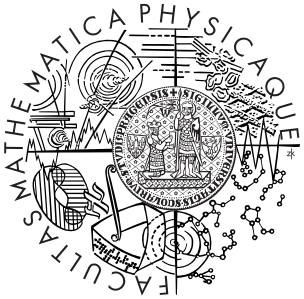
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Lucia Fuchsová

Charakteristiky pravděpodobnostních předpovědí

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: Mgr. Zdeněk Hlávka, Ph.D.
Studijní program: matematika

2009

Rada by som sa poďakovala môjmu vedúcemu pánovi Mgr. Zdeňkovi Hlávkovi, Ph.D. za jeho pomoc, vedenie, ochotu, trpezlivosť a za všetky užitočné rady. Chcela by som sa poďakovať svojim rodičom a bratovi Lukášovi za podporu pri štúdiu a písaní tejto práce. Ďakujem kamarátovi Johnymu za požičanie knihy o programe LaTeX.

Prehlasujem, že som svoju bakalársku prácu napísala samostatne a výhradne s použitím citovaných prameňov. Súhlasím so zapožičiavaním práce a jej zverejňovaním.

V Prahe dňa 27. mája 2009 Lucia Fuchsová

Obsah

Název práce: Charakteristiky pravděpodobnostních předpovědí
Autor: Lucia Fuchsová
Katedra (ústav): Katedra pravděpodobnosti a matematické statistiky
Vedoucí bakalářské práce: Mgr. Zdeněk Hlávka, Ph.D.
e-mail vedoucího: hlavka@karlin.mff.cuni.cz

Abstrakt: V predloženej práci študujeme rozklad združeného rozdelenia predpovedí a pozorovaní. Popisujeme charakteristiky kvality predpovedí a ich miery. Zaoberáme sa rozkladmi strednej štvorcovej chyby a skóra úspešnosti. Ďalej sa venujeme modelom predpovedí založených na pozorovaniach zo známych pravdepodobnostných rozdelení (pre rovnomerné, normálne a beta rozdelenia). Pre tieto rozdelenia definujeme ROC krivku. Určíme združené rozdelenie pravdepodobnostných predpovedí a pozorovaní a študujeme závislosť vybraných charakteristík na rozdiele stredných hodnôt.

Klíčová slova: združené rozdelenie predpovedí a pozorovaní, charakteristiky kvality predpovede, ROC krivka

Title: Characteristics of probability forecasts
Author: Lucia Fuchsová
Department: Department of Probability and Mathematical Statistics
Supervisor: Mgr. Zdeněk Hlávka, Ph.D.
Supervisor's e-mail address: hlavka@karlin.mff.cuni.cz

Abstract: In the present work we study factorization of the joint distribution of forecasts and observations. We describe the characteristics of the forecast quality and their measures. We analyse the mean square error and the skill scores. Next we present the model, where forecasts are based on the observations from known distributions (for uniform, Gaussian and beta distributions). For these distributions we define the ROC curve. We derive the joint distribution of probabilistic forecasts and observations and we study the dependency of some characteristics on the difference of means.

Keywords: joint distribution of forecasts and observations, characteristics of forecast quality, ROC curve

Kapitola 1

Úvod

Overovanie predpovedí má dlhú a pestrú históriu. Prvý mohutný rozvoj metód overovania predpovedí nastal po uverejnení práce J. P. Finleyho v roku 1884. V tejto práci boli zhrnuté výsledky programu predpovedania výskytu tornád. Informácie v tejto kapitole čerpáme z [1].

Zvyčajne sa overovanie pravdepodobnostných predpovedí definuje ako stanovenie stupňa súladu predpovedí a im prislúchajúcich pozorovaní. V praxi overovací proces pozostáva z výpočtu veľkosti jednej alebo dvoch charakteristík predpovedí ako je vychýlenie, presnosť alebo úspešnosť a zostrojenia záverov s prihliadnutím na absolútnu alebo relatívnu charakteristiku na základe numerických hodnôt týchto meraní. Tieto tradičné metódy predstavujú MO (measure oriented) prístup.

V našej práci je overovanie predpovedí definované ako proces stanovenia kvality predpovede, kde kvalitou predpovede rozumieme pozostávanie zo súhrnu štatistických charakteristík popísaných v združenom rozdelení predpovedí a pozorovaní. Táto definícia vedie k DO (distribution oriented) prístupu. DO prístup ponúka teoretické a praktické výhody oproti tradičnému MO prístupu.

DO prístupu je venovaná druhá kapitola, kde sú popísané charakteristiky kvality predpovede a ich miery.

1.1 Označenie

Problém overovania predpovedí zahŕňa ohodnotenie a porovnanie predpovedných systémov. Predpovedným systémom môže byť model alebo metóda, ktorými sa určí predpoveď. My budeme predpovedným systémom rozu-

mieť náhodnú veličinu, ktorou budeme popisovať predpovede. Túto náhodnú veličinu budeme značiť veľkým písmenom F. Predpovede, realizácie náhodnej veličiny F označujeme malými písmenami f, g. Na ohodnotenie a porovnanie predpovedných systémov musia byť dostupné pozorovania. Predpokladáme, že pozorovania sú reprezentované náhodnými veličinami X, Y, ktoré popisujú možné situácie. Konkrétne pozorovania, realizácie náhodných veličín X, Y sú označené malými písmenami x, y.

Veličiny, ktorých sa týkajú predpovede aj pozorovania, môžu byť rôzneho druhu. Niektoré veličiny sú spojité, iné sú diskrétne.

Predpokladáme, že pozorovania x, y môžu nadobúdať iba hodnoty z množiny $\{0, 1\}$. Samotné predpovede môžu byť vyjadrené v pravdepodobnostnej alebo nepravdepodobnostnej forme. Pravdepodobnostné predpovede ukazujú pravdepodobnosť javu. Budeme sa nimi zaoberať v kapitolách 3.3 a 3.5. Nepravdepodobnostné predpovede špecifikujú, či daný jav nastane alebo nenastane. V prípade nepravdepodobnostných predpovedí je množina možných predpovedí a pozorovaní spravidla identická. Ak sú predpovede v pravdepodobnostnej forme, potom je zvyčajne množina jednoznačných predpovedí rozsiahlejšia ako množina pozorovaní, pretože pravdepodobnostné predpovede nadobúdajú hodnoty z intervalu $(0, 1)$, zatiaľčo pozorovania majú hodnotu 0 alebo 1.

Overovaná vzorka, ktorá prislúcha náhodným veličinám F a X, je označená ako systém n dvojíc predpovedí a pozorovaní $\{(f_k, x_k), k = 1...n\}$, pričom nie je vylúčené opakovanie hodnôt. Nech n_f označuje počet možných realizácií náhodnej veličiny F a n_x počet možných realizácií náhodnej veličiny X, v našom prípade je n_x rovné 2. Združené rozdelenie predpovedí a pozorovaní označíme $p(F, X)$. Združené rozdelenie obsahuje informácie o predpovediach, pozorovaniach a vzťahu medzi nimi.

Pre jednoduchosť náhodnú veličinu F "zaokrúhlime" a budeme považovať za diskrétnu. Nech p_{ij} označuje združenú relatívnu četnosť predpovedí $f^{(i)}$ a pozorovaní $x^{(j)}$ v overovanej vzorke $(i = 0, ..., n_f - 1,\ j = 0, ..., n_x - 1)$. Matica $P = (p_{ij})_{i=0,...,n_f-1, j=0,...,n_x-1}$ obsahuje združené relatívne četnosti všetkých dvojíc $(f^{(i)}, x^{(j)})$. My budeme považovať združené relatívne četnosti za združené pravdepodobnosti teda $p_{ij} = P(F = f^{(i)}, X = x^{(j)}) = P(F_k = f^{(i)}, X_k = x^{(j)}), k = 1...n$ potom matica P obsahuje združené pravdepodobnosti všetkých dvojíc $(f^{(i)}, x^{(j)})$. Tieto pravdepodobnosti môžeme chápať ako jednoduchý model združeného rozdelenia $p(F, X)$.

Kapitola 2

Charakteristiky kvality predpovedí

Táto kapitola vychádza z [1], z prác [2], [3] a z bakalárskej práce [4]. Venujeme sa v nej rozkladom združeného rozdelenia predpovedí a pozorovaní na podmienené a marginálne rozdelenia. Popisujeme charakteristiky kvality predpovedí a ich miery. Zaoberáme sa vzťahmi charakteristík a rozkladmi strednej štvorcovej chyby a skóra úspešnosti.

2.1 Rozklad združeného rozdelenia

Združené rozdelenie $p(F, X)$ hrá základnú úlohu v procese overovania. V podstate, overovanie predpovedí pozostáva z opisu a zhurnutia štatistických charakteristík združeného rozdelenia. Hoci združené rozdelenie predpovedí a pozorovaní obsahuje všetky podstatné informácie o kvalite predpovede, rozkladom na marginálne a podmienené rozdelenie sprístupníme tieto informácie. Definujeme dve faktorizácie (rozklady)

$$p(F, X) = p(X|F)p(F) \tag{2.1}$$

$$p(F, X) = p(F|X)p(X) \tag{2.2}$$

Rozklad (2.1) sa nazýva CR (calibration - refinement) rozklad a zahŕňa podmienené rozdelenie pozorovaní za podmienky predpovede $F = f$ teda $p(X|F)$ a marginálne rozdelenie predpovedí, $p(F)$. Podmienené rozdelenie prepovedí je definované pre všetky možné predpovede takto $p(X|F) = P(X =$

$x^{(j)}|F = f^{(i)}$). Marginálne rozdelenie $p(F)$ určuje nepodmienené pravdepodobnosti predpovedí t.j. $p(F) = P(F = f^{(i)}) = \sum_{j=1}^{n_x} p_{ij}$.

Rozklad (2.2) nazývaný LBR (likelihood - base rate) zahŕňa podmienené rozdelenie predpovedí za podmienky pozorovania $X = x$, $p(F|X)$ a marginálne rozdelenie pozorovaní, $p(X)$. Podmienené rozdelenie $p(F|X)$ je definované pre všetky možné pozorovania takto $p(F|X) = P(F = f^{(i)}|X = x^{(i)})$. Marginálne rozdelenie $p(X)$ určuje nepodmienené pravdepodobnosti pozorovaní t.j. $p(X) = P(X = x^{(j)}) = \sum_{i=1}^{n_f} p_{ij}$.

2.2 Popis charakteristík kvality predpovedí

Kvalita predpovedí je definovaná ako súhrn štatistických charakteristík predpovedí, pozorovaní a ich vzťahu, ktorý je založený na združenom rozdelení $p(F, X)$. Charakteristiky kvality predpovedí sa môžu vzťahovať k združenému rozdeleniu alebo k podmienenému a marginálnemu rozdeleniu, ktoré je spojené s rozkladom združeného rozdelenia.

Vychýlenie (bias, systematické alebo nepodmienené vychýlenie) sa vzťahuje k stupňu zhody medzi strednou hodnotou predpovedí μ_X a strednou hodnotou pozorovaní μ_F. Obvykle je vychýlenie definované ako rozdiel medzi μ_X a μ_F.

Združenie (association) zachytáva silu lineárneho vzťahu predpovedí a pozorovaní. Tento vzťah popisuje korelačný koeficient $\rho_{F,X}$.

Správnosť (accuracy) popisuje priemerný stupeň súladu predpovedí a pozorovaní jednotlivo v overovanej vzorke. Obvykle je definovaná pomocou združeného rozdelenia $p(F, X)$. Ale môže byť definovaná aj pomocou podmieneného a marginálneho rozdelenia. V súvislosti s overovaním predpovedí sa správnosť meria pomocou strednej štvorcovej chyby (MSE) alebo strednej absolutnej chyby (MAE).

Úspešnosť (skill) je definovaná ako správnosť predpovedí, ktoré nás zaujímajú, relatívne ku správnosti predpovedí, ktoré boli vytvorené primitívnym predpovedným systémom. Úspešnosť odhaduje skóre úspešnosti (skill score), ktoré je definované ako miera relatívnej správnosti. Kladné (záporné) skóre úspešnosti značí, že správnosť predpovedí, ktoré nás zaujímajú, je vyššia (nižšia) ako správnosť predpovedí vytvorených primitívnym systémom.

Spoľahlivosť (reliability, calibration, conditional bias type 1) charakterizuje stupeň zhody strednej hodnoty podmieneného rozdelenia X za podmienky $F = f$ teda $\mu_{X|F}$ a predpovede F. Uprednostňujeme malé rozdiely $\mu_{X|F}$ a F. Predpovede, ktoré vykazujú dokonalú zhodu $\mu_{X|F}$ a F vo všetkých hodnotách $F = f$ sa nazývajú úplne spoľahlivé (dobre kalibrované alebo podmienene nevychýlené). Predpovede, ktoré sú úplne spoľahlivé sú nevyhnutne nevychýlené (opak všeobecne neplatí).

Rozlíšenie (resolution) vystihuje rozdiel strednej hodnoty podmieneného rozdelenia X za podmienky $F = f$, $\mu_{X|F}$ a strednej hodnoty pozorovaní μ_X. Overovaná vzorka, pre ktorú platí $\mu_{X|F} = \mu_X$ pre všetky hodnoty f náhodnej veličiny F, je úplne bez rozlíšenia. Väčší rozdiel $\mu_{X|F}$ a μ_X je výhodnejší. Rozlíšenie ako charakteristika kvality predpovede je založená na koncepcii, že po rôznych predpovediach by mali nasledovať rôzne pozorovania.

Ostrosť (sharpness, refinement) je charakteristika, ktorá používa iba pravdepodobnostné predpovede. Predpovede sú perfektne ostré, ak je v nich používaná iba hodnota nula a jedna. Na druhej strane konštantná predpoveď pravdepodobnosti je úplne neostrá. Ak sú predpovede, ktoré nás zaujímajú úplne spoľahlivé, potom ostrosť a rozlíšenie sú identické charakteristiky.

Podmienené vychýlenie typu 2 (coditional bias type 2) popisuje stupeň zhody strednej hodnoty podmieneného rozdelenia F za podmienky $X = x$, teda $\mu_{F|X}$ a pozorovania X. Predpovede, ktoré vykazujú úplnú zhodu $\mu_{F|X}$ a X ($\mu_{F|X} = X$) pre všetky x sú úplne podmienene nevychýlené (pravdepodobnostné predpovede, ktoré vyhovujú tejto podmienke, sú dokonalé predpovede).

Diskriminácia (discrimination) charakterizuje rozdiel strednej hodnoty podmieneného rozdelenia F za podmienky $X = x$, $\mu_{F|X}$ a strednej hodnoty predpovedi μ_F. Ak $\mu_{F|X} = \mu_F$ pre každé x, potom overovaná vzorka je úplne nediskriminovaná. Výhodnejší je väčší rozdiel $\mu_{F|X}$ a μ_F. Diskrimináciu by sme mohli neformálne interpretovať ako schopnosť rozlišovať medzi pozorovaniami. Nízka diskriminácia by predstavovala situáciu, v ktorej sa podmienené rozdelenia $p(F|X = 1)$ a $p(F|X = 0)$ prevažne zhodujú.

Neistota (uncertainty) popisuje rozptýlenosť pozorovaní (ako primitívny popis predpovedanej situácie). Ako miera neistoty sa používa rozptyl pozorovaní σ_X^2. Situácia, v ktorej sú udalosti približne rovnako pravdepodobné, značí relatívne vysokú hodnotu neistoty, zatiaľ čo situácia, v ktorej prevláda jedna alebo dve udalosti, naznačuje relatívne nízku hodnotu neistoty. Hoci neistota vôbec nezávisí na predpovediach, je charakteristikou pozorovaní, môže mať jej hodnota značný vplyv na iné charakteristiky predpovedí.

2.3 Spoľahlivý a ostrý predpovedný systém

CR rozkladom združeného rozdelenia dostaneme podmienené rozdelenie pozorovaní za podmienky predpovede $p(X|F)$ a marginálne rozdelenie predpovedí $p(F)$. Podmienené rozdelenie pozorovaní $p(F|X)$ súvisí so spoľahlivosťou predpovedí. Ideálne by bolo, keby sme mali $P(X = 1|F = f) = f$. Ak je táto podmienka splnená, potom predpovedný systém je dokonale spoľahlivý. Marginálne rozdelenie predpvedí $p(F)$ udáva ako často sú používané rôzne hodnoty predpovedí. Súvisí s ostrosťou predpovedí. Konštantná prepoveď je úplne neostrá, nula-jedničkové predpovede sú perfektne ostré.

Môžeme mať predpovedný systém, ktorý je perfektne spoľahlivý ale nevykazuje žiadnu ostrosť. Alebo môžeme mať predpovedný systém, ktorý je ostrý ale nie je vôbec spoľahlivý. Avšak ani jeden z týchto dvoch extrémnych prípadov nie je veľmi užitočný. Uprednostňujeme predpovedné systémy, ktoré majú obe vlastnosti, sú vysoko spoľahlivé a dosť ostré. Inak povedané chceli by sme predpovedný systém, ktorý by bol tak ostrý ako je len možné bez toho aby sme sa vzdali jeho spoľahlivosti. Udržanie spoľahlivosti znamená, že môžeme používať predpovede a lepšia ostrosť znamená, že predpovede rozlišujú efektívne medzi situáciami vedúcimi k rôznym pozorovaniam. Ak je predpovedný systém prefektne ostrý a spoľahlivý, potom pravdepodobnosti $P(F = f|X = 1)$ a $P(F = f|X = 0)$ musia byť dokonale diskriminujúce, opak však všeobecne neplatí.

2.4 Miery charakteristík kvality predpovedí

V tejto časti definujeme základné kvantitatívne miery rôznych charakteristík kvality predpovede ako sú stredná chyba (ME), stredná štvorcová chyba (MSE) a skóre úspešnosti (SS).

Stredná chyba (ME) predpovedí je definovaná ako rozdiel strednej hodnoty predpovedí a strednej hodnoty pozorovaní

$$ME(F, X) = \mu_F - \mu_X$$

Stredná chyba je mierou nepodmieneného vychýlenia. Kladné (záporné) hodnoty ME svedčia o tom, že predpovede sú nadhodnotené (podhodnotené). Ak je $ME(F, X) = 0$, predpovede sú nepodmienene nevychýlené.

Strednú štvorcovú chybu (MSE) definujeme ako

$$MSE(F, X) = \sum_f \sum_x P(F = f, X = x)(f - x)^2$$

MSE je mierou správnosti a jej hodnoty sú nezáporné (nulová je len v prípade, ak $p(F, X) = 0$ pre všetky hodnoty náhodnej veličiny F rôzne od hodnôt náhodnej veličiny X). Vyššiu správnosť majú predpovede, ktoré vykazujú malú hodnotu MSE. Ako vidíme z definície, MSE nerozlišuje pravdepodobnostné a nepravdepodobnostné predpovede.

Skóre úspešnosti (SS) je zvyčajne definované ako zlepšenie správnosti predpovedí, ktoré nás zaujímajú, oproti správností predpovedí vyrobených primitívnym predpovedným systémom. Ak vezmeme MSE za mieru správnosti, potom SS môže byť definované takto

$$SS(F, \mu_X, X) = 1 - [MSE(F, X)/MSE(R, X)]$$

kde $MSE(R, X)$ je MSE predpovedí R založených na porovnácacom štandarde. Za $MSE(R, X)$ môžeme voliť MSE konštantnej predpovede s hodnotou μ_X, čo je stredná hodnota pozorovaní, potom $MSE(\mu_X, X)$ sa rovná rozptylu pozorovaní σ_X^2

$$SS(F, \mu_X, X) = 1 - [MSE(F, X)/\sigma_X^2]$$

Podľa tohto predpokladu je úspešnosť kladná (záporná), keď MSE predpovedí je menšia (väčšia) ako rozptyl pozorovaní. Pretože $MSE(F, X) = 0$ keď $p(F, X) = 0$ pre všetky F rôzne od X, potom pre dokonalé predpovede je hodnota SS rovná jednej.

2.5 Rozklad MSE a SS

MSE môže byť rozložená niekoľkými spôsobmi, členy rozkladu popisujú kvantitatívne miery rozličných charakteristík pravdepodobnostných predpovedí. Základný rozklad MSE je

$$MSE(F, X) = (\mu_F - \mu_X)^2 + \sigma_F^2 + \sigma_X^2 - 2\sigma_F\sigma_X\rho_{F,X}$$

dôkaz:

$$\begin{aligned}
& \sum_f \sum_x P(F=f, X=x)(f-x)^2 \\
= & \sum_f \sum_x P(F=f, X=x)(f^2 - 2fx + x^2) \\
= & \sum_f P(F=f)f^2 + \sum_x P(X=x)x^2 - 2\sum_f \sum_x P(F=f, X=x)fx \\
= & E(F^2) + E(X^2) - 2E(F,X) \\
= & (\sigma_F^2 + \mu_F^2) + (\sigma_X^2 + \mu_X^2) - 2(\sigma_F \sigma_X \rho_{F,X} + \mu_F \mu_X) \\
= & \mu_F^2 - 2\mu_F\mu_X + \mu_X^2 + \sigma_F^2 + \sigma_X^2 - 2\sigma_F\sigma_X\rho_{F,X} \\
= & (\mu_F - \mu_X)^2 + \sigma_F^2 + \sigma_X^2 - 2\sigma_F\sigma_X\rho_{F,X}
\end{aligned}$$

MSE je rozložená na mieru nepodmieneného vychýlenia $[(\mu_F - \mu_X)^2]$, mieru ostrosti(σ_F^2), mieru neistoty (σ_X^2) a mieru združenia ($2\sigma_F\sigma_X\rho_{F,X} = 2\sigma_{F,X}$, kde $\sigma_{F,X}$ označuje kovarianciu pozorovaní a predpovedí). Posledné tri výrazy na pravej strane spolu tvoria rozptyl chýb predpovedí (σ_{F-X}^2).

Ďalšie rozklady MSE súvisia s CR a LBR faktorizáciami. CR rozklad MSE

$$MSE_{CR}(F,X) = \sigma_X^2 + E_F(\mu_{X|F} - f)^2 - E_F(\mu_{X|F} - \mu_X)^2$$

Táto rovnica ukazuje rozklad MSE na mieru neistoty (σ_X^2), mieru spoľahlivosti $[E_F(\mu_{X|F} - f)^2]$ a mieru rozlíšenia $[E_F(\mu_{X|F} - \mu_X)^2]$.

LBR rozklad MSE

$$MSE_{LBR}(F,X) = \sigma_F^2 + E_X(\mu_{F|X} - x)^2 - E_X(\mu_{F|X} - \mu_F)^2$$

Výrazy v rovnici reprezentujú rozklad MSE na mieru ostrosti (σ_F^2), mieru podmieneného vychýlenia typu 2 $[E_X(\mu_{F|X} - x)^2]$ a mieru diskriminácie $[E_X(\mu_{F|X} - \mu_F)^2]$. Posledné dva výrazy naznačujú, že týmto rozkladom je možné sprístupniť pozorovania $x = 0$ a $x = 1$ pomocou podmienenej strednej hodnoty predpovedí ($\mu_{F|X=1}$ a $\mu_{F|X=0}$). A to znížením podmieneného vychýlenia typu 2 a súčasne zvýšením diskriminácie (zvýšiť rozdiel podmienených stredných hodnôt a nepodmienenej strednej hodnoty predpovedí).

Rozklad SS založený na MSE

$$SS_{MSE}(F, \mu_X, X) = \rho_{F,X}^2 - [\rho_{F,X} - \sigma_F/\sigma_X]^2 - [(\mu_F - \mu_X)/\sigma_X^2]$$

Všetky výrazy na pravej strane sú nezáporné a môžeme ich interpretovať s odvolaním sa na lineárny regresný model. V tomto zmysle je $\rho^2_{F,X}$ mierou združenia F a X. Predpovede sú úplne spoľahlivé, keď regresná priamka pretína nulu a má smernicu jedna. Za tohto predpokladu je σ_F rovné $\rho_{F,X}\sigma_X$ a to znamená, že druhý člen na pravej strane je mierou spoľahlivosti. Tento člen vypadne pre úplne spoľahlivé predpovede inak znižuje hodnotu SS. Tretí člen je miera celkového (nepodmieneného) vychýlenia. Vypadne pre úplne nepodmienené predpovede, inak znižuje hodnotu SS. Ak sú predpovede podmienene nevychýlené, potom SS_{MSE} a $\rho^2_{F,X}$ sú si rovné. Overovaná vzorka, ktorá je podmienene nevychýlená pre všetky predpovede, je aj nepodmienene nevychýlená (opak všeobecne neplatí). V tomto zmysle môže byť $\rho^2_{F,X}$ považované za mieru potenciálnej úspešnosti.

Kapitola 3

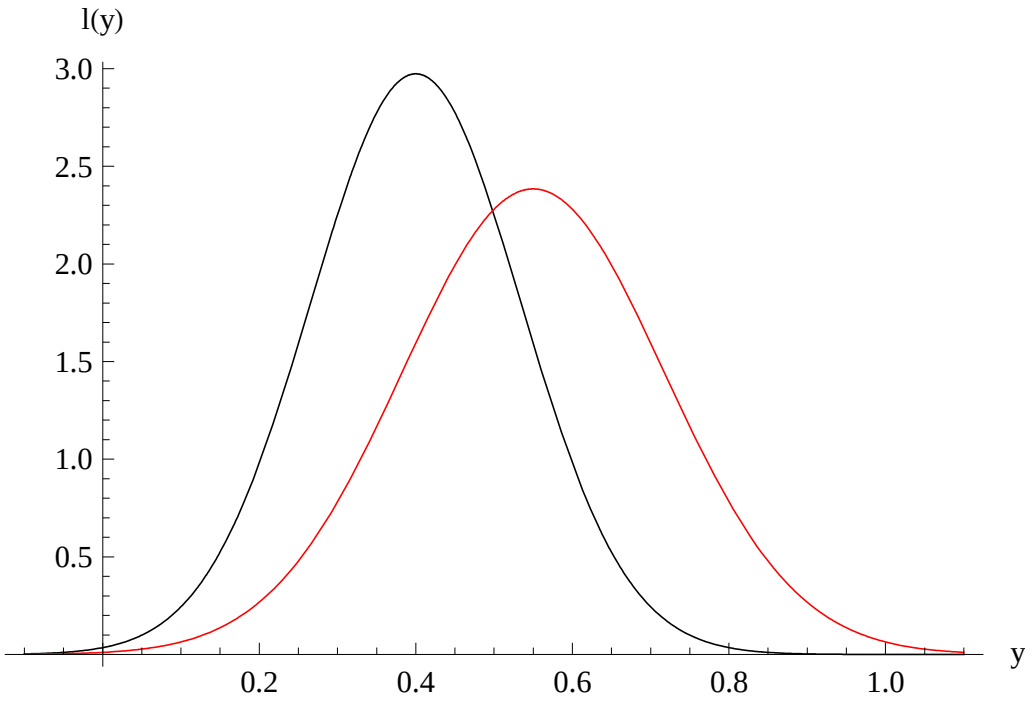
ROC krivky

V tretej kapitole sa budeme zaoberať modelom, kde poznáme podmienené rozdelenia náhodnej veličiny Y. Pre tento model študujeme ROC krivky a charakteristiky definované v druhej kapitole. V časti 3.1 definujeme ROC krivku na základe [6]. V častiach 3.2, 3.4 a 3.6 vychádzame z práce [5], v týchto častiach sú popísané ROC krivky pre rovnomerné, normálne a beta rozdelenie. V časti 3.3 sa venujeme združenému rozdeleniu predpovedí a pozorovaní a charakteristikám kvality predpovedí, ktoré sú vytvorené na základe rovnomerných rozdelení. V časti 3.5 vytvoríme pravdepodobnostnú predpoveď na základe normálnych rozdelení, určíme združené rozdelenie predpovedí a pozorovaní a venujeme sa závislosti vybraných charakteristík na rozdiele stredných hodnôt.

Pri skúmaní problému odhadovania kvality predpovedí sa meteorológovia začali zaujímať o procedúru často používanú v medicíne. Procedúra je založená na ROC krivke (receiver operating characteristic curve). Nech náhodná veličina Y popisuje výsledok nejakého merania alebo testu. Predpokladáme, že poznáme podmienené rozdelenia Y za podmienky $X = 0$ a $X = 1$, teda $Y|X = 0$ a $Y|X = 1$. Hustoty náhodných veličín $Y|X = 0$ a $Y|X = 1$ označíme l_0 a l_1. Pre ilustráciu uvádzame obrázok 3.1, na ktorom sú hustoty dvoch normálnych rozdelení $N(0.4; 0.018)$ a $N(0.55; 0.028)$. Normálnemu rozdeleniu sú venované časti 3.4 a 3.5.

3.1 Definícia ROC krivky

Parameter t, ktorý prebieha všetky možné hodnoty náhodnej veličiny Y, nazveme prah rozhodnuteľnosti (decision threshold). Pomocou parametra t



Obrázek 3.1: Hustoty normálnych rozdelení $N(0.4; 0.018)$ čierna a $N(0.55; 0.028)$ červená.

definujeme predpoveď náhodnú veličinu F takto

$$F = 0 \text{ pre } Y < t$$

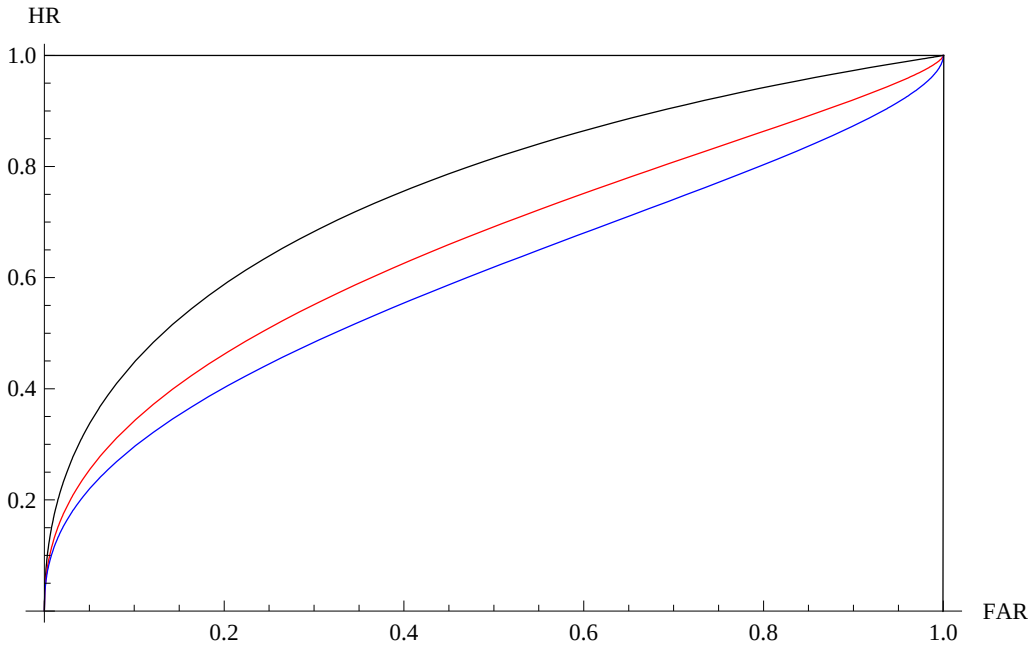
$$F = 1 \text{ pre } Y \geq t$$

Ďalej definujeme hit rate = true positive rate (HR) a false alarm rate = false positive rate (FAR)

$$HR = P(Y \geq t | X = 1) = \int_t^\infty l_1(y)dy$$

$$FAR = P(Y \geq t | X = 0) = \int_t^\infty l_0(y)dy$$

ROC krivka je potom parametrický nákres, kde sa na horizontálnu osu vynáša false alarm rate a na vertikálnu osu hit rate, pričom parameter t prebieha celú škálu možných hodnôt náhodnej veličiny Y. Na obrázku 3.2 môžeme vidieť tri ROC krivky pre tri rôzne dvojice normálnych rozdelení.



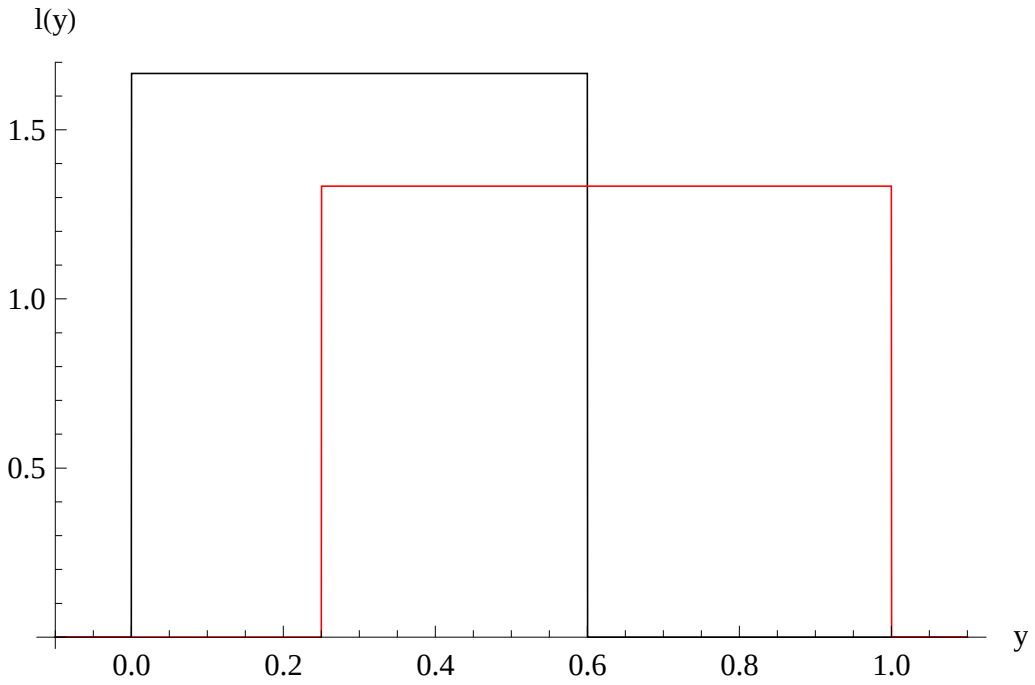
Obrázek 3.2: ROC krivky vytvorené na základe normálnych rozdelení $N(0.4; 0.018)$, $N(0.55; 0.028)$ čierna, $N(0.4; 0.02)$, $N(0.5; 0.04)$ červená a $N(0.4; 0.03)$, $N(0.48; 0.07)$ modrá.

Množina $[0,1]^2 = \{(FAR, HR), FAR \in (0,1), HR \in (0,1)\}$ sa nazýva ROC priestor. Uhlopriečka zodpovedá náhodnej predpovedi. Stupeň konkávnosti môžeme považovať za mieru kvality predpovede.

Plocha pod ROC krivkou (AUC, area under ROC curve) je často považovaná za skalárnu mieru kvality predpovede. AUC môžeme vyjadriť pomocou vzorca $AUC = \int_0^1 HR(FAR)dFAR$. Hodnota AUC rovná 0.5 vyjadruje náhodnú predpoveď, kým AUC rovné 1 znamená dokonalú predpoveď. Je dokázané, že AUC je blízko späté s ekonomickou hodnotou prepovedného systému.

3.2 ROC krivka pre rovnomerné rozdelenie

Na obrázku 3.3 sa nachádzajú hustoty dvoch rovnomerných rozdelení. Nech náhodná veličina $Y|X = 0$ má rovnomerné rozdelenie $R(c_0 - w_0, c_0 + w_0)$, náhodná veličina $Y|X = 1$ má $R(c_1 - w_1, c_1 + w_1)$. Tento popis obsahuje štyri parametre a to dve stredné hodnoty c_0, c_1 a dve polšírky w_0, w_1. Bez ujmy



Obrázek 3.3: Hustoty rovnomerných rozdelení $R(0; 0.6)$ čierna a $R(0.25; 1)$ červená.

na všeobecnosti budeme predpokladať, že $c_1 \geq c_0$. Potom hit rate a false alarm rate sú dané rovnicami

$$HR = \frac{c_1 + w_1 - t}{2w_1}$$

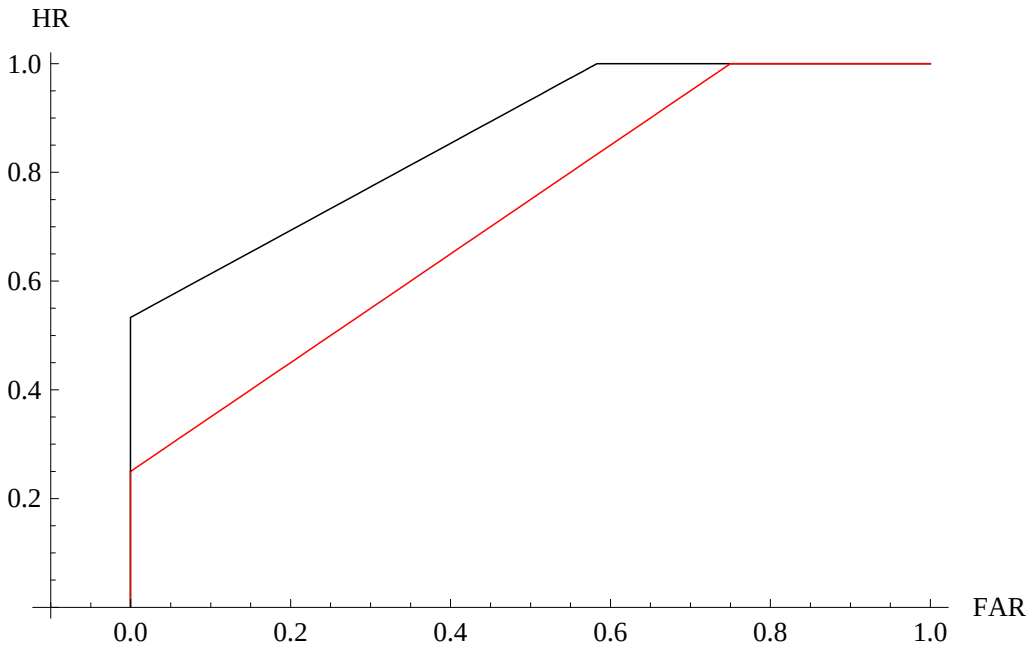
$$FAR = \frac{c_0 + w_0 - t}{2w_0}$$

odvodenie:

$$l_1(y) = \frac{1}{2w_1} \mathrm{I}_{(c_1 - w_1, c_1 + w_1)}(y)$$

$$HR = \int_t^\infty l_1(y)dy = \frac{1}{2w_1} \int_t^\infty \mathrm{I}_{(c_1 - w_1, c_1 + w_1)}(y)dy = \frac{c_1 + w_1 - t}{2w_1}$$

kde $\mathrm{I}_{(a,b)}(y)$ je indikátor javu $y \in (a, b)$.
Odvodíme rovnicu pre ROC krivku. Z rovnice pre FAR vyjadríme t a dosadíme do rovnice pre HR.



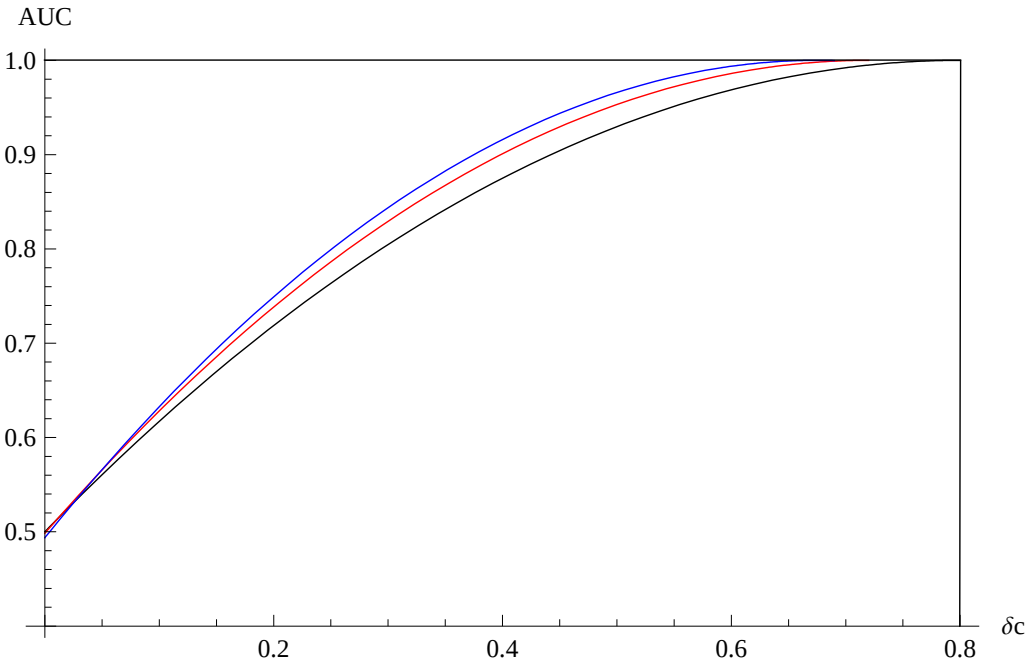
Obrázek 3.4: ROC krivky vytvorené na základe rovnomerných rozdelení $R(0; 0.6)$, $R(0.25; 1)$ čierna a $R(0; 0.8)$, $R(0.2; 1)$ červená.

$$\begin{aligned} t &= c_0 + w_0 - 2FARw_0 \\ HR(FAR) &= \frac{c_1 + w_1 - (c_0 + w_0 - 2FARw_0)}{2w_1} \\ HR(FAR) &= \frac{c_1 - c_0 + w_1 - w_0 + 2FARw_0}{2w_1} \\ HR(FAR) &= \frac{w_0}{w_1} FAR + \frac{\delta c + \delta w}{2w_1} \end{aligned}$$

kde $\delta c = c_1 - c_0$ a $\delta w = w_1 - w_0$

Na obrázku 3.4 vidíme, že ROC krivka sa skladá z troch lineárnych častí, pričom stredná časť je daná rovnicou, ktorú sme odvodili. Rovnica naznačuje, že dva modely s rôznymi strednými hodnotami a šírkami možu viesť k rovnakej ROC krivke, ak majú rovnakú smernicu a priesečník s vertikálnou osou. Všeobecne platí, že ROC krivky neurčujú jednoznačne parametre podkladových rozdelení. Existujú rozdelenia s rôznymi parametrami,

ktoré dávajú vznik rovnakým ROC krivkám. Dĺžka zvislej časti, ktorá prekrýva vertikálnu osu, je určená veličinami δc a w_0/w_1. Smernica závisí iba na pomere polšírok. Ak w_1 nie je rovné w_0, potom sa táto nerovnosť prejaví ako asymetrickosť ROC krivky.



Obrázek 3.5: AUC ako funkcia δc pre rovnomerné rozdelenia s parametrami $w_0 = 0.3$, $w_1 = 0.375$ modrá, $w_0 = 0.34$, $w_1 = 0.38$ červená a $w_0 = 0.4$, $w_1 = 0.4$ čierna.

Z analytického vyjadrenia ROC krivky môžeme spočítať plochu pod krivkou.

$$AUC = 1 - \frac{1}{8}\frac{[\delta c - (w_0 + w_1)]^2}{w_0 w_1}$$

Pre rozdelenia na obrázku 3.3 je $\delta c \leq w_0 + w_1$, potom vidíme, že zväčšovanie hodnoty δc vedie k vyššej hodnote AUC. Znázornené je to na obrázku 3.5. Rovnako aj znižovanie w_0 a w_1 môže viesť k lepšiemu výsledku. Ak budeme vyberať model na základe týchto výsledkov, volíme rozdelenia, pre ktoré bude rozdiel stredných hodnôt čo najväčší a ktoré nebudú mať veľké rozptyly.

3.3 Charakteristiky pre rovnomerné rozdelenie

Budeme predpokladať, že náhodná veličina X má alternatívne rozdelenie s parametrom q, teda $P(X = 1) = q$ a $P(X = 0) = 1 - q$. Náhodná veličina $Y|X = i$ má rozdelenie $R(c_i - w_i, c_i + w_i)$, kde $i = 0, 1$. Ďalej predpokladáme

$$\begin{aligned} c_0 &< c_1 \\ c_0 + w_0 &> c_1 - w_1 \\ c_0 + w_0 &< c_1 + w_1 \\ c_0 - w_0 &< c_1 - w_1 \end{aligned}$$

Potom pravdepodobnostnú predpoveď náhodnú veličinu F môžeme vytvoriť takto

$$F(y) = \frac{l_1(y)}{l_1(y) + l_0(y)}$$

Pravdepodobnostná predpoveď F je potom rovná podmienenej pravdepodobnosti $P(X = 1|y)$

$$\begin{aligned} F(y) &= \frac{\frac{1}{2w_1}\mathrm{I}_{(c_1-w_1,c_1+w_1)}(y)}{\frac{1}{2w_1}\mathrm{I}_{(c_1-w_1,c_1+w_1)}(y) + \frac{1}{2w_0}\mathrm{I}_{(c_0-w_0,c_0+w_0)}(y)} \\ &= 0 \times \mathrm{I}_{(0,c_1-w_1)}(y) + \frac{w_0}{w_0 + w_1} \times \mathrm{I}_{(c_1-w_1,c_0+w_0)}(y) + 1 \times \mathrm{I}_{(c_0+w_0,1)}(y) \end{aligned}$$

Vidíme, že náhodná veličina F nadobúda len 3 hodnoty a to 0, $\frac{w_0}{w_0+w_1}$ a 1. Určíme podmienené pradvepodobnosti $P(F = f|X = x)$, kde $f = 0, \frac{w_0}{w_0+w_1}, 1$ a $x = 0, 1$.

$$\begin{aligned} P(F = 0|X = 0) &= P(Y \in (0, c_1 - w_1)|X = 0) \\ &= \frac{1}{2w_0}(c_1 - w_1 - c_0 + w_0) \\ P(F = \frac{w_0}{w_0 + w_1}|X = 0) &= P(Y \in (c_1 - w_1, c_0 + w_0)|X = 0) \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{2w_0}(c_0 + w_0 - c_1 + w_1) \\
P(F = 1|X = 0) &= P(Y \in (c_0 + w_0, 1)|X = 0) \\
&= 0 \\
P(F = 0|X = 1) &= 0 \\
P(F = \frac{w_0}{w_0 + w_1}|X = 1) &= \frac{1}{2w_1}(c_0 + w_0 - c_1 + w_1) \\
P(F = 1|X = 1) &= \frac{1}{2w_1}(c_1 + w_1 - c_0 - w_0)
\end{aligned}
$$

Pomocou vzťahu $P(F = f, X = x) = P(F = f|X = x)P(X = x)$ určíme združené rozdelenie F a X.

$$
\begin{aligned}
P(F = 0, X = 0) &= (1 - q)\frac{1}{2w_0}(c_1 - w_1 - c_0 + w_0) \\
P(F = \frac{w_0}{w_0 + w_1}, X = 0) &= (1 - q)\frac{1}{2w_0}(c_0 + w_0 - c_1 + w_1) \\
P(F = 1, X = 0) &= 0 \\
P(F = 0, X = 1) &= 0 \\
P(F = \frac{w_0}{w_0 + w_1}, X = 1) &= q\frac{1}{2w_1}(c_0 + w_0 - c_1 + w_1) \\
P(F = 1, X = 1) &= q\frac{1}{2w_1}(c_1 + w_1 - c_0 - w_0)
\end{aligned}
$$

Teraz môžeme vyjadriť charakteristiky predpovedí, ktoré sme popísali v druhej kapitole.

neistota

$$\sigma_X^2 = q(1 - q)$$

ostrosť

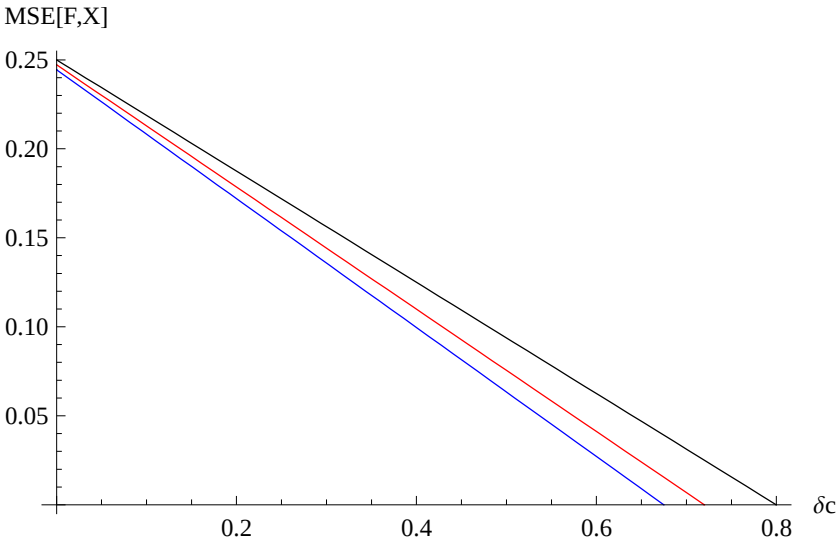
$$\sigma_F^2 = \frac{(w_0 - w_1)(1 - 2q)}{4(w_0 + w_1)} + \delta c\frac{qw_0 + (1 - q)w_1}{2(w_0 + w_1)^2} - (\delta c)^2\frac{(1 - 2q)^2}{4(w_0 + w_1)^2}$$

združenie

$$\sigma_{F,X} = \frac{1}{\sqrt{\sigma_X^2}\sqrt{\sigma_F^2}}\left[q\frac{w_0}{w_1} - \frac{2q^2}{w_0 + w_1} + \frac{3\delta c}{2(w_0 + w_1)}\right]$$

správnosť

$$MSE(F,X) = [(1-q)w_0 + qw_1]\left[\frac{1}{2(w_0+w_1)} - \frac{\delta c}{2(w_0+w_1)^2}\right]$$



Obrázek 3.6: MSE(F,X) ako funkcia δc pre rovnomerné rozdelenia s parametrami $w_0 = 0.3$, $w_1 = 0.375$ modrá, $w_0 = 0.34$, $w_1 = 0.38$ červená a $w_0 = 0.4$, $w_1 = 0.4$ čierna, parameter q je rovný 0.4.

úspešnosť

$$SS(F,\mu_X,X) = 1 - \left(\frac{w_0}{q} + \frac{w_1}{1-q}\right)\left[\frac{1}{2(w_0+w_1)} - \frac{\delta c}{2(w_0+w_1)^2}\right]$$

vychýlenie

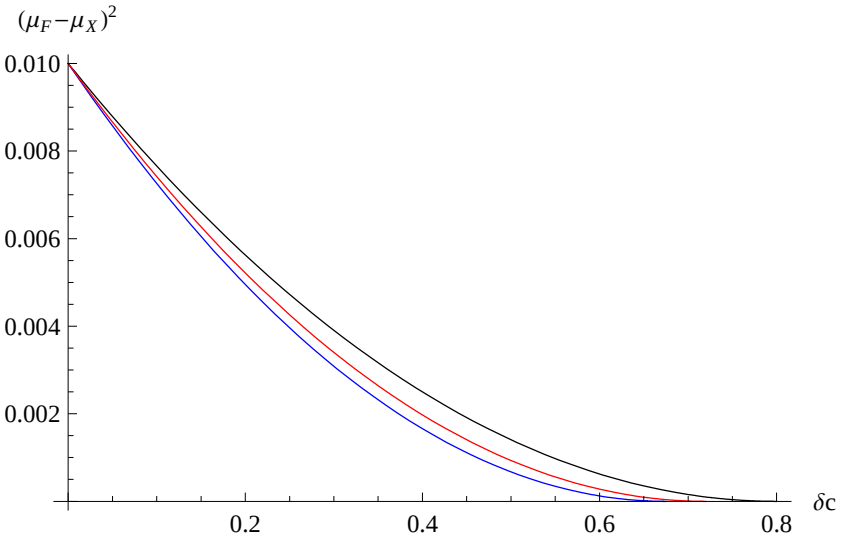
$$(\mu_F - \mu_X)^2 = \left[\frac{1}{2}\left(1 - \delta c\frac{1-2q}{w_0+w_1}\right) - q\right]^2$$

spoľahlivosť

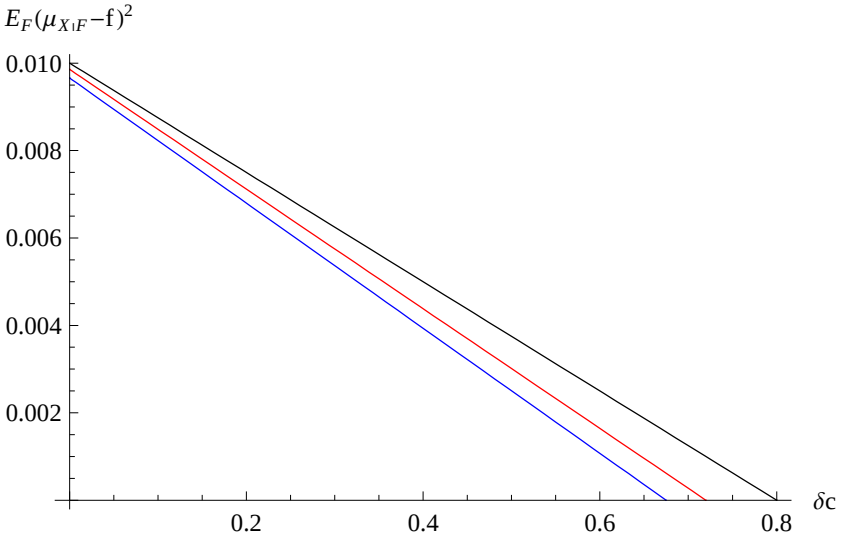
$$E_F(\mu_{X|F} - f)^2 = (w_0 + w_1 - \delta c)\frac{w_0 w_1 (2q-1)^2}{2(w_0+w_1)^2[(1-q)w_1 + qw_0]}$$

rozlíšenie

$$E_F(\mu_{X|F} - \mu_X)^2 = \frac{q(1-q)}{2[(1-q)w_1 + qw_0]}[\delta c + (w_1 - w_0)(1-2q)]$$



Obrázek 3.7: Vychýlenie ako funkcia δc pre rovnomerné rozdelenia s parametrami $w_0 = 0.3$, $w_1 = 0.375$ modrá, $w_0 = 0.34$, $w_1 = 0.38$ červená a $w_0 = 0.4$, $w_1 = 0.4$ čierna, parameter q je rovný 0.4.



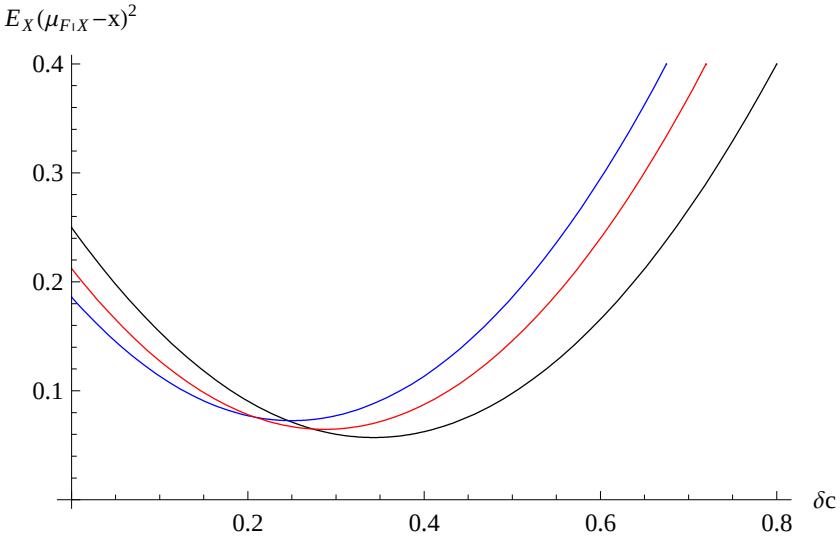
Obrázek 3.8: Spoľahlivosť ako funkcia δc pre rovnomerné rozdelenia s parametrami $w_0 = 0.3$, $w_1 = 0.375$ modrá, $w_0 = 0.34$, $w_1 = 0.38$ červená a $w_0 = 0.4$, $w_1 = 0.4$ čierna, parameter q je rovný 0.4.

diskriminácia

$$E_X(\mu_{F|X} - \mu_F)^2 = q\frac{w_0^2}{w_1^2} - \delta c\frac{2qw_0(qw_1 + w_0)}{w_1^2(w_0 + w_1)} + (\delta c)^2\frac{q(qw_1^2 + 2qw_1w_0 + w_0^2)}{w_1^2(w_0 + w_1)^2}$$

podmienené vychýlenie

$$E_X(\mu_{F|X} - x)^2 = \frac{1-q}{4}\left(1 - \frac{\delta c}{w_0 + w_1}\right)^2 + q\left[\left(1 - \frac{\delta c}{w_0 + w_1}\right)\left(\frac{w_0}{w_1} + \frac{1}{2}\right) - 1\right]^2$$



Obrázek 3.9: Podmienené vychýlenie ako funkcia δc pre rovnomerné rozdelenia s parametrami $w_0 = 0.3$, $w_1 = 0.375$ modrá, $w_0 = 0.34$, $w_1 = 0.38$ červená a $w_0 = 0.4$, $w_1 = 0.4$ čierna, parameter q je rovný 0.4.

Na obrázku 3.6 vidíme, že MSE je klesajúca funkcia rozdielu stredných hodnôt δc. Keďže vyššiu správnosť majú predpovede, pre ktoré je hodnota MSE nižšia, uprednostňujeme väčší rozdiel stredných hodnôt.

Na obrázku 3.7 je znázornená závislosť vychýlenia na δc. Vidíme, že je to opäť klesajúca funkcia. Chceme aby predpovede boli nevychýlené (aby vychýlenie bolo čo najnižšie), preto je pre nás lepší väčší rozdiel stredných hodnôt.

Vyššiu spoľahlivosť majú predpovede, ktoré vykazujú väčšiu zhodu $\mu_{X|F}$ a F , teda nízku hodnotu $E_X(\mu_{X|F} - f)^2$. Ako vidíme z obrázka 3.8, opäť uprednostňujeme vyššiu hodnotu δc.

Pravdepodobnostné predpovede, pre ktoré platí, že $\mu_{F|X} = X$ pre všetky x, sú dokonalé predpovede. Chceme, aby hodnota $E_X(\mu_{F|X} - x)^2$ bola minimálna. Na obrázku 3.9 je znázornená závislosť podmieneného vychýlenia na rozdiele stredných hodnôt δc. Vidíme, že v skúmaných prípadoch sa minimum pohybuje v intervale $(0.2, 0.4)$ a pre hodnoty $\delta c > 0.6$ je podmienené vychýlenie vysoké.

3.4 ROC krivka pre normálne rozdelenie

Normálne rozdelenie nám ponúka realistickejšiu aproximáciu skutočnosti. Avšak výrazy pre ROC krivku a AUC nie sú názorné, pretože obsahujú integrál z funkcie $e^{-x^2/2}$, ktorý nevieme vyjadriť pomocou elementárnych funkcií v konečnom tvare. Tento integrál vyjadríme pomocou distribučnej funkcie normovaného normálneho rozdelenia.

Hustotu náhodnej veličiny Y za podmienky, že pozorovanie je z i-tej skupiny zapíšeme takto

$$l_i(y) = \frac{1}{\sqrt{2\pi w_i^2}} \exp\left\{-\frac{(y - c_i)^2}{2w_i^2}\right\}$$

kde c_i je stredná hodnota a w_i^2 je rozptyl rozdelenia $\mathrm{N}(c_i, w_i^2)$, $i = 0, 1$. Potom $FAR = \Phi((c_0 - t)/w_0)$, $HR = \Phi((c_1 - t)/w_1)$, kde $\Phi(y) = \frac{1}{\sqrt{2\pi}} \int_\infty^y e^{-z^2/2} dz$ je distribučná funkcia normálneho rozdelenia $\mathrm{N}(0, 1)$. Vylúčením prahu t z rovníc pre HR a FAR dostaneme formálny výraz pre ROC krivku

$$HR(FAR) = \Phi\left(\frac{\delta c}{w_1} - \frac{w_0}{w_1}\Phi^{-1}(FAR)\right)$$

kde Φ^{-1}je definovaná ako inverzná funkcia k Φ, $\Phi^{-1}\Phi = 1$. ROC krivky vytvorené na základe normálnych rozdelení sú znázornené na obrázku 3.2.

Použitím substitúcie $HR^* = \Phi^{-1}(HR)$, $FAR^* = \Phi^{-1}(FAR)$ dostávame podobnú rovnicu ako je rovnica ROC krivky pre rovnomerné rozdelenie

$$HR^*(FAR^*) = \frac{\delta c}{w_1} - \frac{w_0}{w_1} FAR^*$$

Predpoklad, že teoretická ROC krivka založená na normálnom rozdelení spĺňa základné vlastnosti ROC kriviek ako je konkávnosť nad aj pod diagonálou, nie je správny. Hoci platí, že v symetrickom prírade, teda keď

$w_0 = w_1$ nie je ROC krivka vyložene konkávna, je ľahké ukázať, že ak $w_0 \neq w_1$ potom ROC krivka pretne diagonálu presne v jednom bode (inom ako koncový bod). ROC krivka pretne diagonálu v bode kde

$$\begin{aligned} \Phi\left(\frac{c_0 - t}{w_0}\right) &= \Phi\left(\frac{c_1 - t}{w_1}\right) \\ \frac{c_0 - t}{w_0} &= \frac{c_1 - t}{w_1} \\ \frac{c_0}{w_0} - \frac{c_1}{w_1} &= \left(\frac{1}{w_0} - \frac{1}{w_1}\right) t \end{aligned}$$

Táto rovnica má iba jedno netriviálne riešenie pre $w_0 \neq w_1$ a to

$$t = \left(\frac{c_0}{w_0} - \frac{c_1}{w_1}\right)\left(\frac{1}{w_0} - \frac{1}{w_1}\right)^{-1} = \frac{c_0 w_1 - c_1 w_0}{w_1 - w_0}$$

Hodnota *FAR* v priesečníku je $\Phi(\delta c/\delta w)$. Ale tento výsledok musíme interpretovať opatrne. Netvrdí, že konkávna empirická ROC krivka svedčí, že $w_0 = w_1$. Aj pre $w_0 \neq w_1$ ROC krivka môže byť väčšinou konkávna (bez pretínania). Pretože $\Phi(x)$ je rýchlo rastúca funkcia x, je približne 0 resp. 1 keď x je približne 2 resp. -2. Preto konkávna empirická ROC krivka poukazuje na jednu z dvoch možností buď $w_0 = w_1$ alebo $w_0 \neq w_1$ a súčasne $|\delta c/\delta w| \geq\sim 2$.

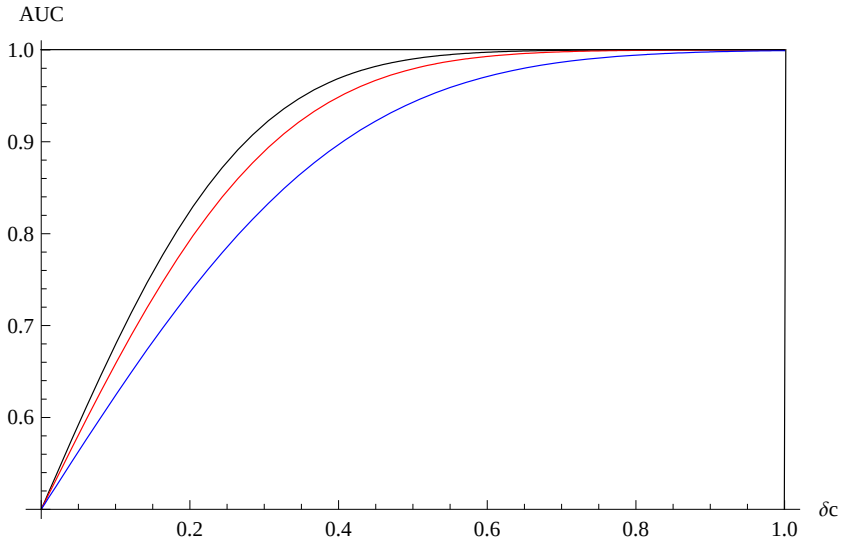
Hodnotu AUC určíme takto

$$AUC = \Phi\left(\frac{\delta c}{\sqrt{w_0^2 + w_1^2}}\right)$$

AUC je opäť nelineárna funkcia podkladových parametrov δc, w_0^2, w_1^2. Na obrázku 3.10 vidíme AUC ako funkciu δc.

3.5 Charakteristiky pre normálne rozdelenie

Opäť predpokladáme, že náhodná veličina X má alternatívne rozdelenie s parametrom q. Nech náhodná veličina $Y|X = 0$ má normálne rozdelenie $N(c_0, w^2)$, náhodná veličina $Y|X = 1$ má rozdelenie $N(c_1, w^2)$, pre jednoduchosť predpokladáme, že obe rozdelenia majú rovnaký rozptyl w^2. Pravdepodobnostnú predpoveď F vytvoríme opäť ako $F(y) = l_1(y)/(l_1(y) + l_0(y))$.



Obrázek 3.10: AUC ako funkcia δc pre normálne rozdelenia s parametrami $w_0 = \sqrt{0.018}$, $w_1 = \sqrt{0.028}$ čierna, $w_0 = \sqrt{0.02}$, $w_1 = \sqrt{0.04}$ červená a $w_0 = \sqrt{0.03}$, $w_1 = \sqrt{0.07}$ modrá.

$$\begin{aligned} f = F(y) &= \frac{\frac{1}{\sqrt{2\pi w^2}} \exp\left\{-\frac{(y-c_1)^2}{2w^2}\right\}}{\frac{1}{\sqrt{2\pi w^2}} \exp\left\{-\frac{(y-c_1)^2}{2w^2}\right\} + \frac{1}{\sqrt{2\pi w^2}} \exp\left\{-\frac{(y-c_0)^2}{2w^2}\right\}} \\ &= \frac{1}{1 + \exp\left\{-\frac{y(c_1-c_0)}{w^2} - \frac{c_0^2-c_1^2}{2w^2}\right\}} \end{aligned}$$

Podmienené hustoty $h_{F|X=0}$ a $h_{F|X=1}$ určíme pomocou vety o transformácií náhodnej veličiny. Túto vetu môžeme násť napríklad v [7, veta 3.5, strana 48].

$$h_{F|X=0}(f) = l_0(G(f))|G'(f)|$$
$$h_{F|X=1}(f) = l_1(G(f))|G'(f)|$$

kde $G(f)$ je inverzná funkcia k $F(y)$.

$$G(f) = \frac{w^2}{c_1 - c_0} \log\left(\frac{f}{1-f}\right) + \frac{c_1 + c_0}{2}$$

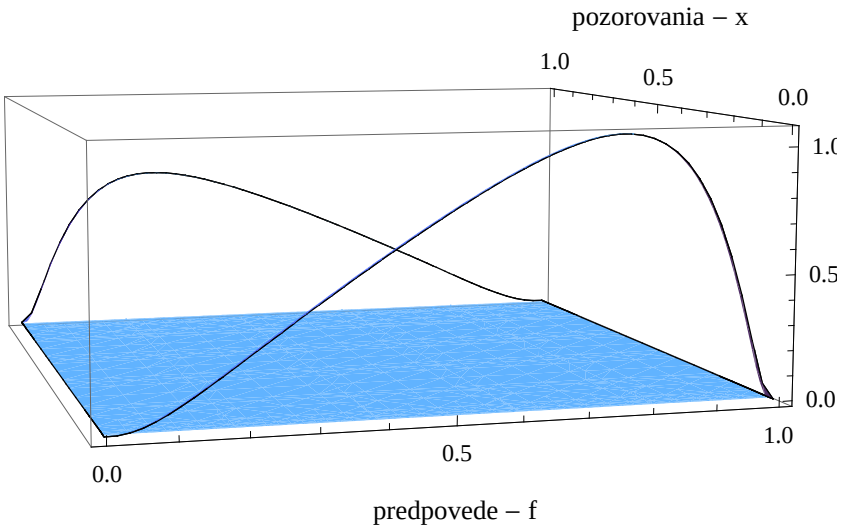
$$G'(f) = \frac{w^2}{(c_1 - c_0) f(1-f)}$$

Vidíme, že pre f z intervalu $(0,1)$ je derivácia $G'(f)$ vždy kladná.

$$h_{F|X=0}(f) = \sqrt{\frac{w^2}{2\pi}} \frac{1}{(c_1 - c_0) f(1-f)} \times \exp\left\{ -\frac{1}{2w^2} \left[\frac{w^2}{c_1 - c_0} \log\left(\frac{f}{1-f} \right) + \frac{c_0 - c_1}{2} \right]^2 \right\} \quad (3.1)$$

$$h_{F|X=1}(f) = \sqrt{\frac{w^2}{2\pi}} \frac{1}{(c_1 - c_0) f(1-f)} \times \exp\left\{ -\frac{1}{2w^2} \left[\frac{w^2}{c_1 - c_0} \log\left(\frac{f}{1-f} \right) + \frac{c_1 - c_0}{2} \right]^2 \right\} \quad (3.2)$$

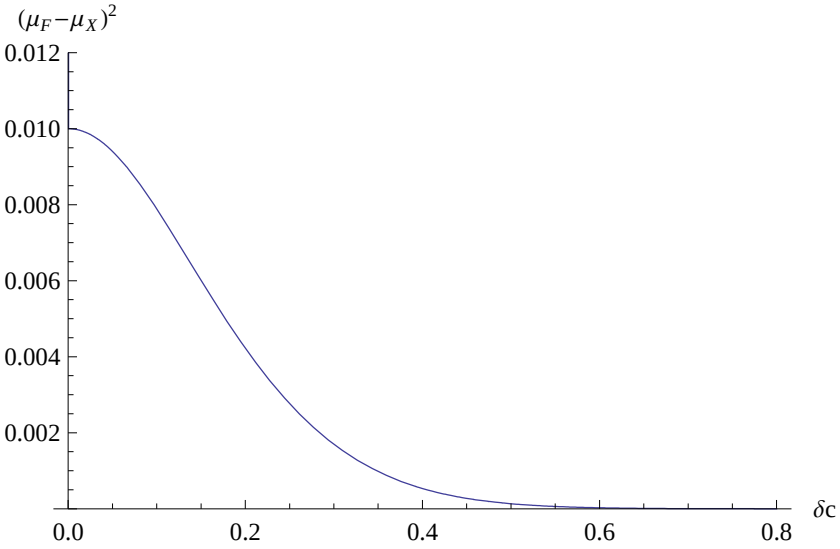
Združené rozdelenie predpovedí a pozorovaní je určené marginálnym rozdelením pozorovaní a podmienenými rozdeleniami (3.1) a (3.2).



Obrázek 3.11: Združené rozdelenie predpovedí a pozorovaní, parametre $c_0 = 0.4$, $c_1 = 0.55$, $w^2 = 0.02$, $q = 0.4$.

Na obrázku 3.11 vidíme združené rozdelenie predpovedí a pozorovaní.

Ďalej študujeme závislosť vybraných charakteristík na rozdiele stredných hodnôt. Hodnoty charakteristík sme počítali pomocou numerickej integrácie v programe Wolfram Mathematica 6.0.

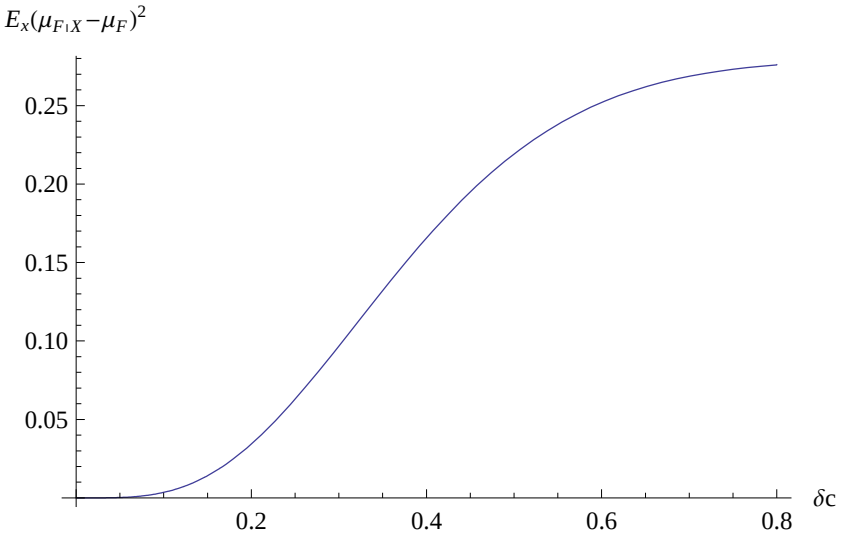


Obrázek 3.12: Vychýlenie ako funkcia δc pre normálne rozdelenia s rozptylom $w^2 = 0.02$, parameter $q = 0.4$.

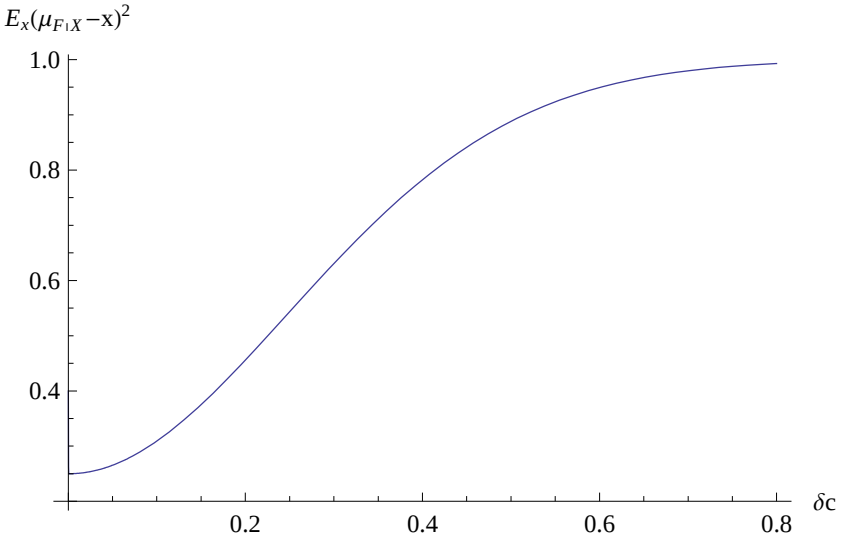
Z obrázka 3.12 vidíme, že vychýlenie je klesajúca funkcia rozdielu stredných hodnôt δc. Lepšie sú predpovede, ktoré sú nevychýlené, preto uprednostňujeme väčšiu hodnotu δc.

Na obrázku 3.13 je znázornená závislosť diskriminácie na δc. Keďže diskrimináciu môžeme chápať ako schopnosť rozlišovať medzi pozorovaniami, je jej vyššia hodnota výhodnejšia a to dosiahneme pri väčšom rozdiele stredných hodnôt.

Obrázok 3.14 znázorňuje podmienené vychýlenie ako funkciu δc. Chceli by sme aby naše predpovede boli podmienene nevychýlené. Pretože podmienené vychýlenie je rastúca funkcia δc, je pre nás tentokrát výhodnejšia nízka hodnota δc.



Obrázek 3.13: Diskriminácia ako funkcia δc pre normálne rozdelenia s rozptylom $w^2 = 0.02$, parameter $q = 0.4$.



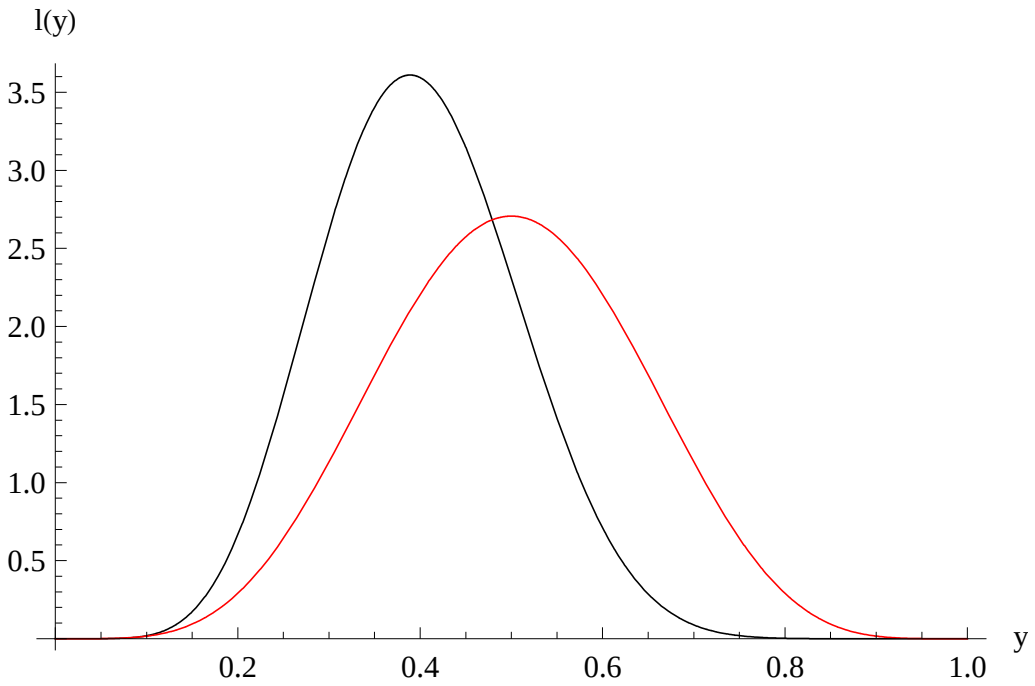
Obrázek 3.14: Podmienené vychýlenie ako funkcia δc pre normálne rozdelenia s rozptylom $w^2 = 0.02$, parameter $q = 0.4$.

3.6 ROC krivka pre beta rozdelenie

Náhodná veličina $Y|X = i$ môže mať aj beta rozdelenie $\beta(a_i, b_i)$ potom

$$l_i(y) = \frac{1}{\beta(a_i, b_i)} y^{a_i - 1}(1-y)^{b_i - 1}, y \in (0,1), i = 0, 1$$

kde β je beta funkcia $\beta(a_i, b_i) = \int_0^1 x^{a_i-1}(1-x)^{b_i-1}dx$. Na obrázku 3.15 sú zobrazené hustoty dvoch beta rozdelení.



Obrázek 3.15: Hustoty beta rozdelení $B(8; 12)$ čierna a $B(6; 6)$ červená.

Strednú hodnotu a rozptyl vypočítame takto $c_i = \frac{a_i}{a_i+b_i}$, $w_i^2 = \frac{a_i b_i}{(a_i+b_i)^2(a_i+b_i+1)}$ Výrazy pre *HR*, *FAR* nie sú opäť názorné

$$HR = \int_t^1 \frac{1}{\beta(a_0, b_0)} y^{a_0-1}(1-y)^{b_0-1}dy$$

$$FAR = \int_t^1 \frac{1}{\beta(a_1, b_1)} y^{a_1-1}(1-y)^{b_1-1}dy$$

Smernica ROC krivky je

$$s(t) = \frac{\beta(a_0, b_0)}{\beta(a_1, b_1)} t^{a_1-a_0}(1-t)^{b_1-b_0}$$

Aby bola ROC krivka symetrická, budeme vyžadovať, aby boli smernice v koncových bodoch krivky nepriamo úmerné. To platí, práve vtedy keď

$$a_1 + b_1 = a_0 + b_0$$

Z toho plynie, že ROC krivka je symetrická ak $a_1 + b_1 = a_0 + b_0$, čo sa dá zapísať aj pomocou stredných hodnôt a rozptylov

$$\frac{c_1(1-c_1)}{w_1^2} = \frac{c_0(1-c_0)}{w_0^2}$$

Zjavná asymetria v empirickej ROC krivke vyjadruje, že táto rovnosť neplatí. Poznamenajme, že v prípade symetrickej ROC krivky sa výrazy $a_1 - a_0$ a $b_1 - b_0$ líšia iba v znamienku.

ROC krivka pretne diagonálu keď $a_1 > a_0$ a $b_1 > b_0$, pretože smernice v dvoch extrémoch sú menšie ako jedna. Tieto dve nerovnosti spolu implikujú, že

$$\frac{c_1(1-c_1)}{w_1^2} > \frac{c_0(1-c_0)}{w_0^2}$$

Porovnaním s predchádzajúcou podmienkou pre symetrickú ROC krivku vidíme, že veličina $c(1-c)/w$ podmieňuje oba javy, symetriu aj pretínanie diagonály.

Kapitola 4

Záver

Cieľom práce bolo predstaviť a porovnať prístupy k hodnoteniu kvality predpovedí založené na ROC krivke a na združenom rozdelení pravdepodobnostných predpovedí a pozorovaní.

Mojim prínosom v tejto práci je odvodenie charakteristík pravdepodobnostnej predpovede v časti 3.3 pre rovnomerné rozdelenia a v časti 3.5 pre normálne rozdelenia a štúdium týchto charakteristík ako funkcií rozdielu stredných hodnôt náhodných veličín na základe, ktorých boli predpovede vytvorené.

Ďalej by sme mohli v práci pokračovať štúdiom charakteristík kvality predpovede a ROC kriviek pre iné pravdepodobnostné rozdelenia. Mohli by sme skúmať a znázorniť tieto charakteristiky ako funkcie iných parametrov než δc.

Literatura

[1] Murphy A. H.: *Forecast Verification. In: Economic Value of Weather and Climate Forecast, ed. by R. W. Katz and A. H. Murphy*, Cambridge University Press, 1997.

[2] Murphy A. H.: *What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting*,Weather and forecasting, Vol. 8, No. 2, 1993, P. 281-293.

[3] Murphy A. H., Winkler R. L.: *A General Framework for Forecast Verifications*, Monthly weather review, Vol. 115, No. 7, 1987, P. 1330-1338.

[4] Zymáková I.: *Bakalárska práca na MFF UK: Ověřování pravděpodobnostních předpovědí*, 2007.

[5] Marzban C.: *A Comment on the ROC Curve and Area Under it as Performance Measures*, Weather and Forecasting, Vol. 19, No. 6, 2004.

[6] Pepe M. S.: *The Statistical Evaluation of Medical Test for Classification and Prediction*, Oxford University Press, 2003.

[7] Anděl J.: *Základy matematické statistiky*, Matfyzpress, 2007.