Typy atribútov

- numerické
    - spojité (napr. vzdialenosti)
    - vieme merať vzdialenosti medzi dátami
    - diskrétne (napr. vek)
- ordinálne
    - vieme ich zoradiť
    - napr. vriaci, horúci, vlažný, studený
- nominálne (kategorické)
    - nevieme ich zoradiť
    - napr. slnečno, zamračené

Pravidlo palca pri tvorbe histogramu

- počet intervalov je odmocnina počtu pozorovaní

Kontingenčná tabuľka

- tabuľka vyjadrujúca vzťah dvoch alebo viacerých štatistických znakov

Charakteristiky atribútu

- početnosť výskytu
    - pre nominálne hodnoty je to jediný spôsob ako zistiť charakter dát
    - histogram
- modus (najčastejšie sa vyskytujúca hodnota)
- aritmetický priemer (stredná hodnota)
- medián
    - vhodný pre data s outliermi
- rozptyl (štandardná odchýlka)
- percentily
    - napr. 25% percentil (označovaný aj dolný kvartil) znamená, že 25% hodnôt je pod určitou hodnotou

Boxplot

- cieľom je sumarizovať dáta a rýchlo zobraziť, či sú dáta symetrické a či majú outlierov
- zakreslíme rámčekom dolný kvartil, medián a horný kvartil
- úsečkou zobrazíme maximum a minimum, ktoré sú maximálne vzdialené 1,5x šírky (rozdiel medzi spodným a horným kvartilom) rámčeka
- outliery sú nad/pod minimum/maximom a sú zobrazené krúžkami

Scatter plot

- zisťujú sa vzťahy medzi dvojicou premenných
- scatter matrix sú grafy všetkých kombinácií dvojíc premenných

Príprava dát sa používa na

- nekonzistentné dáta
- nekompletné dáta
- zašumené dáta
- mnohorozmerné dáta

Príprava vstupu (dát) – predspracovanie

- integrácia dát z rôznych zdrojov
    - rôzne štýly ukladania záznamov
    - rôzne konvencie
    - rôzne časové obdobia
    - rôzne typy chýb
    - rôzne formáty
    - rôzne veličiny
    - ** je potrebné identifikovať rovnaké entity s inými ID a redundancia prebytočných atribútov, ktoré sa dajú vypočítať z iných atribútov

Vysporiadanie sa s chýbajúcimi hodnotami

- ignorovať záznam
- vyplniť manuálne
- použiť konštantnú hodnotu
- použiť heuristiku
- použiť aritmetický priemer hodnôt
- použiť hodnotu najbližšieho suseda
- použiť náhodnú hodnotu
- použiť interpoláciu

Knowledge Discovery in Databases

- je celý proces hľadania znalostí vrátane predspracovania dát aj data miningu.

Data Mining

- je jadrom KDD, jedným z krokov KDD.
- delí sa na (podľa úloh a typov modelov):
    - prediktívne úlohy
        - *Typy modelov:*

        - Klasifikácia
        - Regresia
        - Analýza časových radov
        - Predikcia
        - Detekcia anomálií
    - deskriptívne úlohy
        - *Typy modelov:*
        - Zhlukovanie
        - Asociačné pravidlá
        - Objavovanie sekvencií
        - Detekcia anomálií

Disciplíny, z ktorých čerpá Data Mining

- databázy
- algoritmy
- štatistika
- vyhľadávanie informácií
- strojové učenie

Data science

- zastrešuje viacero oblastí ako je strojové učenie, data mining, big data, štatistika, vizualizácia dát, dátová analytika

Occamova britva

- vraví, že je potrebné používať čo najmenej predpokladov, eliminovať tie, ktoré nemenia predikcie
- ak sú všetky veci rovnaké, jednoduchšie riešenie je lepšie

Kliatba dimenzionality

- pridávanie dimenzií rapídne zvyšuje potrebné množstvo trénovacích dát na vytvorenie zmysluplného modelu
- pravidlo palca
    - aspoň 5 trénovacích inštancií na každú dimenziu

Spôsoby redukcie dimenzií

- výber atribútov (feature selection)
    - manuálne, ak dátam rozumieme
    - automatické metódy
        - regularizovaný strom
            - penalizujú použitie atribútov podobných už vybraným atribútom
        - t-test

        - štatistický test na porovnanie dvoch populácií
- transformácia
    - PCA
        - použitie súradnicovej sústavy, ktorá maximalizuje rozptyl
        - kubická zložitosť
    - náhodná projekcia
        - rýchlejšia ako PCA, ale dáva horšie výsledky
- normalizácia (škálovanie hodnôt do <0, 1>)
    - min-max
    - z-score (zero mean)
        - aritmetický priemer a štandardná odchýlka
        - výhoda oproti min-max, keď v budúcnosti príde číslo mimo hraníc min-max, tak s tým nie je problém
- vzorkovanie
- vyhladzovanie (odstránenie zašumených dát)
- agregácia (dni do mesiacov)
- zovšeobecnenie (nízkoúrovňové dáta nahradiť vyššie-úrovňovými konceptami)
- skonštruovanie atribútu z iných atribútov
- diskretizácia numerických atribútov

Predikcia

- na základe predošlých meraní hodnôt atribútov odhadujeme budúce hodnoty atribútov

Regresia

- učenie sa funkcie zobrazovania
- lineárna regresia
    - vyjadrenie triedy ako lineárna kombinácia atribútov (funkcia)
    - dokáže modelovať len lineárne závislosti
    - citlivá na outlierov
    - dáta musia byť nezávislé
    - je to jednoduchá metóda
- polynomiálna regresia
    - polynomiálne koeficienty vystupujú vo funkcii
    - môže sa preučiť, hlavne pri vysokých polynómoch – rieši sa regularizáciou – penalizáciou za vysoké rády
- metóda najmenších štvorcov
    - minimalizuje sumu štvorcových chýb

Klasifikácia

- učenie s učiteľom
- z trénovacej množiny sa naučí charakteristiky tried, do ktorých potom klasifikuje dáta
- Bayesovská klasifikácia
    - používa Bayesovo pravidlo $P(h|d) = P(d|h) * P(h) / P(d)$

    - je to podmienená pravdepodobnosť
    - naivný Bayesov klasifikátor
        - predpokladá, že všetky atribúty sú rovnako dôležité
        - predpokladá, že atribúty sú navzájom nezávislé
- rozhodovacie stromy
    - prístup rozdeľuj a panuj
    - orezávanie stromu – na odstránenie preučenia, vetiev reprezentujúcich šum
    - výhody
        - sú jednoduché na porozumenie a interpretáciu
        - sú podobné spôsobu uvažovania človeka
        - implicitne vykonávajú výber atribútov
        - nemajú žiadne predpoklady ohľadom charakteru dát
        - dokážu pracovať so všetkými typmi atribútov a aj s chýbajúcimi hodnotami
        - nemá takmer žiadne hyperparametre
    - nevýhody
        - sú nestabilné
        - oproti iným dávajú horšie výsledky, riešením sú náhodné lesy
        - nevedia zachytiť komplexné vzťahy medzi atribútmi
        - preučenie
    - Algoritmus ID3
        - aby sme vytvárali čo najjednoduchšie stromy
        - na rozhodnutie využíva meranie informácie v zlomkoch bitu podľa entrópie
        - nevie pracovať s chýbajúcimi hodnotami
        - nevie pracovať so spojitými atribútmi
        - nerobí orezávanie stromu
        - preferuje rozdelenie na veľa malých množín
    - Algoritmus C4.5
        - rieši chýbajúce hodnoty – ignoruje ich pri tvorbe stromu
        - rieši spojité hodnoty atribútov – rozdeľuje ich na intervaly
        - rieši orezávanie – pre-prunning, post-prunning
        - pri klasifikácii sa ide podľa iných atribútov
- k-najbližších susedov
    - na urýchlenie hľadania susedov je možné použiť mriežku, Voronoi diagram alebo kD strom
    - výhody
        - žiadne predpoklady o dátach
        - jednoduchý algoritmus
        - vhodný aj pre klasifikáciu aj pre regresiu
    - nevýhody
        - zložitosť
        - veľká pamäťová náročnosť
        - fáza predikcie je pomalá
        - citlivý na nerelevantné dáta
- PRISM
    - je to metóda klasifikačných pravidiel
    - pridáva testy do podmienky pravidla tak, aby sa maximalizovala pravdepodobnosť klasifikácie
- Neurónové siete

- typy:
  - dopredné
    - veľké množstvo parametrov
    - nezachytávajú štruktúru
    - trpia preučením
    - sú pamäťovo náročné
  - konvolučné
    - využívajú konvolúciu a pooling, čím znižujú počet parametrov
    - modelujú priestorové vzťahy
  - rekurentné
    - modelujú časovú informáciu
    - vhodné pre spracovanie sekvencií
- výhody
  - so zväčšujúcim množstvom dát zvyšujú svoju úspešnosť
  - zvyšovanie výkonu počítačov umožňuje vytvárať zaujímavé hlboké NN
  - aj po naučení je možné sieť ďalej učiť
  - jednoducho sa paralelizuje
  - riešia veľa problémov
- nevýhody
  - je to čierna skrinka – ťažké porozumieť a interpretovať
  - nezaručujú nájdenie globálneho optima
  - náchylné na preučenie
  - štruktúra grafu musí byť daná
  - veľké množstvo parametrov
  - vstupy musia byť numerické
  - hlboké NN – ich trénovanie trvá veľmi dlho
  - potrebujú veľké objemy dát
  - náročné na HW

- Support Vector Machine
  - je to lineárny diskriminátor s nulovou trénovacou chybou
  - je optimálny
  - vhodný pre klasifikáciu aj regresiu
  - vytvára deliacu nadrovinu, ktorá oddeľuje prvky od seba
  - hľadá deliacu nadrovinu s maximálnym okrajom od podporných vektorov (body)
  - pre lineárne neseparovateľné problémy je potrebný priemet do vyššej dimenzie
  - využíva kernelové funkcie na zníženie výpočtovej náročnosti
    - polynomiálny kernel
    - RBF kernel
    - sigmoidálny kernel
  - pri klasifikácii do viacerých tried sa problém pretransformuje na problém binárnej klasifikácie
    - jeden voči všetkým – n klasifikátorov, pre každú triedu je jeden klasifikátor, ktorý rozhodne, či dané dáta patria do tejto triedy alebo nie
    - jeden voči jednému – $n(n-1)/2$ klasifikátorov pre všetky dvojice tried, kde každá priradí inštanciu do jednej z dvoch tried
  - výhody
    - vhodný, keď dátam nerozumieme
    - vhodný pre text a obrázky

        - s vhodným kernelom je možné riešiť akokoľvek komplexný problém
        - nemá problém s uviaznutí v lokálnom extréme
        - dobre škálovateľný pre veľa rozmerné dáta
        - nie je citlivý na preučenie
        - dobre funguje aj pre malé trénovacie množiny
    - nevýhody
        - výber tej správnej kernelovej funkcie nie je jednoduchý
        - pre veľké množiny dát je potrebný dlhý čas na natrénovanie
        - je ťažké porozumieť a interpretovať model
        - nedáva dobré výsledky pre zašumené a prekrývajúce sa triedy

Zhlukovanie

- učenie bez učiteľa
- z dát rovno určí triedy, do ktorých dáta priradí
- single link – najkratšia vzdialenosť medzi inštanciami z dvoch zhlukov
- complete link – najdlhšia vzdialenosť medzi inštanciami z dvoch zhlukov
- hierarchické zhlukovanie
    - nie je potrebné dopredu určiť počet zhlukov
    - prístupy:
        - divízny – zhora nadol
        - aglomeratívny – zdola nahor
        - vytvárajú dendogram – kompletný strom
            - koreň je 1 zhluk obsahujúci všetky inštancie
            - list je zhluk obsahujúci jednu inštanciu
- podľa rozdeľovania
    - min. kostra
        - odstraňujú sa dlhé hrany z minimálnej kostry tak, že vzniknú zhluky pospájaných bodov
    - k-means
        - potrebuje určiť počet zhlukov
        - iteruje kým sa zhluky menia
        - inicializácia je náhodná – vyberie sa pár prvkov ako centroidy
        - následne sa centroidy počítajú z okna a sú to neexistujúce body, vypočítané
    - PAM
        - potrebuje určiť počet zhlukov
        - podobný k-means ale zhluk je reprezentovaný medoidom
        - vďaka medoidu sa dobre vysporiadáva s outliermi
    - CLARA
        - potrebuje určiť počet zhlukov
        - zlepšuje výpočtovú zložitosť PAM
        - aplikuje PAM na podmnožinu dát
        - nájdené medoidy používa ako medoidy v rámci celej množiny
- podľa hustoty
    - DBSCAN
        - jadrové body a ich okolie, jadrový bod ak obsahuje minPts bodov v okolí
        - parametre minPts, e – hraničná vzdialenosť

        - výhody:
            - nie je potrebné dopredu určiť počet zhlukov
            - je robustný voči outlierom
            - je navrhnutý tak aby pracoval s databázami ako Rtree
        - nevýhody:
            - nie je deterministický
            - kvalita závisí od zvolenej vzdialenostnej miery – obyčajne sa používa Euklidovská
            - je problematické zvoliť hraničnú vzdialenosť ak nerozumieme dátam
            - nie je vhodný pre množiny dát s rôznou hustotou zhlukov
    - OPTICS
        - zovšeobecnenie DBSCAN
        - na usporiadanie objektov sa používa prioritná halda
        - výsledkom je hierarchické zhlukovanie – vytvára dendogram
        - výhody
            - rieši problém s rozdielnymi hustotami zhlukov a nastavením hraničnej vzdialenosti
            - nastavuje sa ako parameter iba minPts
        - nevýhody
            - je výpočtovo náročnejší ako DBSCAN
- fuzzy zhlukovanie
    - C-MEANS
        - potrebuje určiť počet zhlukov
        - iteruje kým sa zhluky menia
        - ako k-means ale minimalizuje inú účelovú funkciu
        - iný výpočet centroidu ako k-means
        - iný výpočet príslušnosti k zhlukom oproti k-means

Sekvenčné dáta

- posuvné okno
    - prekonvertuje sekvenčný predikčný problém na klasifikačný alebo regresný problém
- rekurentné posuvné okno
    - predikovaná hodnota je použitá ako vstup pre ďalšiu predikciu
- Markovove modely
    - cieľom je predikovať nasledujúce pozorovanie v čase na základe n predošlých hodnôt – Markovove modely n-tého rádu
    - nedávne pozorovania sú informatívnejšie ako staršie pozorovania
    - skryté Markovove modely
        - model pre sekvencie nelimitovaný rádom s malým množstvom parametrov
        - zavádza diskrétne skryté premenné, ktoré tvoria Markovovu reťaz a určujú stav, ktorý vygeneruje pomocou emisných pravdepodobností pozorovania
        - emisné pravdepodobnosti – podmienené rozdelenie pozorovaných premenných
        - použitie:
            - dáta merané v čase
            - rozpoznávanie reči

- rozpoznávanie písma
- analýza DNA sekvencií

Dolovanie z Webu

- textové dokumenty sú reprezentované vektorom slov
- webové dokumenty sú reprezentované vektorom slov a HTML značkami, tzn. niektoré slová majú väčšiu váhu, napr. nadpisy
    - latentná sémantická analýza
        - vytvára množinu konceptov, ktoré sa vzťahujú ku slovám aj ku dokumentom
        - dáva lepšie výsledky ako tradičné metódy založené na kľúčových slovách
    - singulárny rozklad matice – dokáže aj zmazať nedôležité dimenzie
- dáta čistíme, identifikujeme používateľa (je to ťažké), identifikujeme session (server nepozná sekvenciu stránok, je potrebné cachovanie)
- dolovanie obsahu
    - vyhľadávanie
    - kategorizovanie
    - zhlukovanie
    - personalizácia
    - používajú sa crawlery
        - programy na automatické sťahovanie webových stránok
        - typy crawlerov
            - univerzálne
                - zhromažďujú všetky stránky bez ohľadu na obsah
            - preferenčné
                - zameriavajú sa na stránky len určitého typu
                - delí sa na focused (pre kategóriu) a topical (pre danú tému) prehľadávanie
        - základný crawlovací algoritmus
            - máme množinu stránok – seed pages
            - udržiavame si zoznam liniek na ešte neprezreté stránky
            - vyberieme stránku zo zoznamu, spracujeme ju a extrahujeme z nej linky do zoznamu, ale vymažeme linky s malou prioritou
            - končí, keď prezrie dané množstvo stránok
            - prehľadávanie môže byť do šírky alebo najlepšie prvé
- dolovanie štruktúry
    - PageRank
        - používa ho Google
        - meria prestíž stránky nezávisle na dopyte
        - prestíž stránky je proporcionálna sume prestíží stránok, ktoré na ňu odkazujú
        - zavádza aj tlmiaci faktor v podobe konštanty
        - uzly neodkazujúce sa na žiadnu stránku sú buď odstránené alebo odkazujú na jednu náhodnú stránku
    - HITS
        - na rozdiel od PageRank používa graf pre konkrétne dopyty

        - zavádza autority – stránky s kvalitnými informáciami a Hub – zoznam liniek na autority
        - výhoda – používa podgraf, ktorý je relevantný pre dopyt, tzn. menej chýb
        - nevýhoda – dĺžka trvania
- dolovanie používania
    - asociačné pravidlá
    - sekvenčné vzory

Asociačné pravidlá

- definujú špecifické vzťahy
- príklad: if temperature == cool then humidity = normal
- vhodné pre nenumerické, kategorické dáta
- obyčajne ich je veľké množstvo
- Podpora
    - support(X => Y) = |X U Y| / |N|, kde |..| - je kardinalita (počet)
    - support(X) = |X| / |N|
- Spoľahlivosť
    - confidence(X => Y) = support(X => Y) / support(X)
- kroky dolovania asociačných pravidiel
    - 1.) hľadanie frekventovaných množín položiek, ktoré majú väčšiu podporu ako prahová hodnota s
    - 2.) generovanie pravidiel z frekventovaných množín položiek
- frekventované množiny položiek
    - množiny položiek, ktoré majú väčšiu podporu ako prahová hodnota s
    - pre každé asociačné pravidlo X => Y, musí byť X U Y frekventovaná množina položiek
    - podmnožina každej frekventovanej množiny položiek je frekventovaná množina položiek
    - nachádzanie frekventovaných množín položiek je síce jednoduché, ale časovo náročné
    - Apriori algoritmus
        - frekventované množiny sú nadol uzavreté
        - ak vieme, že množina nie je frekventovaná, tak z nej netvoríme ďalšie podmnožiny
        - generuje kandidátov do šírky
        - nevýhody
            - množinu dát prechádza veľakrát
            - vysoká časová zložitosť
            - vysoká priestorová zložitosť – dá sa vyriešiť rozdelením na menšie množiny položiek a tie prehľadať samostatne
        - kroky algoritmu:
            - 1.) i = 0, vygeneruj všetky jednopoložkové množiny C1
            - 2.) ak Ci = 0, tak skonči
            - 3.) prejdením databázy sa spočítajú výskyty kandidátov a vylúčia sa tie, ktoré majú menšiu podporu ako prah s
            - 4.) i = i+1

                - 5.) vytvor kandidátov o veľkosti i kombináciou množín položiek z Fi-1 a choď na krok 2.
        - ECLAT
            - prehľadávanie kandidátov do hĺbky
            - vhodný aj na paralelné spracovanie
        - FP-growth (frequent patern)
            - vytvára FP strom
            - vie tak uložiť veľké databázy v kompaktnej štruktúre – FP strom
            - na dolovanie vzorov z FP-stromu používa prístup rozdeľuj a panuj
            - vlastnosti
                - nepočíta kandidátov
                - komprimovaná dátová štruktúra
                - prechádza databázu iba 2x
                - používa menej pamäte ako Apriori a ešte je aj rýchlejší
                - nepracuje dobre s veľkými databázami, keď sa FP strom nezmestí do pamäte
            - kroky vytvárania FP stromu:
                - 1.) prechádza databázu prvýkrát a vytvára zoznam frekventovaných prvkov usporiadaný podľa počtu výskytov
                - 2.) pre každú transakciu usporiada položky podľa usporiadaného zoznamu, prechádza databázu druhýkrát a vytvára FP-strom vkladaním transakcií usporiadaných podľa frekvencií
            - štruktúra FP stromu:
                - koreň stromu je označený null
                - každý vrchol má tri položky
                    - meno položky
                    - počet transakcií obsahujúcich vzor z cesty od koreňa do tohto uzla
                    - bočnú linku na nasledujúci prvok v strome s rovnakým menom
            - princíp fp-growth
                - najprv 1-prvkové vzory
                - pridávanie k vytvoreným množinám prvky, ktoré sú vyššie v strome
- Sekvenčné vzory
    - usporiadaná množina položiek, ktorá spĺňa danú podporu a je maximálna
    - AprioriAll
        - algoritmus:
        - nájdenie všetkých frekventovaných množín položiek
        - nahradenie každej originálnej transakcie množinou všetkých frekventovaných množín položiek
        - nájdenie sekvenčných vzorov
        - nájdenie maximálnych sekvencií
    - GSP
        - vychádza z Apriori
        - algoritmus:
        - nájdenie všetkých frekventovaných položiek (nie množín !!!)
        - repeat

                - generovanie kandidátov spájaním a orezávaním sekvencií – vždy o 1 dlhšie sekvencie
                - počítanie kandidátov
            - until už nie sú žiadne frekventované množiny alebo žiadny kandidáti neboli vygenerovaní
    - PrefixSpan
        - definuje prefix a projekciu pre sekvencie
        - negeneruje kandidátov
        - rekurzívne volá funkciu PrefixSpan() pre menšie podmnožiny
        - vstupom je databáza sekvencií a minSupport
        - výstupom je množina všetkých sekvenčných vzorov
        - je to rekurzívna metóda
        - prekonáva GSP
        - prehľadáva prefixové projekcie
        - doluje kompletnú množinu vzorov pričom negeneruje kandidátov
        - prefixová projekcie redukuje veľkosť databázy a vedie k účinnému spracovaniu

Meranie kvality asociačných pravidiel

Lift – pomer združenej pravdepodobnosti dvoch položiek a súčinu ich pravdepodobností

Conviction – podiel očakávanej frekvencie, že sa A vyskytuje bez B

Vyhodnocovanie metód dolovania znalostí

- vyhodnocovacie schémy
    - ako čo najlepšie využiť dáta pri ich rozdeľovaní do trénovacej a testovacej množiny
    - metódy
        - hold-out
            - rozdelí dáta na 2/3 trénovacia množina, 1/3 testovacia množina
            - testovacia množina nemusí byť reprezentatívna
            - stratifikácia – náhodný výber tak, aby každá trieda bola približne rovnako zastúpená v trénovacej aj testovacej množine
        - krížová validácia
            - rozdelenie množiny na n častí
            - n-krát opakovať trénovanie a testovanie – každá časť je použitá ako testovacia práve raz, ostatné tvoria trénovaciu množinu
            - obyčajne sa n=10
            - error rate – priemer zo všetkých behov + štandardná odchýlka
            - leave one out metóda krížovej validácie
                - trénovacia množina obsahuje n-1 inštancií
                - proces sa opakuje n-krát
                - výhody
                    - trénovacia množina je najväčšia možná
                    - proces netreba opakovať
                - nevýhody

        - rozdelenie nemôže byť stratifikované
        - časovo náročné
    - bootstrap
        - výber s opakovaním
        - dáta sa rozdelia do n častí
        - trénovacia množina sa vytvorí tak, že sa n-krát vyberie náhodne časť dát, tieto časti sa môžu opakovať
        - časti, ktoré neboli vybrané do trénovacej množiny, sa dajú do testovacej množiny
        - 0.632 bootstrap znamená, že 63.2 % dát je v trénovacej množine

Porovnanie metód na dolovanie znalostí

- krížová validácia
- t-test
- minimal description length
- informačné kritériá
    - merajú, ako dobre štatistický model opisuje dáta
    - založené na koncepte entrópie
    - odmeňuje model za dobré výsledky
    - penalizuje model za množstvo parametrov (pomáha proti preučeniu)
    - používajú sa pri výbere modelov
    - typy informačných kritérií
        - Akaike information criterion
            - použiteľné, ak log-likelihood alebo štvorec chyby sú použité na odhadovanie chyby
            - penalizuje modely s veľkým počtom parametrov
            - predstavuje balans medzi presnosťou a komplexitou modelu
            - čím menšia hodnota, tým lepší model
        - Bayesian information criterion
            - aplikovateľné pre modely maximalizujúce vierohodnosť
            - penalizuje modely s veľkým počtom parametrov
            - je to kritérium pre výber parametrického modelu s rôznym počtom parametrov
            - predpokladá, že rozdelenie dát je v exponenciálnej rovine

Vyhodnocovanie modelov predikujúcich pravdepodobnosti

- sensitivity: pomer chorých ľudí, ktorí majú pozitívny výsledok testu (tp)
- specificity: pomer zdravých ľudí, ktorí majú negatívny výsledok testu (1 - fp)
- ROC krivka
    - charakterizuje výkon binárneho klasifikačného systému. Je to 2D graf, ktorého osi sú FPR a TPR (false/true positive rate). V prípade, ak by ROC graf mal tvar x = y - binárny klasifikátor by vôbec nefungoval (lepšie povedane by sa nelíšil od náhodného klasifikátora). Čím je krivka viacej natiahnutá smerom k ľavému hornému rohu, tým je

výkon binárneho klasifikátora lepší. Oplatí sa spomenúť aj AUC (area under curve) - čím je táto plocha pod krivkou väčšia, tým je klasifikačný model kvalitnejší.
    - bod na krivke vyjadruje pomer medzi false pozitive rate a true positive rate
- Lift chart
    - vyjadruje ako dobre predikčný systém pracuje na zvolenej množine dát - napríklad ak je našim cieľom nájsť tých, ktorých bude zaujímať reklamná kampaň, tak aké percento pokrytia pozitívnych hodnôt bude vzhľadom na veľkosť vstupných dát.