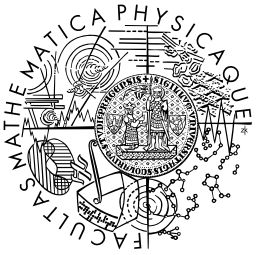


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Zuzana Zatkalíková

ROC křivka

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Petr Lachout, CSc
Studijní program: Matematika
Studijní obor: Finanční matematika

Praha 2023

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ....

Podpis autora

Rada by som sa poďakovala vedúcemu mojej bakalárskej práce docentovi Petrovi Lachoutovi za všetky cenné rady, pripomienky a odpovede na moje otázky počas písania tejto práce. Taktiež by som sa rada poďakovala kamarátovi Tomášovi Šumšalovi za ochotné prediskutovanie niektorých otázok ohľadom práce.

Název práce: ROC křivka

Autor: Zuzana Zatkalíková

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. RNDr. Petr Lachout, CSc, Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Tato bakalářská práce na úvod definuje základní pojmy a poté se zabývá ROC křivkou. Popisuje její význam, vlastnosti a konstrukci i s grafickým znázorněním. Následně je v práci odvozeno vyjádření ROC křivky a její plochy pro normální, exponenciální a rovnoměrné rozdělení, rovněž i s grafickým znázorněním. Dále je dána do souvisu se statistickým testováním. Na závěr je popsáno empirické vyjádření ROC křivky a její aplikace na reálných datech, která jsou zpracována v programovacím jazyce Python.

Klíčová slova: Statistický test, ROC křivka, rozdělení náhodné veličiny, kvantily

Title: ROC curve

Author: Zuzana Zatkalíková

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. RNDr. Petr Lachout, CSc, Department of Probability and Mathematical Statistic

Abstract: This bachelor's thesis firstly defines the basic terms and then describes the ROC curve. Thesis deals with meaning of the ROC curve, its properties and construction with a graphic representation. Subsequently, the expression of the ROC curve and its area for normal, exponential and uniform distribution is derived in the work, also with a graphical representation. Then, it is related to statistical testing. At the end, there are described the empirical expression of the ROC curve and its application to real data processed in the Python programming language.

Keywords: Statistical test, ROC curve, distribution of a random variable, quantiles

Názov práce: ROC krivka

Autor: Zuzana Zatkalíková

Katedra: Katedra pravdepodobnosti a matematickej štatistiky

Vedúci bakalárskej práce: doc. RNDr. Petr Lachout, CSc., Katedra pravdepodobnosti a matematickej štatistiky

Abstrakt: Táto bakalárska práca na úvod zadefinuje základné pojmy a potom sa zaoberá ROC krivkou. Popisuje jej význam, vlastnosti a konštrukciu aj s grafickým znázornením. Následne je v práci odvodené vyjadrenie ROC krivky a jej plochy pre normálne, exponenciálne a rovnomerné rozdelenie, taktiež aj s grafickým znázornením. Ďalej je daná do súvisu so štatistickým testovaním. Na záver je opísané empirické vyjadrenie ROC krivky a jej aplikácia na reálnych dátach, ktoré sú spracované v programovacom jazyku Python.

Kľúčové slová: Štatistický test, ROC krivka, rozdelenie náhodnej veličiny, kvantily

Obsah

Úvod

Táto bakalárska práca sa bude venovať ROC krivkám (Receiver Operating Characteristic Curves), ktoré sú v praxi využívané na vyhodnocovanie dát a určenie kvality nejakého testu. Zo začiatku práce uvedieme základné pojmy a definície, ktoré budú ďalej používané, ako sú binárna klasifikácia, senzitivita a špecificita, kontingenčná tabuľka, či hypotéza a alternatíva.

V druhej kapitole už budeme opisovať ROC krivky. Popíšeme ich históriu, vysvetlíme, čo znamenajú, ako sa konštruujú a uvedieme ich definíciu. Ďalej spomenieme aj základné vlastnosti, ich význam a vysvetlíme, kedy je diagnostický test podľa ROC krivky a jej plochy spoľahlivý.

Tretia kapitola sa bude zaoberať všeobecným vyjadrením ROC krivky a jej plochy pomocou distribučných funkcií, aby mohlo byť toto vyjadrenie následne aplikovateľné na rôzne pravdepodobnostné rozdelenia. Budeme sa venovať normálnemu, exponenciálnemu a rovnomernému rozdeleniu.

Ako posledná bude kapitola s využitím ROC kriviek na reálnych dátach. Najprv dáme ROC krivku do súvisu so štatistickým testovaním a definujeme jej empirické vyjadrenie. Následne spracujeme získané reálne dáta, skonštruujeme z nich empirickú ROC krivku a vypočítame hodnotu jej plochy, čím zistíme, za ako dôveryhodný môžeme daný test považovať.

1. Základné definície a pojmy

Aby sme vedeli ďalej v práci dobre popísať ROC krivky, zavedieme si najprv základné definície a pojmy, ktoré budú v práci spomínané. Všetky uvedené pojmy a definície sú čerpané z kníh (Anděl (1985)), (Anděl (2007)) a zo skrípt (Kulich (2014)).

Definícia 1 (Distribučná funkcia)**.** *Funkcia $F_X(x) = P[X \leq x]$, $x \in \mathbb{R}$ sa nazýva distribučná funkcia náhodnej veličiny X, kde náhodná veličina je merateľné zobrazenie.*

Distribučná funkcia jednoznačne určuje pravdepodobnostné rozdelenie náhodnej veličiny.

Pri výberovom súbore, ktorý môže byť napríklad z reálnych dát, používame empirickú distribučnú funkciu, ktorá je odhadom distribučnej funkcie a konverguje k nej s pravdepodobnosťou rovnej 1. Táto funkcia bude skokovitá a skoky závisia od počtu pozorovaní vo výberovom súbore.

Definícia 2 (Empirická distribučná funkcia)**.** *Pre postupnosť nezávislých a rovnako rozdelených náhodných veličín $X_1, ..., X_n$, nazývanú náhodný výber, nazveme funkciu $\widehat{F_n}(u) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty,u\rangle}(X_i)$ empirická distribučná funkcia náhodného výberu $X_1, ..., X_n$.*

Ďalej si definujeme kvantilovú funkciu a jej hodnoty nazývané kvantily.

Definícia 3 (Kvantilová funkcia a kvantil)**.** *Nech α je predom dané číslo z intervalu (0,1). Kvantilová funkcia rozdelenia F_X je definovaná ako*

$$F_X^{-1}(\alpha) = \inf\{x : F_X(x) \geq \alpha\}.$$

Potom α-kvantilom rozdelenia F_X je číslo $u_X(\alpha) = F_X^{-1}(\alpha)$.

Pre α-kvantil platí

$$F_X(u_X(\alpha)) \geq \alpha \text{ a } F_X(u_X(\alpha) - h) < \alpha \text{ pre } \forall h > 0.$$

Ako empirický odhad použijeme hodnotu α-kvantilu empirickej distribučnej funkcie, teda

$$\widehat{F}_n^{-1}(\alpha) = \inf\{x : \widehat{F}_n(x) \geq \alpha\}.$$

1.1 Testovanie hypotézy

Definícia 4 (Hypotéza a alternatíva)**.** *Nech $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ je náhodný výber nezávislých k-rozmerných náhodných vektorov s rozdelením $F_x \in \mathcal{F}$, kde $\mathcal{F}$ je model. Nech $\theta = t(F) \in \mathbb{R}^d$ je charakteristika rozdelenia, ktorá nás zaujíma (parameter), nech $\Theta = \{t(F), F \in \mathcal{F}\} \subseteq \mathbb{R}^d$ označuje všetky možné hodnoty parametra v modeli $\mathcal{F}$ (nazýva sa parametrický priestor). Zvolíme dve neprázdne disjunktné podmnožiny Θ, ktoré označíme ako Θ_0 a Θ_1. Množinu Θ_0 nazývame hypotéza a množinu Θ_1 alternatíva.*

Hypotézu zvyčajne označujeme ako H_0, alternatívu ako H_1 a testujeme hypotézu $H_0 : \theta_x \in \Theta_0$ proti alternatíve $H_1 : \theta_x \in \Theta_1$.

Definícia 5 (Štatistický test)**.** *Štatistický test je definovaný pomocou vhodne zvolenej funkcie dát $S_n(\boldsymbol{X})$, ktorú nazývame testová štatistika a množine C, ktorú voláme kritický obor. Rozhodujeme sa podľa toho, či testová štatistika padne do kritického oboru alebo nie. Ak $S_n(\boldsymbol{X}) \in C$, potom zamietame hypotézu H_0 v prospech alternatívy H_1. Ak $S_n(\boldsymbol{X}) \notin C$, potom hypotézu H_0 nemôžeme zamietnuť v prospech alternatívy H_1.*

Definícia 6 (Hladina významnosti testu)**.** *Nech $\alpha \in (0,1)$ je predom dané číslo. Ak kritický obor C spĺňa podmienku*

$$\sup_{\theta \in \Theta_0} \mathsf{P}_\theta[S_n(\boldsymbol{X}) \in C] = \alpha,$$

potom hovoríme, že test $S_n(\boldsymbol{X}) \in C$ má hladinu významnosti α.

Definícia 7 (Sila testu)**.** *Nech $\theta \in \Theta_1$, potom*

$$\beta(\theta) = \mathsf{P}_\theta[S_n(\boldsymbol{X}) \in C]$$

sa nazýva sila testu proti alternatíve.

Definícia 8 (Chyba 1. a 2. druhu)**.** *Ak test zamietne platnú hypotézu, hovoríme o chybe 1. druhu. Ak test nezamietne neplatnú hypotézu, hovoríme o chybe 2. druhu.*

	H_0 platí	H_0 neplatí
Nezamietame H_0	OK	chyba 2. druhu
Zamietame H_0	chyba 1. druhu	OK

Tabuľka 1.1: Tabuľkové vzťahy medzi pravdivosťou a nepravdivosťou hypotézy H_0 a výsledkami testu

Pravdepodobnosť chyby 1. druhu je hladina testu α a doplnok k pravdepodobnosti chyby 2. druhu je sila testu β. V obore možných hodnôt náhodnej veličiny sa určí taká časť, do ktorej za platnosti H_0 padne výsledok veličiny s pravdepodobnosťou α. Táto časť oboru možných hodnôt sa nazýva kritický obor. Oddeľujú ho od oboru prijatia kritické hodnoty, čo sú kvantily rozdelenia testovacieho kritéria za platnosti H_0.

1.2 Kontingenčná tabuľka 2x2

Máme dvojrozmerný náhodný vektor $\mathbf{X} = (X, D)^T$ taký, že X môže mať hodnoty 0,1 a D hodnoty 0,1. Označíme $p_{ij} = \mathsf{P}[X = i, D = j]$. Ďalej označíme $p_i = \mathsf{P}[X = i] = \sum_{j=0}^{1} p_{ij}$, $p_j = \mathsf{P}[D = j] = \sum_{i=0}^{1} p_{ij}$. Budeme predpokladať, že platí $p_{ij} > 0$ pre všetky dvojice (i, j). Teraz uvažujeme výber o rozsahu n z rozdelenia s pravdepodobnosťami p_{ij}. Označíme n_{ij} počet tých prípadov, kedy súčasne nastalo $X = i$ a $D = j$. Marginálne početnosti budeme označovať $n_{i+} = \sum_{j=0}^{1} n_{ij}$, $n_{+j} = \sum_{i=0}^{1} n_{ij}$ a celý rozsah ako $n = \sum_{i=0}^{1} \sum_{j=0}^{1} n_{ij}$ Potom môžeme výsledky zapísať do kontingenčnej tabuľky, ktorá obsahuje 2x2 početností a pravdepodobnosti do matice pravdepodobností.

	$D=0$	$D=1$	$\sum$
$X=0$	n_{00}	n_{01}	n_{0+}
$X=1$	n_{10}	n_{11}	n_{1+}
$\sum$	n_{+0}	n_{+1}	n

Tabuľka 1.2: Kontingenčná tabuľka 2x2

	$D=0$	$D=1$	$\sum$
$X=0$	p_{00}	p_{01}	p_{0+}
$X=1$	p_{10}	p_{11}	p_{1+}
$\sum$	p_{+0}	p_{+1}	1

Tabuľka 1.3: Matica pravdepodobností

1.3 Binárna klasifikácia

Binárna klasifikácia je klasifikovanie veličín z náhodného výberu do dvoch tried, nazveme ich negatívna trieda a pozitívna trieda. Na toto klasifikovanie používame binárnu veličinu $D = \{0, 1\}$, ktorá bude určovať správne klasifikovanie do týchto dvoch tried, čiže predstavuje skutočnosť.

$$D = \begin{cases} 0, & \text{pre negatívnu triedu} \\ 1, & \text{pre pozitívnu triedu} \end{cases}$$

Diagnostickým testom vyhodnocujeme zaradenie medzi pozitívnych a negatívnych na základe nejakých údajov. Výsledkom je nejaká veličina, ktorú označíme ako $\tilde{X}$. Predom stanovíme určitú hodnotu tejto veličiny, ktorú budeme nazývať klasifikačný prah a označovať ju budeme ako ϕ. Následne podľa tohoto klasifikačného prahu rozhodneme, či nastal pozitívny alebo negatívny výsledok testu a toto zaradenie označme X, ako sme označovali aj v kontingenčnej tabuľke.

$$X = \begin{cases} 0, & \text{pre negatívnych podľa testu} \\ 1, & \text{pre pozitívnych podľa testu} \end{cases}$$

Diagnostický test ale rozdeľuje do týchto tried s nejakou chybou a jeho výsledok sa nemusí vždy zhodovať s realitou. Výsledky testu môžu byť teda klasifikované ako skutočne pozitívni, skutočne negatívni, nesprávne pozitívni alebo nesprávne negatívni. Používajú sa pre ne skratky z anglických názvov, čiže skutočne negatívni sa označuje ako TN (true negative), nesprávne negatívni ako FN (false negative), nesprávne pozitívni ako FP (false positive) a skutočne pozitívni ako TP (true positive). Ich počty sa zapisujú do kontingenčnej tabuľky ako vidíme v tabuľke 1.4, kde sme vychádzali z tabuľky 1.2.

	$D = 0$	$D = 1$	$\sum$
$X = 0$	$TN = n_{00}$	$FN = n_{01}$	n_{0+}
$X = 1$	$FP = n_{10}$	$TP = n_{11}$	n_{1+}
$\sum$	n_{+0}	n_{+1}	n

Tabuľka 1.4: Klasifikácia výsledkov diagnostického testu

1.4 Senzitivita a špecificita

Senzitivita je citlivosť testu. Vyjadruje pravdepodobnosť, že test dá pozitívny výsledok pri pozitívnom objekte v závislosti na klasifikačnom prahu ϕ, podľa ktorého sme prípady delili na pozitívne a negatívne. Nadobúda hodnoty od 0 do 1. Empiricky je daná pomerom správne vyhodnotených pozitívnych prípadov ku všetkým možnostiam, ktoré sú v skutočnosti pozitívne. Ak by teda mal test 100-percentnú senzitivitu, znamenalo by to, že by test naozaj odhalil všetkých ľudí, ktorí sú pozitívni.

Špecificita vyjadruje pravdepodobnosť, že test dá negatívny výsledok pri negatívnom objekte opäť v závislosti na klasifikačnom prahu ϕ, podľa ktorého sme prípady delili medzi pozitívne a negatívne. Čiže špecificita sa dá popísať ako schopnosť testu správne vybrať všetky negatívne prípady. Empiricky je daná podielom skutočne negatívnych prípadov ku všetkým prípadom, ktoré sú v skutočnosti negatívne. Taktiež nadobúda hodnoty od 0 do 1. 100-percentná špecificita znamená, že všetky negatívne prípady by boli vyhodnotené ako naozaj negatívne. Nasledujúce dve definície sú parafrázované z knihy (Zhou a kol. (2002)).

Definícia 9 (Senzitivita)**.** *Pre danú hodnotu ϕ definujeme senzitivitu testu ako pravdepodobnosť* $\mathsf{P}\left[\tilde{X} \leq \phi | D = 1\right]$.

Definícia 10 (Špecificita)**.** *Pre danú hodnotu ϕ definujeme špecificitu testu ako pravdepodobnosť* $1 - \mathsf{P}\left[\tilde{X} \leq \phi | D = 0\right] = \mathsf{P}\left[\tilde{X} > \phi | D = 0\right]$.

1.5 Terminológia

TP (True positive): počet skutočne pozitívnych, označujeme aj n_{11}
TN (True negative): počet skutočne negatívnych, označujeme aj n_{00}
FP (False positive): počet nesprávne pozitívnych, označujeme aj n_{10}
FN (False negative): počet nesprávne negatívnych, označujeme aj n_{01}
Pre daný klasifikačný prah $\phi \in (-\infty, \infty)$ máme:
$TPR(\phi)$ (True positive rate): Sensitivita = $\mathsf{P}\left[\tilde{X} \leq \phi | D = 1\right]$
$TNR(\phi)$ (True negative rate): Špecificita = $\mathsf{P}\left[\tilde{X} > \phi | D = 0\right]$
$FPR(\phi)$ (True negative rate): 1 - Špecificita = $\mathsf{P}\left[\tilde{X} \leq \phi | D = 0\right]$, pravdepodobnosť chyby 1. druhu
$FNR(\phi)$ (True negative rate): 1 - Senzitivita = $\mathsf{P}\left[\tilde{X} > \phi | D = 1\right]$, pravdepodobnosť chyby 2. druhu
Empirické hodnoty:

$\widehat{TPR}(\phi)$ (True positive rate): Sensitivita $= TP/(TP + FN) = \frac{n_{11}}{n_{+1}}$

$\widehat{TNR}(\phi)$ (True negative rate): Špecificita $= TN/(FP + TN) = \frac{n_{00}}{n_{+0}}$

$\widehat{FPR}(\phi)$ (False positive rate): 1 - Špecificta $= FP/(TN + FP) = \frac{n_{10}}{n_{+0}}$, odhad pravdepodobnosti chyby 1. druhu

$\widehat{FNR}(\phi)$ (False negative rate): 1 - Senzitivita $= FN/(FN+TP) = \frac{n_{01}}{n_{+1}}$, odhad pravdepodobnosti 2. druhu

2. ROC krivka

V tejto kapitole sa zameriame na spojité testy. Predpokladáme, že ich rozdelenia majú kladnú hustotu a teda aj rastúcu, spojitú distribučnú funkciu. Na základe týchto testov budeme ROC krivky konštruovať.

2.1 História

ROC krivka vznikla pre potreby hodnotenia radarových signálov počas druhej svetovej vojny, z čoho vznikol aj jej názov Receiver Operating Characteristic Curve. Po útoku na Pearl Harbor v roku 1941, začala americká armáda nový výskum, ako by sa z radarových signálov dalo presnejšie vyhodnotiť, či ide o skutočné japonské lietadlá a tak ich odlíšiť od falošných poplachov, takže išlo o odlíšenie signálu od šumu. Neskôr, v šesťdesiatych rokoch, sa ROC krivka začala používať aj v medicíne na hodnotenie presnosti diagnostického testu, ktorého výsledkom je nejaká spojitá veličina, ale jej využitie sa rozšírilo až začiatkom osemdesiatych rokov, najmä v rádiológii. Dnes má ROC krivka široké využitie hlavne v oblasti medicíny a riadení rizík v bankovníctve.

2.2 Pojem ROC krivka

ROC krivka, po anglicky Receiver Operating Characteristic Curve, graficky znázorňuje klasifikáciu do dvoch tried a popisuje kvalitu testu v závislosti na nastavení jeho klasifikačného prahu. Používa sa teda na hodnotenie a optimalizáciu binárneho klasifikačného testu, ktorý ukazuje vzťah medzi špecificitou a senzitivitou testu alebo detektoru pre všetky prípustné hodnoty prahu.

2.3 Konštrukcia ROC krivky

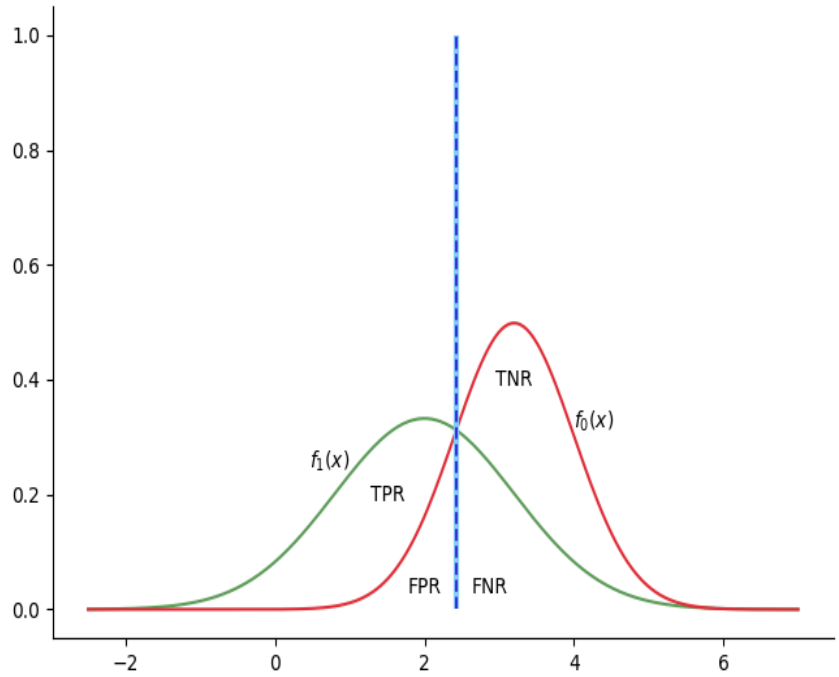
Na príklade si ukážeme ako skonštruovať ROC krivku. Uvažujeme populáciu, kde časť sú zdraví ľudia a časť chorí. Budeme mať nejaký diagnostický test, ktorý použijeme na túto zmiešanú skupinu ľudí. Test je založený na spojitej náhodnej veličine $\tilde{X}$, nazývanej prediktor, podľa ktorého budeme prípady klasifikovať do prvej alebo druhej skupiny. Zvolíme prvú skupinu ako pozitívne prípady a druhú ako negatívne. V praxi by to bolo napríklad pri testovaní množstva kalcia v tele človeka, pretože ľudia s nízkou hodnotou výsledku by boli považovaní za chorých (pozitívnych), tak by boli zaradení do prvej skupiny. Dalo by sa to zvoliť aj opačne, čiže prvú skupinu ako negatívnych a druhú ako pozitívnych, ale my sa v práci budeme venovať iba jednému zaradeniu. Opačné zaradenie by sa odvodzovalo analogicky.

Predpokladáme, že $\tilde{X}$ má v prvej skupine spojité rozdelenie s kladnou hustotou $f_1(x)$ a v druhej skupine má spojité rozdelenie s kladnou hustotou $f_0(x)$. Vykreslíme grafy oboch hustôt a zvolíme deliaci bod ϕ, nazývaný klasifikačný prah. Je to vlastne kvantil oboch rozdelení a zmenou hladiny spoľahlivosti meníme aj veľkosť senzitvity a špecificity. Vo väčšine klasifikačných metód určujeme klasifikačný prah tam, kde sa pretnú grafy hustôt, ale neexistuje žiadne všeobecne

platné pravidlo pre jeho výber. Pri diagnostickom testovaní všetko závisí od choroby, ktorá má byť diagnostikovaná, dostupná možnosť liečby a účel testu.

Plocha pod grafom hustoty $f_1(x)$ ležiaca naľavo od klasifikačného prahu ϕ je senzitivita, čiže vyjadruje pravdepodobnosť, že test dá pozitívny výsledok pri pozitívnom objekte. Tieto prípady označujeme skratkou $TPR(\phi)$. Na pravej strane od klasifikačného prahu pod grafom hustoty $f_1(x)$ je pravdepodobnosť podielu falošne negatívnych prípadov, čiže označíme $FNR(\phi)$. V prípade druhej hustoty $f_0(x)$ je plocha pod jej grafom ležiacej napravo od ϕ pravdepodobnosť podielu správne klasifikovaných negatívnych prípadov. Označujeme ju $TNR(\phi)$, čo je špecificita. Potom zostáva už len pravdepodobnosť falošne pozitívnych prípadov $FPR(\phi)$. Tie predstavuje plocha po grafom hustoty $f_0(x)$ naľavo od klasifikačného prahu.

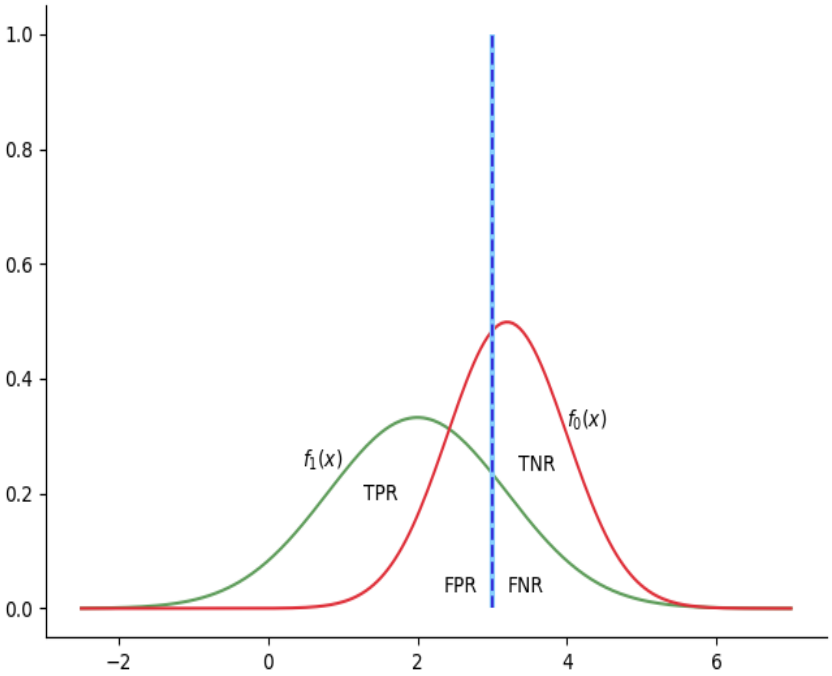
V prípade druhej hustoty $f_0(x)$ je plocha pod jej grafom ležiacej napravo od ϕ pravdepodobnosť podielu správne klasifikovaných negatívnych prípadov. Označujeme ju $TNR(\phi)$, čo je špecificita. Potom zostáva už len pravdepodobnosť falošne pozitívnych prípadov $FPR(\phi)$. Tie predstavuje plocha po grafom hustoty $f_0(x)$ naľavo od klasifikačného prahu.



Obr. 2.1: Graf hustôt normálneho rozdelenia

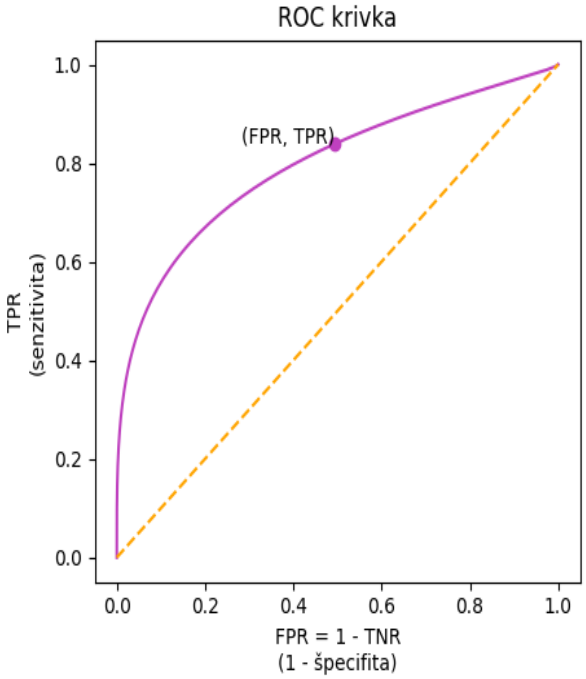
Na obrázku 2.2 môžeme vidieť, ako sa posunutím klasifikačného prahu ϕ, ktorý je znázornený modrou priamkou, zmenili veľkosti $FPR(\phi)$, $TPR(\phi)$, $FNR(\phi)$ a $TNR(\phi)$ v porovnaní s obrázkom 2.1.

Takýmto posúvaním ϕ a získavaním jednotlivých hodnôt môžeme skonštruovať ROC krivku. Znázorňuje vzťah medzi senzitivitou a špecificitou. Je to graf pravdepodobnosti podielu skutočne pozitívnych prípadov ($TPR(\phi)$), čiže senzitivity, na osi y a pravdepodobnosti podielu nesprávne pozitívnych ($FPR(\phi)$) na osi x, ktorá sa zapisuje aj pomocou špecificity ako $FPR(\phi) = 1 - TNR(\phi)$. Teda na vytvorenie celej ROC krivky musíme zakresliť $TPR(\phi)$ proti $FPR(\phi)$ s rozsahom od 0 do 1 pre všetky možné klasifikačné prahy ϕ.



Obr. 2.2: Zmena posunutím klasifikačného prahu ϕ

Začneme v ľavom dolnom rohu, kde $TPR(\phi)$ aj $FPR(\phi)$ musia mať hodnotu 0. Posúvaním klasifikačného prahu, budeme meniť veľkosti senzitivity a špecificity. Ak posunieme ϕ napravo, tak zväčšíme senzitivitu, ale zmenšíme špecificitu. Posunutím naľavo zmenšíme senzitivitu, ale zväčšíme špecificitu. Preto je teda x-ová os častejšie vyjadrovaná pomocou $FPR(\phi)$ a nie pomocou špecificity.



Obr. 2.3: ROC krivka

Posúvaním ϕ tak dostaneme spoločný nárast alebo spoločný pokles dvojice $(FPR(\phi), TPR(\phi))$, ktorou znázorňujeme ROC krivku. Postupne teda posúvame

klasifikačný prah a tým dostávame hodnoty ROC krivky, ako vidíme na obrázku 2.3, až kým neprídeme k hodnote 1 pre $TPR(\phi)$ aj $FPR(\phi)$ nachádzajúcej sa v pravom hornom rohu grafu.

Klasifikácia, ktorá dobre rozdeľuje triedy, bude mať ROC krivku, ktorá sa tiahne blízko k ľavému hornému rohu grafu. Naopak, klasifikácia, ktorá zle rozdeľuje triedy, bude mať ROC krivku blízko diagonálnej krivke, znázornenej prerušovanou žltou čiarou, ktorá reprezentuje klasifikáciu, ktorá nie je lepšia než náhodný odhad. Každá ROC krivka, ktorá sa tiahne pod touto diagonálou predstavuje zlý test, ktorý s väčšou pravdepodobnosťou dá zlý výsledok než dobrý.

Definícia 11 (Definícia ROC krivky, parafrázované vo vlastnom značení z knihy (Pepe (2003)))**.** *Máme súbor dát, je zložený z dvoch tried, ktoré sú disjunktné vzhľadom k vlastnosti $D = \{0, 1\}$. Nech ϕ je klasifikačný prah, potom definujeme výsledok spojitého testu $\tilde{X}$, kde $\tilde{X}$ je spojitá náhodná veličina, ako*

$$\begin{aligned} \textit{negatívny ak} \quad & \tilde{X} > \phi, \\ \textit{pozitívny ak} \quad & \tilde{X} \leq \phi. \end{aligned}$$

Pre danú hodnotu ϕ ďalej určíme odpovedajúce hodnoty skutočne a falošne pozitívnych podielov

$$\begin{aligned} TPR(\phi) &= \mathsf{P}\left[\tilde{X} \leq \phi | D = 1\right] = F_1(\phi), \\ FPR(\phi) &= \mathsf{P}\left[\tilde{X} \leq \phi | D = 0\right] = F_0(\phi), \end{aligned}$$

kde $F_1(\phi)$ a $F_0(\phi)$ sú podmienené distribučné funkcie. ROC krivku potom definujeme ako množinu všetkých možných hodnôt $TPR(\phi)$ a $FPR(\phi)$.

ROC krivka je založená na kontingenčnej tabuľke pravdepodobností správneho a nesprávneho zaradenia pozitívnej a negatívnej vzorky ako vidíme v tabuľke 1.4, parafrázovanej z knihy (Zhou a kol. (2002)).

Pre predom zvolené ϕ platí, že $\tilde{X} > \phi$ je klasifikovaná trieda ako negatívni, pre $\tilde{X} \leq \phi$ ako pozitívni.

	Skutočná trieda	
Klasifikovaná trieda	**D=0**	**D=1**
$\tilde{X} > \phi$	TNR	FNR
$\tilde{X} \leq \phi$	FPR	TPR

Tabuľka 2.1: Kontingenčná tabuľka pravdepodobností výsledkov testu

Z terminológie v podkapitole 1.5 vyplýva, že musí platiť $TPR(\phi)+FNR(\phi) = 1$ a taktiež $TNR(\phi) + FPR(\phi) = 1$.

2.4 Vlastnosti

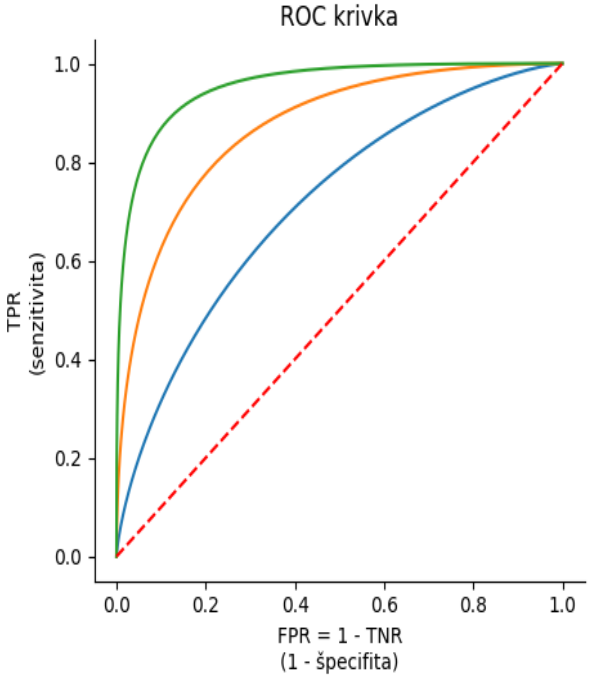
Pre spojité rozdelenia s kladnými hustotami platí, že keď narastá hodnota klasifikačného prahu ϕ, rastú aj hodnoty $FPR(\phi)$ a $TPR(\phi)$ ako môžeme vidieť na obrázku grafu hustôt 2.1. Ak teda $\phi = \infty$, máme $\lim_{\phi\to\infty} TPR(\phi) = 1$ a

$\lim_{\phi\to\infty} FPR(\phi) = 1$. Opačne zase, ak $\phi = -\infty$, máme $\lim_{\phi\to-\infty} TPR(\phi) = 0$ a $\lim_{\phi\to-\infty} FPR(\phi) = 0$. Preto je ROC krivka monotónna rastúca funkcia z intervalu [0,1] na [0,1].

Definičný obor ROC krivky je teda interval [0,1] a obor hodnôt je tiež interval [0,1]. Množina $[0,1]^2 = \{(FPR, TPR), FPR \in (0,1), TPR \in (0,1)\}$ sa nazýva ROC priestor.

ROC krivka nezávisí na konkrétnom umiestnení dát na číselnej ose. Závisí iba na poradí dát, takže je invariantná k monotónnej rastúcej transformácií prediktoru $\tilde{X}$.

Diagnostický test je užitočnejší, čím vyššia je jeho senzitivita a špecificita. Test, ktorý každý pozitívny výsledok určí ako pozitívny a každý negatívny výsledok ako negatívny, je nazývaný perfektný. Hovoríme aj, že test má perfektnú diskriminačnú schopnosť. ROC krivka na grafe kopíruje ľavý horný roh ROC priestoru. Naopak, pre test s nulovou diskriminačnou schopnosťou platí, že teoretická ROC krivka pre náhodný prediktor, čiže prediktor nezávislý od stavu ochorenia, je diagonála idúca z ľavého dolného rohu do pravého horného rohu ROC priestoru. Tento test sa považuje za bezvýznamný. Čím bližšie je teda ROC krivka k ľavému hornému rohu ROC priestoru, tým lepšia je diskriminačná schopnosť testu. V praxi väčšina testov leží medzi perfektným a bezvýznamným testom. Znázornenie týchto dvoch testov pomocou ROC kriviek môžeme neskôr vidieť aj na obrázku 3.5.



Obr. 2.4: Porovnanie ROC kriviek.

Na grafe 2.4 vidíme červenou prerušovanou priamkou znázornenú krivku pre bezvýznamný test, modrou znázornenú krivku pre lepší test, oranžovou pre ešte spoľahlivejší test a zelenou je znázornená krivka pre najspoľahlivejší test.

Aby sme jedným číslom vyjadrili kvalitu testu, počítame veľkosť plochy pod ROC krivkou. Táto plocha sa nazýva AUC z anglického Area Under Curve. Vypočítame ju ako $AUC = \int_0^1 ROC(t)dt$, $t \in [0,1]$. Najlepší diagnostický test bude

mať najväčšiu plochu pod ROC krivkou. Ak $AUC = 1$, tak test je perfektný a má sto percentnú senzitivitu a špecificitu. Naopak, ak je $AUC = 0$, tak je test perfektne nepresný, čiže zdraví pacienti budú mať výsledok, že sú chorí a chorí pacienti budú mať výsledok, že sú zdraví. Diagonála $AUC = 0{,}5$ je považovaná za hranicu odkedy test začína mať nejakú významnosť. Hodnota 0,5 plochy pod ROC krivkou je ešte len úplne náhodný test, ale so zvyšujúcou sa hodnotou narastá aj dôveryhodnosť testu. Ako aj vidíme na grafe, pod zelenou ROC krivkou, ktorá predstavuje najlepší test, je aj najväčšia plocha.

Pre spojité testy, kde máme kladnú diferencovateľnú hustotu, vieme vypočítať sklon ROC krivky a overiť jej konkavitu. Budeme vychádzať zo vzťahu $ROC(t) = F_1(F_0^{-1}(t))$, kde $t \in (0{,}1)$, ktorý odvodíme v nasledujúcej kapitole. Keďže ROC krivka je funkcia na intervale (0,1) a platí $ROC(0) = 0$, $ROC(1) = 1$, čo dobre vidíme aj na obrázku 2.4, tak jej sklon vypočítame pomocou prvej derivácie ako

$$ROC'(t) = (F_1(F_0^{-1}(t)))' = F_1'(F_0^{-1}(t))(F_0^{-1})'(t) =$$
$$= F_1'(F_0^{-1}(t))\frac{1}{F_0'(F_0^{-1})(t))} = \frac{f_1(F_0^{-1}(t))}{f_0(F_0^{-1}(t))},$$

kde vo výpočte sme používali retiazkové pravidlo a vzorec pre deriváciu inverznej funkcie. Používané distribučné funkcie sú za našich predpokladov spojité a rastúce, takže aj kvantilová funkcia $F_0^{-1}(t)$ bude spojitá a rastúca. ROC krivka má teda sklon $\frac{f_1(F_0^{-1}(t))}{f_0(F_0^{-1}(t))}$, ktorý je z predpokladu kladnej hustoty kladný v každom bode. Z matematickej analýzy vieme, že v bodoch, kde je prvá derivácia kladná, je funkcia rastúca. Takže za predpokladu rastúcej kvantilovej funkcie a kladnej hustoty platí, že $\frac{f_1(F_0^{-1}(t))}{f_0(F_0^{-1}(t))} > 0$, pre všetky $t \in (0, 1)$, čiže ROC krivka je za týchto predpokladov rastúca.

Druhá derivácia, podľa vzorca derivácie podielu, bude potom

$$\left(\frac{f_1(F_0^{-1}(t))}{f_0(F_0^{-1}(t))}\right)' =$$
$$= \frac{f_1'(F_0^{-1}(t))\frac{1}{f_0(F_0^{-1}(t))}f_0(F_0^{-1}(t)) - f_1(F_0^{-1}(t))f_0'(F_0^{-1}(t))\frac{1}{f_0(F_0^{-1}(t))}}{[f_0(F_0^{-1}(t))]^2} =$$
$$= \frac{f_1'(F_0^{-1}(t)) - \frac{f_1(F_0^{-1}(t))}{f_0(F_0^{-1}(t))}f_0'(F_0^{-1}(t))}{[f_0(F_0^{-1}(t))]^2}.$$

Z matematickej analýzy opäť vieme, že ak druhá derivácia funkcie je nekladná na otvorenom intervale, tak je na tomto intervale daná funkcia konkávna. Keď má funkcia na otvorenom intervale druhú deriváciu zápornú, tak je daná funkcia na tomto intervale rýdzo konkávna.

Menovateľ v druhej derivácií je druhá mocnina, takže bude vždy kladný. Z odvodenia v podkapitole 3.1 bude platiť, že $\phi = F_0^{-1}(t)$, čiže $F_0^{-1}(t)$ predstavuje osu x na grafe daných hustôt f_1 a f_0. Takže ROC krivka bude konkávna, ak

$$f_1'(F_0^{-1}(t)) < \frac{f_1(F_0^{-1}(t))}{f_0(F_0^{-1}(t))}f_0'(F_0^{-1}(t)),$$

čiže keď smernica hustoty f_1 bude menšia než smernica hustoty f_0 vynásobená podielom týchto hustôt v danom funkčnom bode.

3. ROC krivka pre rôzne pravdepodobnostné rozdelenia

3.1 Vyjadrenie ROC krivky a jej plochy AUC

Označíme $\tilde{X}_0$ ako veličiny z triedy D = 0 a $\tilde{X}_1$ ako veličiny z triedy D = 1. Keďže platí, že $\tilde{X} \leq \phi$ klasifikujeme ako pozitívnych a $\tilde{X} > \phi$ ako negatívnych, vyjadríme súradnice ROC krivky $(FPR(\phi), TPR(\phi))$ ako

$$FPR(\phi) = \mathsf{P}\left[\tilde{X}_0 \leq \phi\right] = F_0(\phi),$$
$$TPR(\phi) = \mathsf{P}\left[\tilde{X}_1 \leq \phi\right] = F_1(\phi),$$

kde vychádzame z definície 11. Vďaka týmto vzťahom odvodíme vzorec pre ROC krivku postupom

$$t = F_0(\phi), \tag{3.1}$$

$$\phi = F_0^{-1}(t), \tag{3.2}$$

$$F_1(\phi) = F_1(F_0^{-1}(t)), \tag{3.3}$$

pre $0 \leq t \leq 1$. Potom platí vzorec

$$ROC(t) = F_1(F_0^{-1}(t)).$$

Pre odvodenie plochy AUC použijeme integrál

$$AUC = \mathsf{P}\left[\tilde{X}_1 < \tilde{X}_0\right] = \int_0^1 ROC(t)\,\mathrm{d}t = \int_0^1 F_1(F_0^{-1}(t))\,\mathrm{d}t. \tag{3.4}$$

Pri takomto odvodení ROC krivky a jej plochy predpokladáme, že dané rozdelenie má spojitú a rastúcu distribučnú funkciu na určitom intervale. Keby nemalo, tak kvantilová funkcia nemusí byť vždy inverznou funkciou k distribučnej funkcií. V nasledujúcich troch podkapitolách sa budeme venovať normálnemu, exponenciálnemu a rovnomernému rozdeleniu, kde si zavedieme ich hustoty aj distribučné funkcie. Všetky tieto rozdelenia majú kladnú hustotu na nejakom intervale, pre normálne rozdelenie je to dokonca celé $\mathbb{R}$, a mimo tohoto intervalu majú hustotu nulovú a ich distribučné funkcie sú rastúce a spojité, takže spĺňajú predpoklady na daných intervaloch.

3.2 Normálne rozdelenie

Pre veličiny rozdelené do dvoch tried, kde obe triedy majú normálne rozdelenie, hovoríme o binormálnom modeli ROC krivky. Tento model hrá dôležitú rolu v ROC analýze, je to klasický model ROC krivky.

Definícia 12 (Normálne rozdelenie, parafrázované z knihy (Anděl (2007))). *Náhodná veličina X so strednou hodnotou $\mu \in \mathbb{R}$ a rozptylom $\sigma^2 > 0$ má normálne rozdelenie, ak jej hustota má tvar*

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Označujeme ako $X \sim \mathcal{N}(\mu,\sigma^2)$. Pre distribučnú funkciu $F(x)$ platí

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

kde Φ je distribučná funkcia normovaného normálneho rozdelenia $\mathcal{N}(0,1)$.

Máme teda

$$\tilde{X}_1 \sim \mathcal{N}(\mu_1,\sigma_1^2),$$
$$\tilde{X}_0 \sim \mathcal{N}(\mu_0,\sigma_0^2).$$

Názorný graf hustôt normálneho rozdelenia sme mohli vidieť na obrázku 2.1.

Pre hocijaký klasifikačný prah ϕ platí, že

$$FPR(\phi) = \mathsf{P}\left[\tilde{X}_0 \leq \phi\right] = \Phi\left(\frac{\phi-\mu_0}{\sigma_0}\right),$$
$$TPR(\phi) = \mathsf{P}\left[\tilde{X}_1 \leq \phi\right] = \Phi\left(\frac{\phi-\mu_1}{\sigma_1}\right).$$

Keďže funkcia ROC krivky sa dá zapísať pre $t \in (0,1)$ ako $ROC(t) = TPR(FPR)$, tak vyjadríme najprv ϕ pomocou $FPR(\phi)$ a následne dosadíme do $TPR(\phi)$ ako v postupe (3.1), (3.2) a (3.3)

$$t = FPR(\phi) = \Phi\left(\frac{\phi-\mu_0}{\sigma_0}\right),$$
$$\Phi^{-1}(t) = \frac{\phi-\mu_0}{\sigma_0},$$
$$\phi = \mu_0 + \sigma_0\Phi^{-1}(t).$$

Vyjde nám teda, že

$$TPR(\phi) = \Phi\left(\frac{\mu_0-\mu_1+\sigma_0\Phi^{-1}(t)}{\sigma_1}\right).$$

Funkciu ROC krivky pre normálne rozdelenie potom zapisujeme

$$ROC(t) = \Phi(a + b\Phi^{-1}(t)),$$

kde $a = \frac{\mu_0-\mu_1}{\sigma_1}$ a $b = \frac{\sigma_0}{\sigma_1}$. Krivka je teda definovaná pomocou dvoch parametrov a a b. Parameter a vyjadruje rozdiel stredných hodnôt rozdelení testov pozitívnych a negatívnych a parameter b vyjadruje pomer pozitívnych a negatívnych. ROC krivku k príslušným hustotám normálneho rozdelenia sme mohli vidieť na obrázku 2.3.

Plochu AUC pod ROC krivkou pre normálne rozdelenie vyjadrujeme pomocou vzorca (3.4) ako

$$AUC = \mathsf{P}\left[\tilde{X}_0 > \tilde{X}_1\right] = \mathsf{P}\left[\tilde{X}_0 - \tilde{X}_1 > 0\right] =$$

$$= 1 - \Phi\left(\frac{\mu_0 - \mu_1}{\sqrt{\sigma_0^2 + \sigma_1^2}}\right) = \Phi\left(\frac{\frac{\mu_0 - \mu_1}{\sigma_1}}{\sqrt{1 + \frac{\sigma_1^2}{\sigma_0^2}}}\right).$$

Z čoho nám vyjde, že plocha pod ROC krivkou je

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

3.3 Exponenciálne rozdelenie

Definícia 13 (Exponenciálne rozdelenie, parafrázované z knihy (Anděl (2007)))**.** *Náhodná veličina X má exponenciálne rozdelenie, ak jej hustota má tvar*

$$f(x) = \lambda e^{-\lambda x}, \qquad x > 0,$$
$$f(x) = 0, \qquad \textit{inak,}$$

kde $\lambda > 0$ a stredná hodnota je $\mathsf{E}\, X = \frac{1}{\lambda}$. Označujeme ako $X \sim Exp(\lambda)$. Distribučná funkcia pre exponenciálne rozdelenie je

$$F(x) = 1 - e^{-\lambda x}, \qquad x > 0,$$
$$F(x) = 0, \qquad \textit{inak.}$$

Potom budeme mať

$$\tilde{X}_1 \sim Exp(\lambda_1),$$
$$\tilde{X}_0 \sim Exp(\lambda_0),$$

kde $\lambda_1 > 0$ aj $\lambda_0 > 0$.

Podobne ako pri normálnom rozdelení platí, že

$$FPR(\phi) = \mathsf{P}\left[\tilde{X}_0 \leq \phi\right] = 1 - e^{-\lambda_0 \phi},$$
$$TPR(\phi) = \mathsf{P}\left[\tilde{X}_1 \leq \phi\right] = 1 - e^{-\lambda_1 \phi},$$

ale len pre $\phi \in (0, \infty)$. Pre $\phi < 0$ by tieto hodnoty boli nulové.

Následne, podľa postupu (3.1), (3.2) vyjadríme ϕ

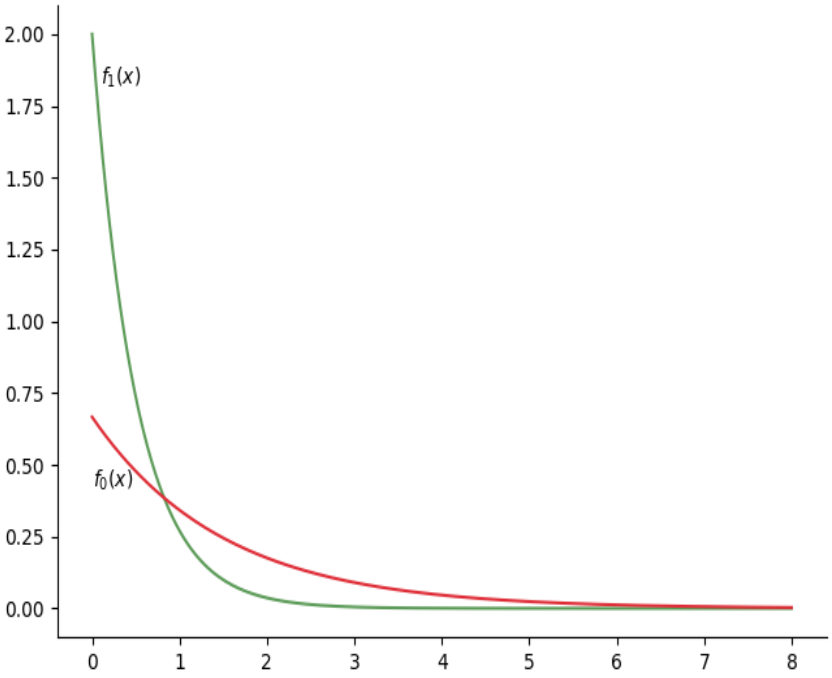
$$t = FPR(\phi) = 1 - e^{-\lambda_0 \phi},$$

$$e^{-\lambda_0 \phi} = 1 - t,$$

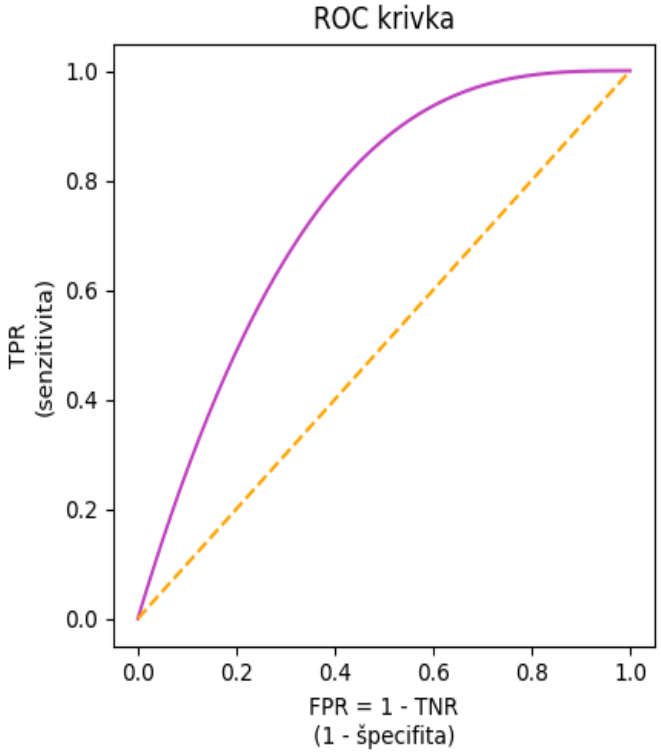
$$\phi = -\frac{\ln(1-t)}{\lambda_0},$$

kde ϕ je kladné ak t je kladné, čo ale platí vždy, keďže $t \in (0{,}1)$. Dosadíme do rovnice (3.3), aby sme dostali predpis ROC krivky

$$F_1(F_0^{-1}(t)) = 1 - e^{-\lambda_1(-\frac{\ln(1-t)}{\lambda_0})} = 1 - e^{\frac{\lambda_1}{\lambda_0}\ln(1-t)} = 1 - (1-t)^{\frac{\lambda_1}{\lambda_0}}.$$



Obr. 3.1: Graf hustôt exponenciálneho rozdelenia



Obr. 3.2: ROC krivka pre dané hustoty exponenciálneho rozdelenia

Funkcia ROC krivky pre exponenciálne rozdelenie je teda

$$ROC(t) = 1 - (1-t)^{\frac{\lambda_1}{\lambda_0}}.$$

Zo vzorca (3.4) vypočítame zintegrovaním $ROC(t)$ plochu AUC.

$$AUC = \int_0^1 ROC(t)\,\mathrm{d}t = \int_0^1 1 - (1-t)^{\frac{\lambda_1}{\lambda_0}}\mathrm{d}t =$$

$$= \frac{\lambda_1}{\lambda_0 + \lambda_1},$$

kde poslednú rovnosť sme dostali pomocou substitúcie.

3.4 Rovnomerné rozdelenie

ROC krivka pre rovnomerné rozdelenie sa odvádza z časti iným postupom, čo je spôsobené tvarom jeho hustoty a jej nenulovosti iba na určitom intervale. Pre rôzne možnosti umiestnenia hustôt, nám vznikne iný tvar ROC krivky. Týchto možností je 11 a teda na lepší opis použijeme aj grafy.

Definícia 14 (Rovnomerné rozdelenie, parafrázované z knihy (Anděl (2007)))**.** *Náhodná veličina X má rovnomerné rozdelenie na intervale (a,b), ak jej hustota má tvar*

$$f(x) = \frac{1}{b-a}, \qquad a < x < b,$$
$$f(x) = 0, \qquad \textit{inak.}$$

Označujeme ako $X \sim \mathcal{R}(a,b)$. Pre distribučnú funkciu $F(x)$ platí

$$F(x) = \frac{x-a}{b-a}, \qquad a < x < b,$$
$$F(x) = 0, \qquad x \leq a,$$
$$F(x) = 1, \qquad x \geq b.$$

Stredná hodnota rovnomerného rozdelenia je $\mathsf{E}\ X = \frac{a+b}{2}$, pre jednoduchosť ju označím ako c. Budeme mať teda

$$\tilde{X_1} \sim \mathcal{R}(c_1 - d_1, c_1 + d_1),$$
$$\tilde{X_0} \sim \mathcal{R}(c_0 - d_0, c_0 + d_0),$$

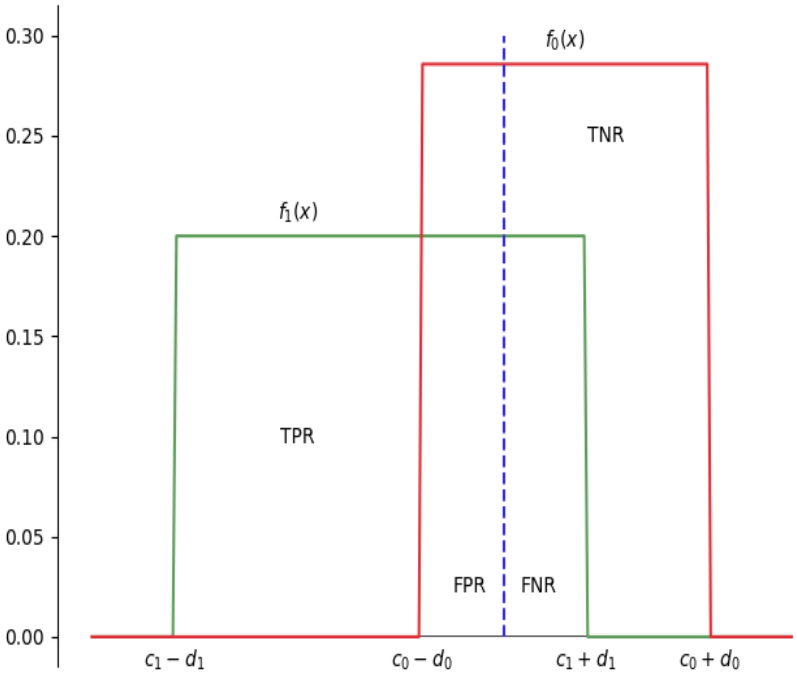
kde c_1, c_0 sú stredné hodnoty a d_1, d_0 sú odchýlky od stredných hodnôt.

Z podkapitoly 3.1 vyjadríme $FPR(\phi)$ a $TPR(\phi)$ ako

$$FPR(\phi) = \mathsf{P}\left[\tilde{X}_0 \leq \phi\right] = F_0(\phi),$$
$$TPR(\phi) = \mathsf{P}\left[\tilde{X}_1 \leq \phi\right] = F_1(\phi),$$

kde klasifikačný prah ϕ je znázornený modrou priamkou na obrázku 3.3 a z definície 14 rovnomerného rozdelenia vieme, že platí

$$f_1(\phi) = \frac{1}{b-a} = \frac{1}{c_1 + d_1 - c_1 + d_1} = \frac{1}{2d_1},$$



Obr. 3.3: Graf hustôt rovnomerného rozdelenia

pre $\phi \in (c_1 - d_1, c_1 + d_1)$. Potom teda

$$F_1(\phi) = \int_{-\infty}^{\phi} f_1(y)\,\mathrm{d}y = \frac{1}{2d_1}\int_{-\infty}^{\phi} \mathbb{1}_{(c_1-d_1,c_1+d_1)}(y)\mathrm{d}y = \frac{\phi - c_1 + d_1}{2d_1}.$$

Funkciu $F_0(\phi)$ dostaneme podobným postupom ako

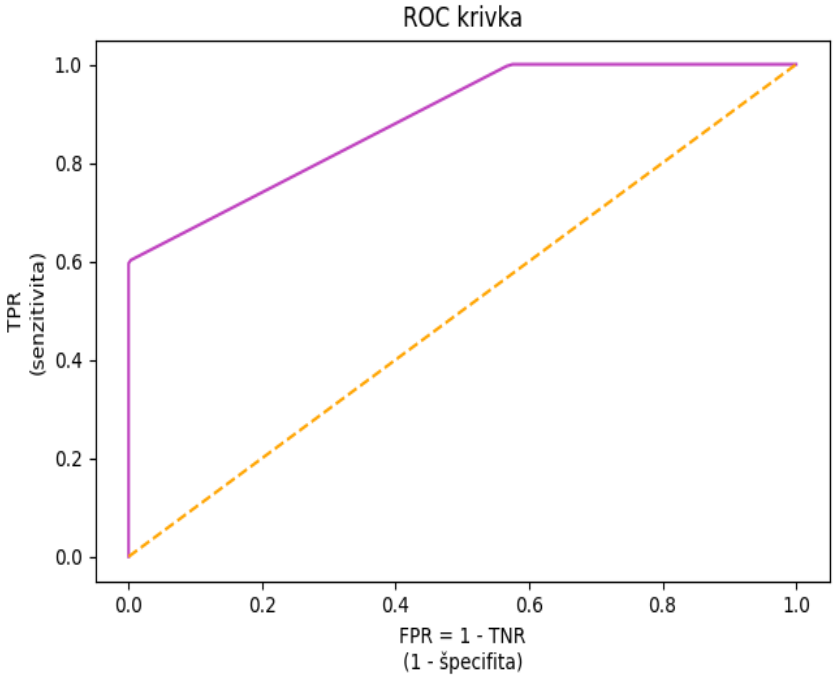
$$F_0(\phi) = \frac{\phi - c_0 + d_0}{2d_0}.$$

Ako sme už spomínali, keďže hustoty rovnomerného rozdelenia sú nenulové iba na určitom intervale a ich graf na tomto intervale sa rovnomerne udržuje na jednej hodnote, ako môžeme vidieť na obrázku 3.3, tak bude ROC krivka vykresľovaná trochu inak, než pri normálnom a exponenciálnom rozdelení. Bude sa skladať z lineárnych častí, ako môžeme vidieť na obrázku 3.4 a tieto lineárne časti budú rôzne pre rôzne umiestnenie hustôt. Pre lepšie pochopenie, popíšme podrobnejšie jednu možnosť a potom podobným postupom znázorníme aj ďalšie možnosti.

Na grafe 3.3 vidíme, že x-ovú os delia hustoty na 5 intervalov. Umiestnime klasifikačný prah ϕ úplne naľavo na grafe, v tomto prípade na interval $(-\infty, c_1 - d_1)$, a budeme ho postupne posúvať smerom doprava, aby prešiel všetkými intervalmi.

Pre prvý interval bude $ROC(t) = ROC(0) = 0$, keďže hodnoty $TPR(\phi)$ aj $FPR(\phi)$ sú nulové. Následne prejde ϕ do intervalu $(c_1 - d_1, c_0 - d_0)$, kde, ako vidíme na grafe 3.3, sa bude zväčšovať iba hodnota $TPR(\phi)$, keďže zasahuje ϕ iba do hustoty $f_1(x)$, ale ešte nezasahuje do plochy pod hustotou $f_0(x)$, čiže $FPR(\phi) = 0$. To znamená, že na grafe ROC krivky sa bude vykresľovať iba zvislá časť na osi y, ktorá predstavuje $TPR(\phi)$. Táto zvislá časť pôjde od bodu $(0, 0)$ do bodu, kde $\phi = c_0 - d_0$, čiže $(0, TPR(c_0 - d_0))$.

Keď ϕ prejde do intervalu $(c_0 - d_0, c_1 + d_1)$, začne tým zasahovať do oboch hustôt a bude vznikať druhá časť ROC krivky, pretože budú rásť hodnoty $TPR(\phi)$ aj



Obr. 3.4: ROC krivka pre dané hustoty rovnomerného rozdelenia

$FPR(\phi)$. Vzorec tejto strednej časti dostaneme postupom (3.1), (3.2) a následne, po vyjadrení ϕ, dosadíme do vzorca (3.3) ako $ROC(t) = F_1(F_0^{-1}(t))$ a dostaneme

$$\phi = c_0 - d_0 + 2td_0,$$

$$F_1(F_0^{-1}(t)) = \frac{c_0 - d_0 + 2td_0 - c_1 + d_1}{2d_1} = \frac{d_0}{d_1}t + \frac{\delta d - \delta c}{2d_1},$$

kde $\delta c = c_1 - c_0$ a $\delta d = d_1 - d_0$. Túto časť funkcie ROC krivky pre rovnomerné rozdelenie teda zapíšeme ako

$$ROC(t) = \frac{d_0}{d_1}t + \frac{\delta d - \delta c}{2d_1}, \tag{3.5}$$

kde $t \in (0, \frac{d_1+d_0+\delta c}{2d_0})$. Keď nastane rovnosť $\phi = c_1 - d_1$, tak $TPR(\phi) = 1$. Horná hodnota t je teda bod, kde sa ROC krivka dostane na úroveň $ROC(t) = 1$, teda

$$1 - ROC(t) = 0,$$

$$1 - \frac{d_0}{d_1}t - \frac{\delta d - \delta c}{2d_1} = 0,$$

$$t = \frac{d_1 + d_0 + \delta c}{2d_0}. \tag{3.6}$$

Pre štvrý interval $\phi \in (c_1 + d_1, c_0 + d_0)$ už ϕ nezasahuje do hustoty $f_1(x)$. Hodnota $TPR(\phi)$ dosiahne hodnoty $TPR(\phi) = 1$ a už sa meniť nebude, pretože zaberá celú plochu pod grafom hustoty $f_1(x)$. Bude sa už len zvyšovať hodnota $FPR(\phi)$, takže zvyšok ROC krivky bude mať predpis $ROC(t) = 1$ pre $t \in$

$(\frac{d_1+d_0+\delta c}{2d_0}, 1)$. Keď ϕ dosiahne hodnotu $\phi = c_0 + d_0$, tak sa ROC krivka dostane do bodu (1,1). Tam ROC krivka aj končí, pretože na poslednom intervale (c_0+d_0, ∞) sa už nemenia hodnoty $TPR(\phi)$ ani $FPR(\phi)$.

Keďže obsah ROC priestoru má hodnotu 1, čo vyplýva z vlastností ROC krivky, tak hodnotu plochy pod ROC krivkou vypočítame odpočítaním veľkosti plochy nad ROC krivkou (vrámci ROC priestoru) od 1. Veľkosť plochy nad strednou časťou ROC krivky dostaneme ako

$$\int_0^{\frac{d_1+d_0+\delta c}{2d_0}} (1 - ROC(t))\,\mathrm{d}t,$$

takže AUC, čiže plochu pod celou ROC krivkou, dostaneme ako

$$AUC = 1 - \int_0^{\frac{d_1+d_0+\delta c}{2d_0}} (1 - ROC(t))\,\mathrm{d}t = 1 - \int_0^{\frac{d_1+d_0+\delta c}{2d_0}} (1 - \frac{d_0}{d_1}t - \frac{\delta d - \delta c}{2d_1})\,\mathrm{d}t =$$

$$= 1 - \int_0^{\frac{d_1+d_0+\delta c}{2d_0}} 1\mathrm{d}t + \frac{d_0}{d_1}\int_0^{\frac{d_1+d_0+\delta c}{2d_0}} t\mathrm{d}t + \frac{\delta d - \delta c}{2d_1}\int_0^{\frac{d_1+d_0+\delta c}{2d_0}} 1\mathrm{d}t$$

$$AUC = \frac{(\delta c + d_1 + d_0)^2}{8 d_0 d_1}, \tag{3.7}$$

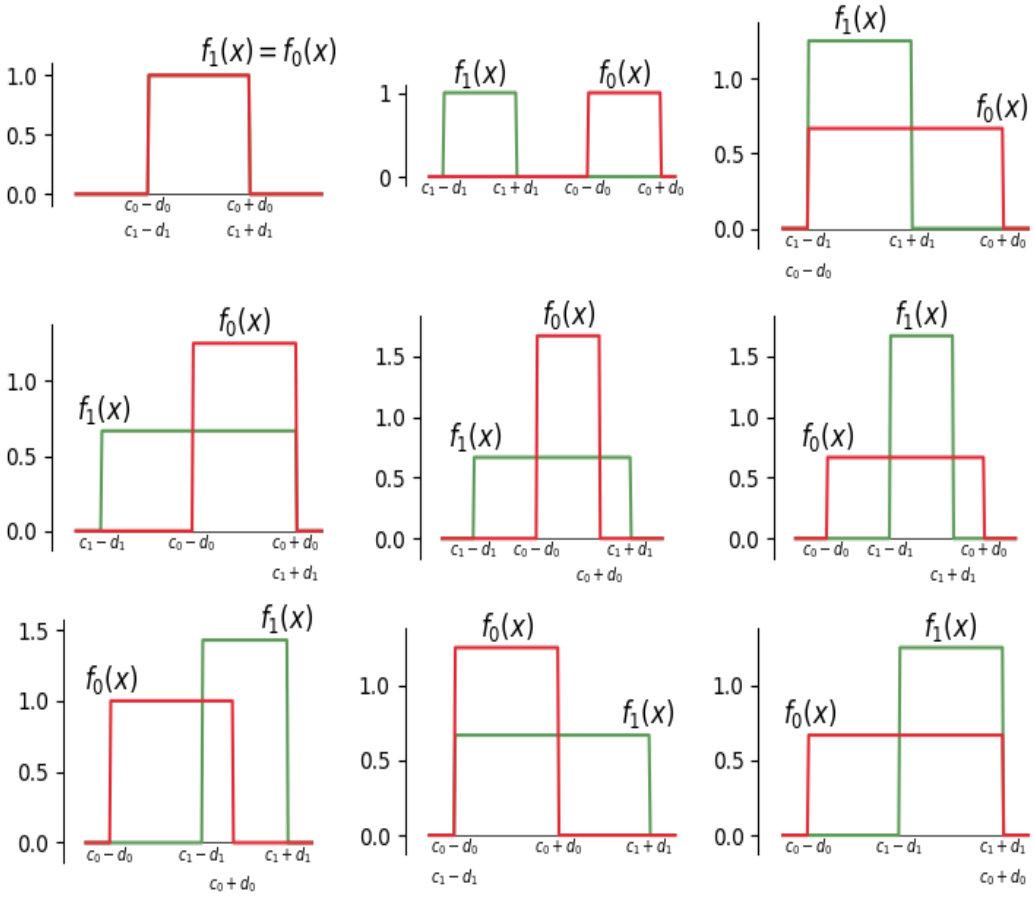
kde poslednú rovnosť sme dostali po ekvivalentných úpravách.

Ako bolo povedané na začiatku tejto podkapitoly, hustoty nemusia byť umiestnené takto, ale môžu nastať rôzne iné prípady, kedy bude ROC krivka pre rovnomerné rozdelenie vyzerať inak. Na nasledujúcich grafoch si ukážeme ďalších 9 možností. Podľa predchádzajúceho postupu si v skratke popíšeme predpis jednotlivých ROC kriviek a výpočet ich plôch pre tieto ďalšie možnosti.

Prvé dva grafy na obrázku 3.5 predstavujú hustoty, z ktorých sú tvary ROC kriviek pre bezvýznamný a perfektný test, ako sme už opisovali v podkapitole 2.4 o vlastnostiach ROC krivky. Ich znázornenie môžeme vidieť na obrázku 3.6. Predpis týchto ROC kriviek a plochou pod nimi je zrejmý.

Nasledujúce dva grafy predstavujú prípad, keď sa zhoduje v daných dvoch rovnomerných rozdeleniach jeden parameter. Pre takto prekrývajúce sa hustoty sa budú ich príslušné ROC krivky skladať z dvoch lineárnych častí, ako vidíme na treťom a štvrtom grafe obrázku na 3.6. V prvom prípade, keď $c_1 - d_1 = c_0 - d_0$ vypočítame prvú časť ROC krivky vzorcom (3.5) pre $t \in (0, \frac{d_1+d_0+\delta c}{2d_0})$ a potom od bodu, kde $t = \frac{d_1+d_0+\delta c}{2d_0}$, bude druhá lineárna časť $ROC(t) = 1$, teda pre $t \in (\frac{d_1+d_0+\delta c}{2d_0}, 1)$. V druhom prípade, keď $c_1 + d_1 = c_0 + d_0$, budeme mať najprv zvislú časť, pretože $FPR(\phi) = 0$, čiže od bodu $(0,0)$ do bodu $(0, TPR(c_0 - d_0))$. Od hodnoty klasifikačného prahu $\phi = c_0 - d_0$ bude druhá časť, opäť vypočítaná vzorcom (3.5) pre $t \in (0,1)$. Plochu pod krivkou v oboch prípadoch vypočítame pomocou odvodeného vzorca 3.7.

Pre ďalšie dve možnosti usporiadania hustôt, sa bude ROC krivka skladať z troch lineárnych častí. V prvej možnosti usporiadania je najprv zvislá časť, pretože $FPR(\phi) = 0$. Táto zvislá časť pôjde od bodu $(0,0)$ do bodu, kde ϕ dosiahne hodnotu $c_0 - d_0$, čiže $(0, TPR(c_0 - d_0))$. Ďalej určíme vzorec pre ROC krivku opäť pomocou vzorca (3.5) pre $t \in (0,1)$. Nakoniec, od bodu, keď $FPR(\phi) = 1$, bude opäť zvislá časť, teda od bodu $(1, TPR(c_0 + d_0))$ do bodu $(1,1)$. V druhej



Obr. 3.5: Rôzne možnosti umiestnenia hustôt rovnomerného rozdelenia

možnosti máme najprv vodorovnú časť $ROC(t) = 0$, z čoho vyjadríme po akú hodnotu t bude daná ROC krivka vodorovná.

$$ROC(t) = 0,$$

$$\frac{d_0}{d_1}t + \frac{\delta d - \delta c}{2d_1} = 0,$$

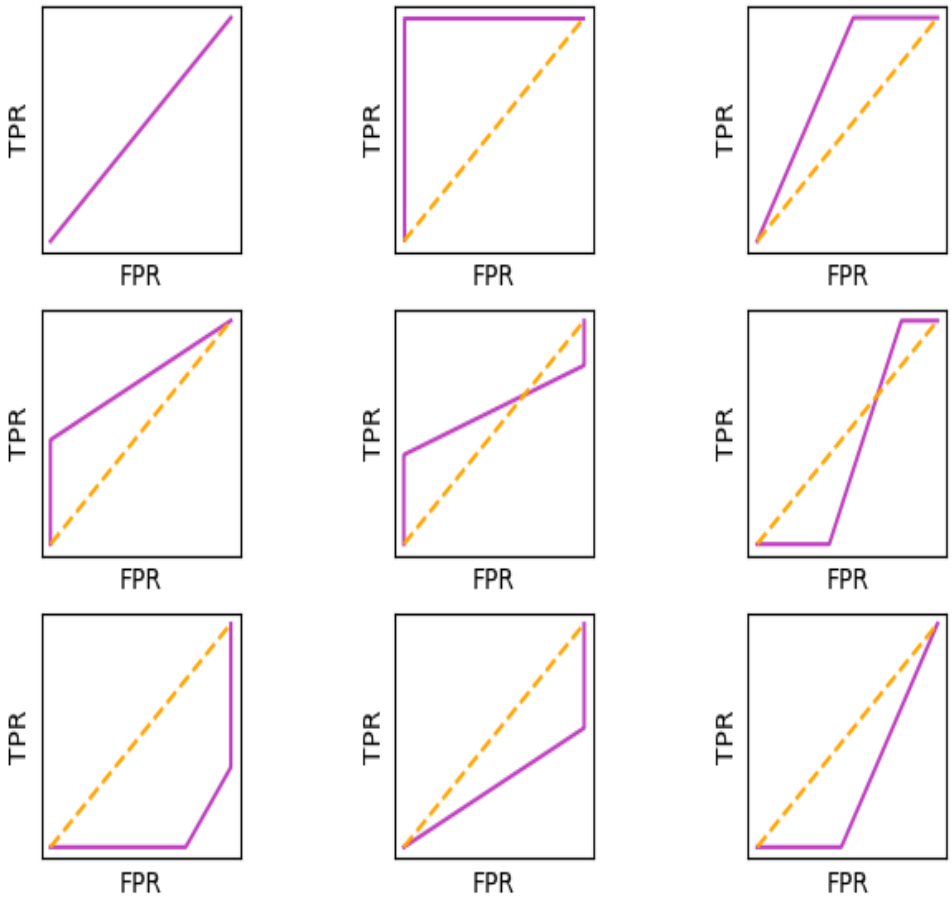
$$t = \frac{\delta c - \delta d}{2d_0}. \tag{3.8}$$

Teda pre $t \in (0, \frac{\delta c - \delta d}{2d_0})$ bude vodorovná časť ROC krivky. Následne, pre $t \in (\frac{\delta c - \delta d}{2d_0}, \frac{d_1 + d_0 + \delta c}{2d_0})$, kde horná hodnota t bola odvodená vyššie ako (3.6), bude predpis ROC krivky podľa (3.5). Ako posledná bude vodorovná časť, čiže $ROC(t) = 1$ pre $t \in (\frac{d_1 + d_0 + \delta c}{2d_0}, 1)$. Plocha pod prvou možnosťou bude vypočítaná pomocou základného vzorca pre výpočet plochy pod ROC krivkou, čiže podľa (3.4). Plochu pod druhou možnosťou odvodíme ako súčet plochy trojuholníka pod danou ROC krivkou a obdĺžnika vedľa neho, čiže

$$AUC = \int_{\frac{\delta c - \delta d}{2d_0}}^{\frac{d_1 + d_0 + \delta c}{2d_0}} ROC(t)\mathrm{d}t + \int_{\frac{d_1 + d_0 + \delta c}{2d_0}}^{1} 1\mathrm{d}t =$$

$$= \int_{\frac{\delta c - \delta d}{2d_0}}^{\frac{d_1 + d_0 + \delta c}{2d_0}} \left(\frac{d_0}{d_1}t + \frac{\delta d - \delta c}{2d_1}\right)\mathrm{d}t + \int_{\frac{d_1 + d_0 + \delta c}{2d_0}}^{1} 1\mathrm{d}t = \frac{d_0 - \delta c}{2d_0}.$$

Poslednú rovnosť sme dostali výpočtom integrálov a následnými ekvivalentnými úpravami.



Obr. 3.6: ROC krivky k daným hustotám z obrázku 3.5

Ako vidíme na obrázku, ďalšie tri grafy ROC krivky predstavujú bezvýznamné testy, keďže sú pod diagonálou, ktorá znázorňuje bezvýznamný test. Veľkosti ich plôch by vyšli menej než 0,5. Preto sa im podrobnejšie venovať nebudeme, ale vyjadrili by sme ich analogicky ako v predchádzajúcich postupoch.

Ako už bolo na začiatku spomenuté, máme jedenásť možností. Najprv sme si podrobne popísali prvú možnosť na grafe 3.4, následne na grafe 3.6 ďalších deväť možností. Nakoniec, je tu ešte jedenásta možnosť, a to keby sa hustoty neprekrývali a boli by vymenené, takže by ROC krivka prechádzala bodom $(1, 0)$ a hodnota AUC by bola 0. Túto možnosť ani nezobrazujeme, z dôvodu nezmyselnosti testu pre danú ROC krivku.

4. Štatistické testovanie a aplikácia na reálne dáta

4.1 Štatistické testovanie

V súvislosti so štatistickým testovaním hovoríme, že keď testovaný jedinec nespĺňa hypotézu H_0 a test H_0 zamieta, tak ho zaraďujeme medzi TP, keď jedinec spĺňa H_0 a test H_0 prijíma, tak ho zaraďujeme medzi TN, keď spĺňa H_0, ale test H_0 zamieta, tak ho radíme medzi FP a keď H_0 nespĺňa, ale test H_0 prijíma, tak ho radíme medzi FN.

Určenie kritického oboru a oboru prijatia sa robí pomocou kritických hodnôt testovacieho kritéria, čo sú konkrétne kvantily príslušných rozdelení súvisiacich so zvolenou hladinou významnosti α.

Pri testovaní štatistických hypotéz je chyba 1.druhu mylné zamietnutie skutočne pravdivej nulovej hypotézy ako výsledok testu, zatiaľ čo chyba 2. druhu je nezamietnutie nulovej hypotézy, ktorá je v skutočnosti nepravdivá ako sme videli v tabuľke 1.1.

Veľká časť štatistickej teórie hovorí o minimalizácii jednej alebo oboch týchto chýb, hoci úplné odstránenie ktorejkoľvek z nich je štatisticky nemožné, ak výsledok nie je určený známym procesom. Výberom nízkej prahovej hodnoty a úpravou hladiny α možno zvýšiť kvalitu testu hypotézy.

Hladina testu α je pravdepodobnosť chyby 1. druhu, čiže pravdepodobnosť zamietnutia platnej hypotézy a pravdepodobnosť chyby 2. druhu je $1 - \beta$, kde β je sila testu.

Test je navrhnutý tak, aby udržal chybu 1. druhu pod vopred určenou hranicou, čiže hladinou významnosti α. Zvyčajne je hladina významnosti nastavená na 0,05, čo znamená, že je prijateľné mať 5% pravdepodobnosť nesprávneho zamietnutia pravdivej hypotézy H_0.

Tieto dva typy chybovosti spolu vzájomne súvisia, čiže pre každý daný súbor vzoriek vedie snaha znížiť jeden typ chyby k zvýšeniu druhého typu chyby. Rôzne prahové hodnoty sa môžu použiť na to, aby bol test špecifickejší alebo senzitívnejší, čo zvyšuje kvalitu testu. Zvyšovaním špecificity testu sa znižuje pravdepodobnosť chyby 1. druhu (FPR), čiže α. Naopak zvyšovaním senzitivity sa znižuje pravdepodobnosť chyby 2. druhu (FNR).

Hypotéza a alternatíva pre diagnostický test môže byť napríklad:

$$H_0 : \text{Ženy nemajú rakovinu prsníka.}$$

$$H_1 : \text{Ženy majú rakovinu prsníka.}$$

Potom chyby budú:

Chyba 1. druhu: Pravdou je, že ženy nemajú rakovinu prsníka, ale podľa údajov usudzujeme, že majú.

Chyba 2. druhu: Pravdou je, že ženy majú rakovinu prsníka, ale podľa údajov sa domnievame, že nemajú.

4.2 Empirický odhad ROC krivky

Empirický odhad ROC krivky môžeme opísať ako aplikáciu ROC krivky na pozorované dáta. Budeme používať vzorce z knihy (Pepe (2003)) prepísané s naším značením z tabuľky 1.4. Každý bod na grafe je generovaný rôznymi hodnotami klasifikačného prahu ϕ. Spojením týchto bodov vytvoríme empirickú ROC krivku. Preto pre každý možný klasifikačný prah ϕ sa počíta $\widehat{FPR}(\phi)$ a $\widehat{TPR}(\phi)$ ako

$$\widehat{FPR}(\phi) = \sum_{j=1}^{n_{+0}} \frac{\mathbb{1}[\tilde{X}_{0_j} \leq \phi]}{n_{+0}}, \tag{4.1}$$

$$\widehat{TPR}(\phi) = \sum_{i=1}^{n_{1+}} \frac{\mathbb{1}[\tilde{X}_{1_i} \leq \phi]}{n_{1+}}. \tag{4.2}$$

Tieto hodnoty vieme zapísať aj ako v podkapitole 1.5.

Empirická ROC krivka je potom vykreslenie $\widehat{TPR}(\phi)$ proti $\widehat{FPR}(\phi)$ pre každú možnú hodnotu $\phi \in (-\infty, \infty)$. Ekvivalentne vieme zapísať empirickú ROC krivku ako

$$\widehat{ROC}(t) = \widehat{F_1}(\widehat{F_0}^{-1}(t)),$$

kde $\widehat{F_0}$ a $\widehat{F_1}$ sú empirické distribučné funkcie pre veličiny $\tilde{X}_0$ a $\tilde{X}_1$.

Technicky je empirická ROC krivka diskrétna funkcia, lebo $\widehat{FPR}(\phi)$ môže nadobúdať iba hodnoty $\{0, \frac{1}{n_{+0}}, \frac{2}{n_{+0}}, ..., 1\}$ a tieto hodnoty spájame lineárne. Funkcia teda bude mať vertikálne skoky veľkosti $\frac{1}{n_{1+}}$ a horizontálne skoky veľkosti $\frac{1}{n_{+0}}$.

Empirická ROC krivka je funkcia zoradenia daných dát, čiže závisí na relatívnom zoradení výsledkov testu a ich zaradenia, či sú pozitívni alebo negatívni. Preto má aj empirická ROC krivka vlastnosť, že je invariantná k striktne rastúcej transformácií dát, čo je jedna z vlastností ROC krivky z podkapitoly 2.4.

Plocha pod empirickou ROC krivkou sa počíta ako

$$\widehat{AUC} = \sum_{j=1}^{n_{+0}} \sum_{i=1}^{n_{1+}} \frac{\mathbb{1}[\tilde{X}_{1_i} > \tilde{X}_{0_j}] + \frac{1}{2}\mathbb{1}[\tilde{X}_{1_i} = \tilde{X}_{0_j}]}{n_{+0} n_{1+}}. \tag{4.3}$$

4.3 Spracovanie dát

V našej práci budeme postupovať pomocou logistickej regresie. Pre klasifikáciu získaných dát pripravíme testovaciu množinu a trénovaciu množinu náhodným rozdelením daných dát. Klasifikácia prebieha v dvoch fázach. Najprv prebieha trénovacia fáza na trénovacej množine, kde sa daný klasifikátor natrénuje, čiže si vytvorí klasifikačný model. Cieľom je, aby sa čo najlepšie naučil klasifikovať dáta.

Keď už máme vytvorený klasifikačný model prejdeme na fázu testovania, kde klasifikátor dostane hodnoty testovacej množiny a ďalej bude predikovať hodnoty podľa toho, ako sa to naučil na trénovacej množine. Čiže dostane na vstup spracované dáta, podľa ktorých určí, či daný človek má skúmanú chorobu alebo nie. Porovnaním reálnych výsledkov pre testovaciu množinu a tých, čo vyšli vyhodnotením pomocou klasifikátora zistíme, ako veľmi je spoľahlivý a či by sme ním mohli testovať aj naďalej alebo by to bolo nedôveryhodné.

Dobré znázornenie porovnania týchto výsledkov sa robí pomocou vykreslenia ROC krivky. Zobrazíme ROC krivku a následne posudzujeme spoľahlivosť podľa veľkosti plochy AUC pod ROC krivkou.

Hodnotíme presnosť testu podľa nasledujúcej tabuľky z odporúčanej literatúry dostupnej na internete z prednášky ROC křivka:

AUC	Hodnotenie testu
0,50 až 0,60	nedostatočný
0,60 až 0,70	dostatočný
0,70 až 0,80	dobrý
0,80 až 0,90	veľmi dobrý
0,90 až 1,00	výborný

Tabuľka 4.1: Približné hodnotenie kvality testu

4.4 Aplikácia na reálne dáta

Dáta budeme brať z internetovej stránky Dáta z mamografie a spracovávať ich budeme v programovacom jazyku Python (viď. A.1). Tento dátový súbor obsahuje údaje o 11183 ženách, kde v každom riadku je 6 nameraných údajov o danej žene a jej skutočný stav, či má rakovinu prsníka alebo nie. Keď je v dátach na konci riadku 0, značí to, že daná žena rakovinu nemá, čiže je negatívna a keď je na konci riadku 1, tak že rakovinu má, čiže je pozitívna.

-0.78441482	-0.47019533	-0.59163147	-0.85955255	-0.37786573	-0.94572324	0
0.15193407	-0.2136264	-0.0055711987	0.42315815	-0.37786573	1.1019319	0
-0.78441482	-0.47019533	-0.59163147	-0.85955255	-0.37786573	-0.94572324	0
0.18530251	0.94535738	0.08459192	0.8379294	-0.37786573	0.54126442	0
-0.012536761	-0.42153571	0.40016284	1.2181012	1.0804627	1.6900233	1
1.2417439	4.2055522	-0.32114211	1.3657817	1.8293491	0.529076	1
0.75425981	0.016400912	0.08459192	4.9421817	1.4216019	1.2512401	1

Tabuľka 4.2: Ukážka reálnych dát

Súbor by mal obsahovať 98% negatívnych, čiže všetky ich údaje by mali byť v norme a 2% pozitívnych, čiže dané údaje sú mimo normy.

Nameraných 6 údajov označím ako Y a klasifikáciu medzi 0 a 1 ako D zo značenia v predchádzajúcich kapitolách.

Tieto dáta software náhodne rozdelí na trénovacie a testovacie. Následne natrénujeme model logistickej regresie na tréningovej skupine dát a spravíme predpovede pre testovacie dáta, čiže ako nám vyšla predikovaná hodnota zaradenia X pre testovací Y.

Budeme testovať hypotézu

$$H_0 : \text{žena je negatívna}$$

proti alternatíve

$$H_1 : \text{žena je pozitívna.}$$

Ak test nezamietne hypotézu H_0, tak predikovaná hodnota X bude 0, ak zamietne H_0, tak predikovaná hodnota X bude 1.

Porovnaním reálnych hodnôt D, ktoré boli zaradené ako testovacie, s predikovanými hodnotami X spravíme kontingenčnú tabuľku pre počty skutočne pozitívnych, skutočne negatívnych, nesprávne pozitívnych a nesprávne negatívnych pri zvolenom klasifikačnom prahu $\phi = 0{,}5$:

	Skutočná trieda	
Klasifikovaná trieda	**D=0**	**D=1**
$X = 0$	2724	41
$X = 1$	8	23

Tabuľka 4.3: Kontingenčná tabuľka pre $\phi = 0{,}5$

S presnosťou 0,9824749642346209 a chybovosťou 0,017525035765379112 nám vyšlo 2732 negatívnych a 64 pozitívnych.

Nesprávne pozitívnych vyšlo 8 a nesprávne negatívnych vyšlo 41.

Odhad senzitivity zo zistených dát spočítame ako

$$\widehat{TPR}(\phi) = TP/(TP + FN) = 23/(23 + 41) \doteq 0{,}359375$$

a odhad špecificity ako

$$\widehat{TNR}(\phi) = TN/(FP + TN) = 2724/(8 + 2724) \doteq 0{,}997071742.$$

Odhad pravdepodobnosti chyby 1. druhu, čiže $\widehat{FPR}(\phi)$, spočítame ako

$$\widehat{FPR}(\phi) = FP/(FP + TN) = 8/(8 + 2724) \doteq 0{,}002928258$$

a odhad pravdepodobnosti chyby 2. druhu, čiže $\widehat{FNR}(\phi)$, ako

$$\widehat{FNR}(\phi) = FN/(FN + TP) = 41/(41 + 23) \doteq 0{,}640625.$$

Pravdepodobnosť chyby 1. druhu nám vyšla nízka, takže táto voľba klasifikačného prahu sa dá považovať za dobrú.

Znížime hodnotu klasifikačného prahu na $\phi = 0{,}3$, aby sme zvýšili senzitivitu. Kontingenčná tabuľka potom bude:

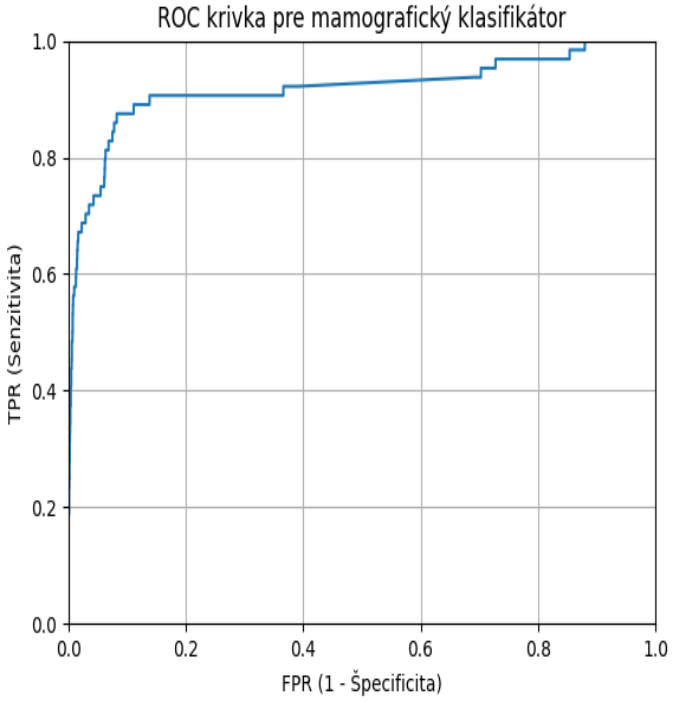
	Skutočná trieda	
Klasifikovaná trieda	**D=0**	**D=1**
$X = 0$	2711	29
$X = 1$	21	35

Tabuľka 4.4: Kontingenčná tabuľka pre $\phi = 0{,}3$

Odhad senzitivity nám teraz vyšiel $\widehat{TPR}(\phi) \doteq 0{,}625$ a odhad špecificity vyšiel ako $\widehat{TNR}(\phi) \doteq 0{,}992313323$. Pravdepodobnosť chyby 1. druhu vyšla ako $\widehat{FPR}(\phi) \doteq 0{,}007686676$ a pravdepodobnosť chyby 2. druhu ako $\widehat{FNR}(\phi) \doteq$

0,453125. Klasifikačný prah teda môžeme meniť, aby sme zvýšili senzitivitu alebo špecificitu alebo aby sme dostali lepšie hodnoty chýb 1. a 2. druhu.

Teraz si zo získaných dát zostrojíme empirickú ROC krivku. Ako sme už opisovali, zostrojíme ju vypočítaním hodnôt $\widehat{FPR}(\phi)$ a $\widehat{TPR}(\phi)$ pre každú možnú hodnotu klasifikačného prahu ϕ a ich zakreslením na graf $\widehat{FPR}(\phi)$ proti $\widehat{TPR}(\phi)$. Tieto hodnoty lineárne pospájame a dostaneme empirickú ROC krivku. Následne dopočítame hodnotu plochy pod krivkou.



Obr. 4.1: Empirická ROC krivka pre dáta z mamografie

Obsah plochy pod empirickou ROC krivkou nám vyšiel podľa vzorca (4.3) ako $\widehat{AUC} = 0{,}91818$, čiže daný klasifikátor sa podľa tabuľky 4.1 môže považovať za výborný a mohli by sme ním posudzovať testy aj naďalej.

Záver

Cieľom práce bolo preštudovať dostupnú literatúru o ROC krivkách a zistené poznatky spísať, vysvetliť pojem ROC krivky a popísať jej vlastnosti. Najprv teda, v prvej kapitole, boli zadefinované používané pojmy. Následne bola ROC krivka popísaná v druhej kapitole aj s jej vlastnosťami. V práci je zahrnutá aj jej prvá a druhá derivácia, vďaka ktorým bola popísaná konkavita ROC krivky.

Ďalej mala byť ROC krivka ukázaná na rôznych pravdepodobnostných rozdeleniach. Toto bolo popísané v tretej kapitole, kde sa pre normálne, exponenciálne a rovnomerné rozdelenie vyjadril vzorec pre ROC krivku aj jej plochu AUC. Ku každému rozdeleniu bolo aj grafické znázornenie jeho hustôt a príslušnej ROC krivky. Pre rovnomerné rozdelenie boli popísané všetky možnosti, ktoré môžu nastať pri rôznom spôsobe prekrývania daných hustôt. Tieto možnosti boli v práci taktiež graficky znázornené.

Na záver boli ROC krivky uvedené do súladu s teóriou štatistického testovania a aplikovanie ROC krivky na dátach, čo zahŕňalo definovanie empirického vyjadrenia ROC krivky a jej plochy. Ukázané bolo využitie ROC krivky na ilustratívnom príklade s reálnymi dátami z mamografie. Dáta boli spracované v programovacom jazyku Python a z grafu bolo vidieť, že daný test bol spoľahlivý.

Zoznam použitej literatúry

Dáta z mamografie. (Date of the latest version: 21.8.2020; Accessed: 20.3.2023; Bachelor Thesis). URL `https://raw.githubusercontent.com/jbrownlee/Datasets/master/mammography.csv`.

Anděl, J. (1985). *Matematická statistika*. Druhé vydání. SNTL - Nakladatelství technické literatury, Praha.

Anděl, J. (2007). *Základy matematické statistiky*. Druhé opravené vydání. Matfyzpress, Praha. ISBN 80-7378-001-1.

Kulich, M. (2014). *Přehledový větník pro předmět statistika pro finanční matematiky*. Matematicko-fyzikální fakulta Univerzity Karlovy, Praha.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press, New York. ISBN 978-0-19-850984-4.

ROC křivka (2009). URL `https://is.muni.cz/el/sci/jaro2009/MAS02/um/7421238/Prednaska_c._13.pdf`.

Zhou, X., Obuchowski, N. A. a McClish, D. K. (2002). *Statistical Methods in Diagnostic Medicine*. John Wiley and Sons, New York. ISBN 0-471-34772-8.

Zoznam obrázkov

Zoznam tabuliek

A. Prílohy

A.1 Prvá príloha

Elektronická príloha obsahuje spracovanie reálnych dát v programovacom jazyku Python. Opis dát je v štvrtej kapitole tejto bakalárskej práce.