1}{n}S_{01}$$

Gadījumā, ja $\lim_{N\to\infty}\frac{m}{n}$ ir ierobežots un nav vienāds ar nulli, tad $N^{1/2}\left[g(\hat{\theta}) - g(\theta)\right]$ ir asimptotiski normāli sadalīts ar viedējo vērtību 0 un dispersiju σ_g^2, kur

$$\sigma_g^2 = \lim_{N\to\infty} N\sum_{j=1}^{k}\sum_{i=1}^{k}\frac{\delta g}{\delta\theta^i}\frac{\delta g}{\delta\theta^j}\left(\frac{1}{m}\xi_{10}^{i,j} + \frac{1}{n}\xi_{01}^{i,j}\right).$$

σ_g^2 novērtējums ir

$$s_g^2 = N\sum_{j=1}^{k}\sum_{i=1}^{k}\frac{\delta g}{\delta\theta^i}\frac{\delta g}{\delta\theta^j}\left(\frac{1}{m}s_{10}^{i,j} + \frac{1}{n}s_{01}^{i,j}\right).$$

2.3. Nogludinātā JEL metode ROC līkņu starpībai

Yangs un Zhao [13] piedavāja jaunu metodi ROC līkņu starpības novērtēšanai un ticamības intervālu konstruēšanai.

Pieņemsim, ka $X = (X_1, X_2)$ un $Y = (Y_1, Y_2)$ ir divdimensiju gadījuma lielumi. X un Y ir neatkarīgi. $F(x_1, x_2)$ un $G(y_1, y_2)$ ir attiecīgi slimo un neslimo populācijas sadalījuma funkcijas. Apzīmēsim marginālās sadalījuma funkcijas ar $F_1(x_1)$, $F_2(x_2)$, $G_1(y_1)$ un $G_2(y_2)$. Pieņemsim, ka tiek veikts diagnostiskais test uz X_1 un Y_1. Tad nepārtraukta ROC līkne būs $R_1(p) = 1 - F_1(G_1^{-1}(1-p))$, kur $0 < p < 1$ un G_1^{-1} ir Y_1 kvantiļu funkcija. Līdzīgi definējam otro nepārtrauktu ROC līkni $R_2(p) = 1 - F_2(G_2^{-1}(1-p))$, kur G_2^{-1} ir Y_2 kvantiļu funkcija. Tātad divu korelētu ROC līkņu starpība noteiktā punktā p ir

$$\Delta(p) = F_1(G_1^{-1}(1-p)) - F_2(G_2^{-1}(1-p))$$

Pieņemsim, ka divdimensiju dati $X = (X_{1i}, X_{2i}),\ \ i = 1,\ ...\ , m$ ir saistīti ar slimo subjektu populāciju un $Y = (Y_{1j}, Y_{2j}),\ \ j = 1,\ ...\ , n$ ir saistīta ar neslimo subjektu populāciju. (X_{1i}, X_{2i}) un (Y_{1j}, Y_{2j}) ir neatkarīgi un vienādi sadalīti. Apzīmēsim ar $F_m(x_1, x_2) = 1/m\sum_{j=1}^{m} I_{X_{1,j}\le x_1, X_{2,j}\le x_2}$ un $G_n(y_1, y_2) = 1/n\sum_{j=1}^{n} I_{Y_{1,j}\le y_1, Y_{2,j}\le y_2}$ divdimensiju sadalījumu funkciju empīrisko novērtējumus. Marginālie sadalījumu empīriskie novērtējumi ir $F_{m,1}(x_1) = 1/m\sum_{i=1}^{m} I_{X_{1,i}\le x_1}$, $F_{m,2}(x_2) = 1/m\sum_{i=1}^{m} I_{X_{2,i}\le x_2}$, $G_{n,1}(y_1) = 1/n\sum_{i=1}^{n} I_{Y_{1,i}\le y_1}$, $G_{n,2}(y_2) = 1/n\sum_{i=1}^{n} I_{Y_{2,i}\le y_2}$.

ROC līkņu nogludinātie novērtējumi ir

$$\hat{R}_{m,n,1}(p) = 1 - \frac{1}{m}\sum_{j=1}^{m} K\left(\frac{1 - p - G_{n,1}(x_{1,j})}{h}\right)$$

$$\hat{R}_{m,n,2}(p) = 1 - \frac{1}{m}\sum_{j=1}^{m} K\left(\frac{1-p-G_{n,2}(x_{2,j})}{h}\right)$$

kur $h = h(n) > 0$ ir joslas platums un $K(p) = \int_{u \le p} w(u)du$ ir nogludinātā sadalījuma funkcija, $w(u)$ ir simetriskā blīvuma funkcija $[-1, 1]$. un ROC līkņu starpību apzīmēsim ar Δ. Tad ROC līkņu starpības npvērtējums ir

$$\hat{\Delta}_{m,n}(p) = \hat{R}_{m,n,1}(p) - \hat{R}_{m,n,2}(p),$$

Lai ģenerēt pseido izlasi, apskatīsim sekojošu procedūru: ja $1 \le i \le m$, tad divu līkņu starpība būs

$$\hat{\Delta}_{m,n,i}(p) = \frac{1}{m-1}\sum_{j=1,\, j\neq i}^{m}\left[K\left(\frac{1-p-G_{n,m,2}(x_{1,j})}{h}\right) - K\left(\frac{1-p-G_{n,m,1}(x_{2,j})}{h}\right)\right],$$

bet ja $m+1 \le i \le m+n$, tad līkņu strapība būs

$$\hat{\Delta}_{m,n,i}(p) = \frac{1}{m-1}\sum_{j=1}^{m}\left[K\left(\frac{1-p-G_{n,m-i,2}(x_{1,j})}{h}\right) - K\left(\frac{1-p-G_{n,m-i,1}(x_{2,j})}{h}\right)\right]$$

kur

$$G_{n,-i,k}(y) = \frac{1}{n}\sum_{j=1,\, j\neq i}^{n} I_{(y_{k,j} \le y)},\ i = 1, \ldots, m+n$$

Džeknaifa pseido-izlase ir definēta šādi

$$\hat{V}_i(p) = (m+n)\hat{\Delta}_{m,n}(p) - (m+n-1)\hat{\Delta}_{m,n,i}(p),\ i = 1, \ldots, m+n. \tag{5}$$

Apzīmēsim ar N visu novērojumu skaitu, t.i. $N = n + m$, tad empīriskās ticamības funkcijas log-attiecība punktā p un starpības vērtības $\tilde{\Delta}$, kura ir balstīta uz pseido-izlasi ir

$$L(\tilde{\Delta}, p) = \frac{\sup\left\{\prod_{i=1}^{N} p_i : \sum_{i=1}^{N} p_i = 1, \sum_{i=1}^{N} p_i \hat{V}_i(p) = \tilde{\Delta},\ p_i > 0,\ i = 1, \ldots, N\right\}}{\sup\left\{\prod_{i=1}^{N} p_i,\ \sum_{i=1}^{N} p_i = 1,\ p_i > 0,\ i = 1, \ldots, N\right\}}.$$

Izmantojot Lagranža reizinātāju metodi, tiek iegūts

$$l(\tilde{\Delta}, p) = -2\log L(\tilde{\Delta}, p) = 2\sum_{i=1}^{N}\log(1 + \lambda(\hat{V}_i(p - \tilde{\Delta})), \tag{6}$$

kur Lagranža reizinātājs λ apmierina izteiksmi

$$\sum_{i=1}^{N}\frac{\hat{V}_i(p) - \tilde{\Delta}}{1 + \lambda(\hat{V}_i(p - \tilde{\Delta})} = 0.$$

Pseido- izlases dispersija ir

$$v_{m,n}(p) = \frac{1}{m+n}\sum_{i=1}^{N}\left(\hat{V}_i(p) - \frac{1}{m+n}\sum_{i=1}^{N}\hat{V}_i(p)\right)^2 \tag{7}$$

Apskatīsim dažus pieņēmumus:

1. $F_1(x_1)$, $F_2(x_2)$, $G_1(y_1)$, $G_2(y_2)$, $F(x_1, x_2)$ un $G(y_1, y_2)$ ir nepārtrauktās funkcijas un tām ir nepārtraukts un ierobežots pirmais atvasinājums;

2. ROC līknes $R_1(p)$ un $R_2(p)$ un pirmie atvasinājumi $R_1^{'}(p)$ un $R_2^{'}(p)$ ir nepārtraukti un ierobežoti $p \in (0, 1)$;

3. $w(u)$ ir simetriskā blīvuma fnkcija intervālā $[-1, 1]$ un $w^{'}(u)$ ir ierobežota, nepārtraukta, ja $u \in [-1, 1]$;

4. $h = h(n) \to 0$, $nh^2/\log n \to \infty$, $nh^4 \to 0$, ja $n \to \infty$;

5. $p \in (a, b)$ $\forall (a, b) \subset (0, 1)$;

6. $m/n \to r$, $r > 0$.

Teorēma 3. Ņēmot vērā pieņēmumus 1. - 6., $p \in (a, b)$, pseido izlases dispersijai ir spēkā asimptotiskā īpašība:

$$v_{m,n}(p) \xrightarrow{p} \sigma^2(p),$$

kur

$$\sigma^2(p) = \sigma_1^2(p) + \sigma_2^2(p) + 2\sigma_{12}^2(p),$$

$$\sigma_1^2(p) = \frac{1+r}{r} R_i(p)(1 - R_i(p)) + (1+r)(1-p)p(R_i^{'}(p))^2, i = 1, 2$$

$$\sigma_{12}^2(p) = \frac{1+r}{r}(F(G_1^{-1}(p), G_2^{-1}(p)) - (1 - R_1(p))(1 - R_2(p))) + (1+r)(G(G_1^{-1}(p), G_2^{-1}(p)) - p^2)R_1^{'}(p)R_2^{'}(p).$$

Teorēma 4.

$$l(\Delta(p), p) \xrightarrow{d} \chi_1^2),$$

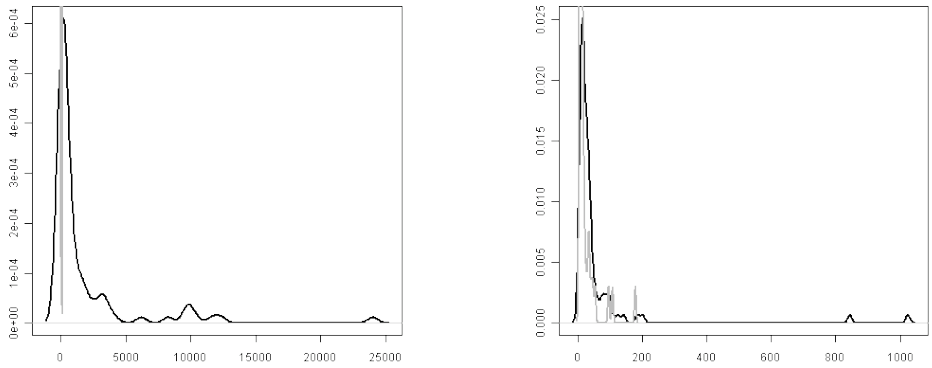
kur $\Delta(p)$ ir divas savstarpēji saistĪtas ROC līknes starpības reāla vertība fiksētā punktā $p \in (a, b)$. Asimptotisks $100(1-\alpha)\%$ nogludināts Džeknaifa empīriskas paticamības ticamības intervāls $\Delta(p)$:

$$I(p) = (\tilde{\Delta} : l(\tilde{\Delta}, p) \leq \chi_1^2(\alpha)),$$

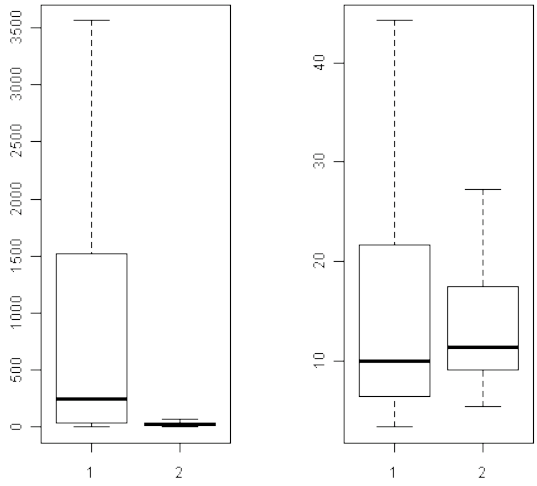
kur $\chi_1^2(\alpha)$ ir χ_1^2 $\alpha-$ kvantīle.

3. Praktiskais piemērs

Lai pārbaudītu kā strādā teorija praksē, tika izmantoti dati "aizkuņģa dziedzera vēža seruma biomarkeru dati" (1989) [12] ("the Pancreatic Cancer Serum Biomarkers"). Dati satur divu biomarkeru: vēžā antīgēnu $CA-125$ (cancer antigen) un karbohidrāta antīgēnu $CA-19-9$ (carbohydrate antigen) un bināro mainīgo, kura vērtības ir 1 (pacienti, kuriem ir aizkuņģa vēzis) vai 0 (pacienti, kuriem ir pankreatīts). Mērījumi tika veikti, apsekojot 90 pacientus ar aizkuņģa dziedzera vēzi un 51 pacientus kam nebija vēzis, bet pankreatīts. 5. attēlā ir atspoguļotās attiecīgās sadalījuma blīvuma funkcijas.



5. att.: Pirmā un otrā biomarkera sadalījuma blīvuma funkcija slimiem un neslimiem subjektiem



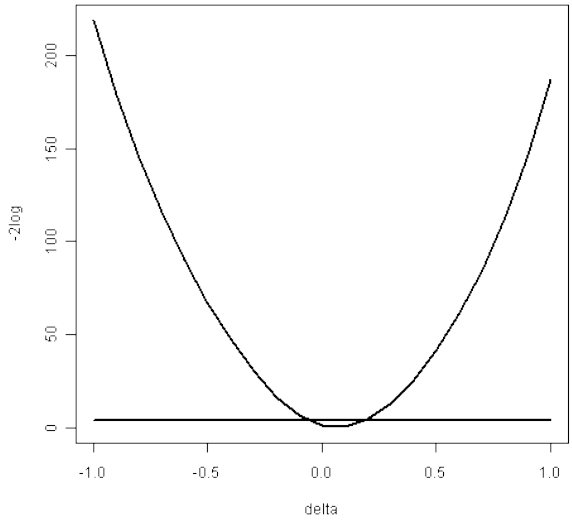
6. att.: Slimo (X_1 un X_2) un neslimo (Y_1 un Y_2) populācijas biomarkeru kastu grafiki, attiecīgi

6. attēlā ir attiecīgi slimo un neslimo populācijas biomarkeru kastu grafiki (*boxplot*) bez izlēcējiem. Var ievērot, ka otra biomarkera vērtības ir daudz mazākās nekā pirmā biomarkera vērtības slimo subjektu izlasei, neslimo subjektu izlasei tik lielas atšķirības starp pirma un otra biomarkera vērtībam nav.

Apzīmēsim ar $R_1(p)$ pirmā biomarkera ROC līkni un ar $R_2(p)$ otrā biomarkera ROC līkni. Pieņemsim, ka AUC_1 ir laukums zem $R_1(p)$ un AUC_2 ir ir laukums zem $R_2(p)$. Novērtējot laukumus zem ROC līknem, iegūvam rezultātu, ka $\hat{AUC}_1 = 0.8614$ un $\hat{AUC}_2 = 0.7056$. Izmantojot DeLonga testu, iegūvam, ka statistika $Z = 2.7221$, un $p - value = 0.006488$. Tātad nulles hipotēze, ka AUC vērtība ir 0 tiks noraidīta.

Varam secināt, ka testam ar pirmo biomarkeru ir labākās klasificēšanas spējas. Apskatīsim, kādu rezultātu dos džeknaifa empīriskas ticamības metode ROC līkņu starpībai.

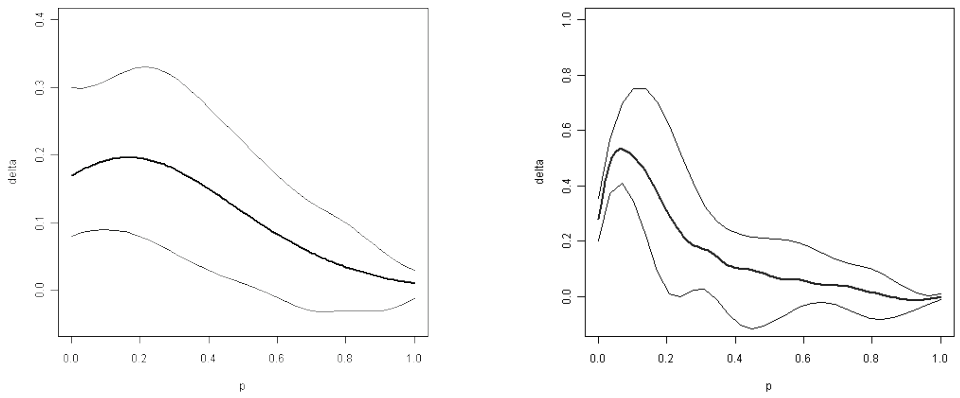
Lai novērtētu ROC līknes un ROC līkņu starpību tika izvēlēts Gausa kodols un joslas platums, kas ir vienāds ar $n^{-1/3} = 0.2697$. Tālāk tiek veidota džeknaifa pseido-izlase. Un ar empīriskās ticamības metodi tiks konstruēti ticamības joslas. Katrai p vērtībai tiek konstruēta $-2\log$ funkcija.



7. att.: $l(\tilde{\Delta}, p)$ ticamības līkne pie noteiktā $p = 0.5$, $\alpha = 0.05$

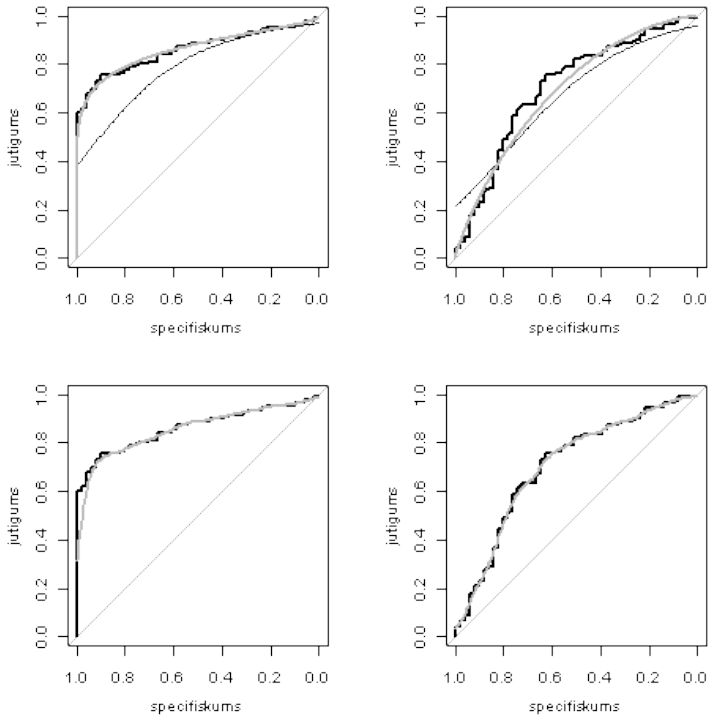
Ticamības intervāli tiek konstruēti, ņemot vērā, ka $I(p) = \{\tilde{\Delta} : l(\tilde{\Delta}, p) \leq \chi^2_1(\alpha)\}$. Tātad pie nozīmības līmeņa 95% konstruējam ticamības intervālus ROC līkņu starpībai.

8. attēlā ir redzamās ticamības intervāli ROC līkņu starpībai. Varam redzēt, ka nevienā punktā funkcija neiziet aiz ticamības josliem un 95% ticamības intervāls atrodas augstāk par 0 katrā $(1 - s_p)$ punktā no 0 līdz aptuvēni 0.7. Tātad nevaram noraidīt hipotēzi, ka pirmajam biomarkeram, $CA - 125$ patiešām ir labākās spējas atšķirt subjektus, kuriem ir pankreātiskais vēzis un subjektus, kuriem ir pankreatīts.



8. att.: Ticamības intervāli ROC līkņu starpībai, konstruēti ar nogludināto JEL metodi. Joslas platums $h = 0.2697$ (pa labi) un joslas platums $h = 0.03$ (pa kreisi)

Kā tika pieminēts teorētiskā daļa, kodolu gludināšanās metodē ir ļoti svarīgi pareizi izvēlēties joslas platumu. Uzkonstruējot attiecīgus grafikus, varam ievērot, ka mūsu apskatītajā gadījumā ROC līkne izskatās pārgludināta, tāpēc būtu lietderīgi apskatīt arī citu joslas platumu.



9. att.: Nogludinātās ROC līknes pirmajam biomarkerim (pa kreisi) un otrajam biomarkerim (pa labi)

9. attēlā ir redzamās pirmā (pa kreisi) un otrā biomarkeru ROC līknes. Augšējos attēlos ar tievu melno līniju ir attēlota nogludinātā ROC līkne ar jau apskatīto joslas

platumu $h = n^{-1/3} = 0.2697$ un ar pelēku ir ar R [16] iebūvētas paketes *pROC* [15] palīdzību iegūts nogludināts novērtējums. Apakšā ir grafiki, ar "uz aci" izvēlēto joslas platumu $h = 0.03$.

Līdzīgi kā iepriekšējā gadījumā konstruēsim ticamības intervālus ROC līkņu starpībai. Rezultāts ir līdzīgs pirmajām gadījumām un ir apskatāms 8. attēlā. ROC līkņu starpībai ir mazliet lielākās vērtības nekā iepriekšējā gadījumā.

4. Nobeigums

Maģistra darbā tika apskatīta teorija par ROC līkņu galvenam raksturojošiem lielumiem, ROC līkņu novērtēšanu un salīzdināšanu. Tika apskatītas un salīdzinātās neparametriskās un parametriskās metodes.

Darba mērķis bija apskatīt jauno nogludinātās džeknaifa empīriskās ticamības funkcijas metodi, ar kuru palīdzību ir iespējams konstruēt ticamības intervālus ROC līknem un ROC līkņu starpībai. Teorija tika pārbaudīta praksē ar programmā R uzrakstītu algoritmu palīdzību. Tika secināts, ka iegūtie rezultāti nav pretrunā ar citam metodem. Ticamības joslas dod iespēju veikt grafiski hipotēžu pārbaudi par divu ROC līkņu vienādību.

Turpmākais solis varētu būt apskatīt un izpētīt teoriju par klasifikācijas uzdevumiem ar trim kategorijam, ROC virsmas un tilpumu zem ROC virsmam (VUS); apskatīt, vai ir iespējams pielietot metodes, kuras tiek lietotas ROC līkņu novērtēšanai un salīdzināšanai arī ROC virsmas gadījumā; un meklēt jaunas metodes ROC līkņu un ROC virsmu pētīšanai.

Literatūra

[1] **Claeskens G., Jing B., Peng L., Zhou W.** *Empirical likelihood confidence regions for comparison distributions and ROC curves*, The Canadian Journal of Statistics 31 (2) (2003) 173-190.

[2] **DeLong E., DeLong D., Daniel L., Clarke-Pearson D.L.** *Comparing the areas under two or more correlated Receiver Operating Charasteristic curves: a nonparametric approach*, Biometrics 44(1988) 837-845.

[3] **Dumberg L., Rufibach K.** *loncondens: Computational Related to Unvariate LogConcave Density Estimation*, REVSTAT 39(2011).

[4] **Fabsic P.** *Comparing the accuracy of ROC curve estimation methods*, (2012). Mathematics Theses.

[5] **Gong Y., Peng L., Qi Y.** *Smoothed jackknife empirical likelihood for ROC curve*, Journal of Multivariate Analysis 101(2010) 1520-1531.

[6] **Gonzcalves L., Subtil A., Rosario Oliveira M., Bermudez P. Z.** *ROC Curve Estimation: An Overview*, REVSTAT 12(2014) 1-20.

[7] **Horova I., Forbelska M., Zelinka J.** *Comparative Study of the Estimation of the Area Under the ROC curves*, Masaryk University. (2004).

[8] **Jing B., Yuan J., Zhou W.** *Jackknife empirical likelihood*, Journal of the American Statistical Association 104(2009) 1224-1232.

[9] **Kumar R., Indrayan A.** *Receiver Operating Characteristic (ROC) Curve for Medical Researchers*, INDIAN PEDIATRICS. 48(2011)

[10] **Lloyd C. J.** *Using Smoothed Receiver Operating Characteristic Curves to Summarize and Compare Diagnostic Systems*, Journal of the American Statistical Association 93(444)(1998), 1356-1364.

[11] **Rufibach K.** *A smooth ROC curve estimator based on log-concave density estimates*, Biostatistics (2011).

[12] **Wieand S., Gail M., James B.** *A family of nonparametric statistics diagnostic markers with paired or unpaired data*, Biometrika 76(1989) 585-592.

[13] **Yang H., Zhao Y.** *Smoothed jackknife empirical likelihood inference for the difference of ROC curves*, Journal of Multivariate Analysis 115(2013) 270-284.

[14] **Zhang D.** *Statistical Inferences of Comparison between two Correlated ROC Curves with Empirical Likelihood Approaches*, (2012). A Dissertation.

[15] **Robin X., Turck N., Hainard A., Tiberti N.** *pROC: display and analyze ROC curves*, 2014, http://cran.r-project.org/web/packages/pROC/index.html

[16] **R Development Core Team** *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. 2009. http://www.R-project.org.

Pielikumi

1. Pielikums. R kods

```
#jackknife empirical likelihood confidence interval
emp.fun <- function(y.data, x.data){
fn <- ecdf(y.data)
fn(x.data)}
emp.fun <- Vectorize(emp.fun, "x.data")

Rmn <- function(y.data, x.data, p, h){
1 - .Internal(mean(pnorm((1 - p - (emp.fun(y.data, x.data)))/h)))}
Rmn <- Vectorize(Rmn, vectorize.args="p")

Delta.mni2 <- function(y.dati1, y.dati2, x.dati1, x.dati2, p, h, i){
Rmn(y.dati1[-i], x.dati1, p, h) - Rmn(y.dati2[-i], x.dati2, p, h)}
Delta.mni2 <- Vectorize(Delta.mni2, vectorize.args="i")

V.hat <- function(y.dati1, y.dati2, x.dati1, x.dati2, p, h, i1, i2){
m <- length(x.dati1);n <- length(y.dati1)
V.hat1 <- (m + n)*Delta.mn(y.dati1, y.dati2, x.dati1, x.dati2, p, h) -
   (m + n - 1)*Delta.mni1(y.dati1, y.dati2, x.dati1, x.dati2, p, h, i1)
V.hat2 <- (m + n)*Delta.mn(y.dati1, y.dati2, x.dati1, x.dati2, p, h) -
   (m + n - 1)*Delta.mni2(y.dati1, y.dati2, x.dati1, x.dati2, p, h, i2)
if (length(p)==1)  V.hat <- c(V.hat1,V.hat2)
   else    V.hat <-cbind(V.hat1, V.hat2)
V.hat}

lambda.fun <- function(lambda, delta, y.dati1, y.dati2, x.dati1,
 x.dati2, p, h, i1, i2)
{sum((V.hat(y.dati1, y.dati2, x.dati1, x.dati2, p, h, i1, i2) - delta)/(1 +
      lambda*(V.hat(y.dati1, y.dati2, x.dati1, x.dati2, p, h, i1, i2) - delta)))}
lambda.fun <- Vectorize(lambda.fun, vectorize.args="lambda")

log.fun <- function(delta, y.dati1, y.dati2, x.dati1, x.dati2, p, h, i1, i2)
{lb <- (1 - 1/length(V.hat(y.dati1, y.dati2, x.dati1, x.dati2, p, h, i1, i2)))/
(delta - max(V.hat(y.dati1, y.dati2, x.dati1, x.dati2, p, h, i1, i2)))
ub <- (1 - 1/length(V.hat(y.dati1, y.dati2, x.dati1, x.dati2, p, h, i1, i2)))/
(delta - min(V.hat(y.dati1, y.dati2, x.dati1, x.dati2, p, h, i1, i2)))
lambda <- uniroot(function(lambda) lambda.fun(lambda, delta, y.dati1,
 y.dati2, x.dati1,x.dati2, p, h, i1, i2), interval=c(lb,ub),tol=0.001)$\$$root
log.fun <- 2*sum(log(1 + lambda*(V.hat(y.dati1, y.dati2, x.dati1, x.dati2,
 p, h, i1, i2) - delta))); log.fun}
log.fun <- Vectorize(log.fun, vectorize.args=c("delta"))

lower.b <- function(delta, y.dati1, y.dati2, x.dati1, x.dati2, p, h, i1, i2, Q)
```

```
{uniroot(log.fun1, interval =
c(min(V.hat(y.dati1, y.dati2, x.dati1, x.dati2, p, h, i1, i2))+0.1,
optimize(log.fun(delta, y.dati1, y.dati2, x.dati1, x.dati2, p, h, i1, i2),
c(min(V.hat(y.dati1, y.dati2, x.dati1, x.dati2, p, h, i1, i2)),
max(V.hat(y.dati1, y.dati2, x.dati1, x.dati2, p, h, i1, i2))),
delta=delta, y.dati1 = y.dati1, y.dati2=y.dati2, x.dati1=x.dati1,
x.dati2=x.dati2, p=p, h=h, i1=i1, i2=i2, Q=Q)$\$$minimum))$\$$root}

upper.b <- function(delta, y.dti1, y.dati2, x.dati1, x.dati2, p, h, i1, i2, Q)
{uniroot(log.fun1, interval= c(optimize(log.fun1(delta, y.dati1, y.dati2,
 x.dati1, x.dati2, p, h, i1, i2),c(min(V.hat(y.dati1, y.dati2, x.dati1,
  x.dati2, p, h, i1, i2)),max(V.hat(y.dati1, y.dati2, x.dati1, x.dati2,
   p, h, i1, i2))))$\$$minimum,
max(V.hat(y.dati1, y.dati2, x.dati1, x.dati2, p, h, i1, i2))-0.1),
delta=delta, y.dati1 = y.dati1, y.dati2=y.dati2, x.dati1=x.dati1,
 x.dati2=x.dati2,p=p, h=h, i1=i1, i2=i2, Q=Q)$\$$root}

#splaini
lines(spline(y, x2))
lines(spline(y, x1))

#pROC
library(pROC)
roc1 <- roc(data[,"d"], data[,"A"])
roc2 <- roc(data[,"d"], data[,"B"])
par(mfrow=c(2,2))
plot(roc1, xlab="specifiskums, ylab="jutigums")
lines(1-pp, Rmn(Y1,X1,pp, h))
lines(smooth.roc(roc1), col="grey")
plot(roc2, xlab="specifiskums", ylab="jutigums")
lines(1-pp, Rmn(Y2,X2,pp, h))
lines(smooth.roc(roc2), col="grey")
plot(roc1, xlab="specifiskums", ylab="jutigums")
lines(1-pp, Rmn(Y1,X1,pp, 0.03))
plot(roc2, xlab="specifiskums", ylab="jutigums")
lines(1-pp, Rmn(Y2,X2,pp, 0.03))
coor1 <- coords(roc1, "best", ret=c("threshold",
 "specificity", "sensitivity", "accuracy",
"tn", "tp", "fn", "fp", "npv", "ppv", "1-specificity",
"1-sensitivity", "1-accuracy", "1-npv", "1-ppv"))
coor2 <- coords(roc2, "best", ret=c("threshold",
 "specificity", "sensitivity", "accuracy",
"tn", "tp", "fn", "fp", "npv", "ppv", "1-specificity",
"1-sensitivity", "1-accuracy", "1-npv", "1-ppv"))
auc(roc1)
auc(roc2)
roc.test(roc1, roc2, method="bootstrap")
```

```
roc.test(roc1, roc2, method="delong")
roc.test(roc1, roc2, alternative="less")
roc.test(roc1, roc2, method="specificity", specificity=0.8)

plot(roc1, print.thres="best", print.thres.best.method="closest.topleft",
col="black")
plot(roc1, print.thres="best", print.thres.best.method="youden",add=T)
lines(c(1,0.902), c(1,0.756))
lines(c(0.902,0.902),c(0.756,0.1))
plot(roc2, print.thres="best", print.thres.best.method="closest.topleft",
 col="black", xlab="specifiskums", ylab="jutigums", xilim=c(0,1), ulim=c(0,1))
plot(roc2, print.thres="best", print.thres.best.method="youden",
 col="black", add=T)
lines(c(1,0.647), c(1,0.733))
lines(c(0.627,0.627),c(0.756,0.38))

#simulacijas ROC liknem
library(logcondens)
library(caTools)
roc <-  function(contr, case, p)
{F  <-   c(function(p) punif (p, 0, 3),function(p) pnorm (p, 2, 1),
function(p) pexp (p, 2),function(p) pgamma (p, 4, 1.5),function(p)
 plogis (p, 2, 1))
G  <-  c(function(p) qunif (p, -2, 1),function(p) qnorm (p, 0, 1),
function(p) qexp (p, 1),function(p) qgamma (p, 2, 0.5),function(p)
 qlogis (p, 0, 1))
return (1 - sapply ( sapply ((1 - p), G[[contr]]), F[[case]]) )}

ASE <- function (nov, nov_emp, roc){sqrt(mean((nov - roc)^2) /
 mean((nov_emp - roc)^2))}
IAD <-  function (nov, roc, p){y  <-  abs(roc - nov); return (trapz(p,y))}

sim.data <-  function (type, class, n){
if ( type == 1) {if( class == 0) return (runif(s.size , -2, 1))
else return ( runif (s.size , 0, 3))}
else if ( type == 2) {if( class == 0) return ( rnorm (s.size ,
 mean = 0, sd = 1))
else return ( rnorm (s.size , mean = 2, sd = 1))}
else if ( type == 3) {if( class == 0) return ( rexp (s.size , 1))
else return ( rnorm (s.size , 2))}
else if ( type == 4) {if( class == 0) return ( rgamma (s.size , shape = 2,
 rate = 0.5) )
else return ( rgamma (s.size , shape = 4, rate = 1.5) )}
else if ( type == 5) {if( class == 0) return ( rlogis (s.size , location = 0,
 scale = 1))
else return ( rlogis (s.size , location = 2, scale = 1))}}
```

```
sim  <-  function(sad, s.size)
{p  <-  seq(0, 1, by = 0.01); n <- length(p)
pp  <-  p[n:1];true <- roc(sad[1], sad[2], p)
ASE.res <-   matrix(NA , ncol = 5, nrow = n)
IAD.res  <-   matrix(NA , ncol = 6, nrow = n )
for (i in 1:n) {controls <-  sim.data(sad[1], 0, s.size )
cases  <-  sim.data(sad[2], 1, s.size)
emp <-  coords(roc(controls = controls , cases = cases),
p = pp , input = "spec", ret = "sens") [1:n]
param <-  coords ( roc(controls = controls , cases = cases , smooth = T,
smooth.method = "fitdistr"), p = pp , input = "specificity",
 ret = "sensitivity")[1:n]
binorm <-  coords ( roc(controls = controls , cases = cases , smooth = T,
smooth.method = "binormal"), p = pp , input = "spec", ret = "sens") [1:n]
log <-  coords ( roc(controls = controls , cases = cases , smooth = T,
smooth.method = "logcondens"), p = pp , input = "spec", ret = "sens") [1:n]
kern  <- coords( roc(controls = controls , cases = cases , smooth = T,
smooth.method = "density"), p = pp , input = "spec", ret = "sens") [1:n]
nov  <- cbind(emp, param, binorm, log, kern)
ASE.res [i,] <-  apply(est[ ,2:6] , 2, ASE , emp = emp , true = true )
IAD.res [i,]  <- apply(est , 2, IAD , true = true , p = p)}
return ( cbind ( ASE.res , IAD.res ))}
```

Maģistra darbs "Divu ROC līkņu salīdzināšana" 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: Jeļena Vaļkovska

Rekomendēju darbu aizstāvēšanai

Vadītājs: Docents Dr. Math. Jānis Valeinis

Recenzents: doktorante Māra Vēliņa

Darbs iesniegts Matemātikas nodaļā

Dekāna pilnvarota persona: vecākā metodiķe Dzintra Holsta

Darbs aizstāvēts maģistra gala pārbaudījuma komisijas sēdē

_____ 06.2014. prot. Nr._____

Komisijas sekretāre: docente Silvija Čerāne