PRIMENJENI ALGORITMI

Logistička regresija

Logistička regresija

- Predstavlja jednu od osnovnih tehnika mašinskog učenja (ML, Machine learning)
- U pitanju je statistički model koji koristi logističku funkciju za modelovanje sistema
- Koristi se za određivanje verovatnoće da neki podatak zadovoljava određeni uslov, kao što je na primer pripadnost nekoj grupi
- Može se koristiti za kalsifikaciju podataka u unapred određene grupe
- U slučaju klasifikacije sa dve grupe, daje verovatnoću da podatak pripada jednoj grupi. Ukoliko je verovatnoća veća od 0.5, tada se uzima da je podatak iz posmatrane grupe, a u suprotnoj podatak pripad drugoj grupi
- Najčešće se vrši klasifikacija podataka u dve grupe, a koristeći ovu osobinu možemo da implementiramo i klasifikaciju za više klasa

Logistički model

- U logističkoj regresiji, koristi se zavisna promenljiva Y koja može da ima dva stanja, npr 0 i 1, tacno i netačno, bolestan i zdrav, ...
- Posmatramo linearnu zavisnost $Y = B_1*X+B_0$. Tada tražimo verovatnoću p da je Y=1.
- Verovatnoću da je Y=1 pišemo kao $p = P(Y|1)$
- Ako koristimo Bernulijevu raspodelu, tada dobijamo da je $Y=\log(p/(1-p))$, pa je odavde $p = e^y/(1-e^y)$, tj. $p =1/(1-e^{-y})$
- Ovo se naziva sigmoid funkcija i po ovoj funkciji se računa verovatnoća da je Y=1, tj. da podatak pripada klasi 1. verovatnoća da podatak pripada klasi 0 je tada 1-p

Logistička regresija u Juliji

- Da bi koristili logističku regresiju u Juliji potrebni su nam:
  - *Paket GLM, isto kao i za linearnu regresiju*
  - *Funkcija glm(formula, data, family, link) koja vraća logistički regresor*
    - formula – formula na osnovu koje vršimo klasifikaciju, analogna formuli u linearnoj refresiji
    - data – podaci na osnovu kojih se kreira regresor
    - family – jedna od opcija: Bernoulli(), Binomial(), Gamma(), Normal() ili Poisson()
    - link – link koji zavisi od familije. Može biti CauchitLink(), CloglogLink(), IdentityLink(), InverseLink(), LogitLink(), LogLink(), ProbitLink() ili SqrtLink()
  - *Funkcija predict(regresor, dataFrame) koja vraća niz verovatnoća da određeni element dataFrame-a pripada klasi X1*
    - Ukoliko je verovatnoća veća od 0.5, najčešće kažemo da podatak pripada klasi 1, a u suprotnom pripada klasi 0
- Najčešće kombinacije familije i linka su Bernoulli() i LogitLink(), Binomial() i LogitLink(), Gamma() i InverseLink(), Normal() i InverseLink(), Poisson() i LogLink()

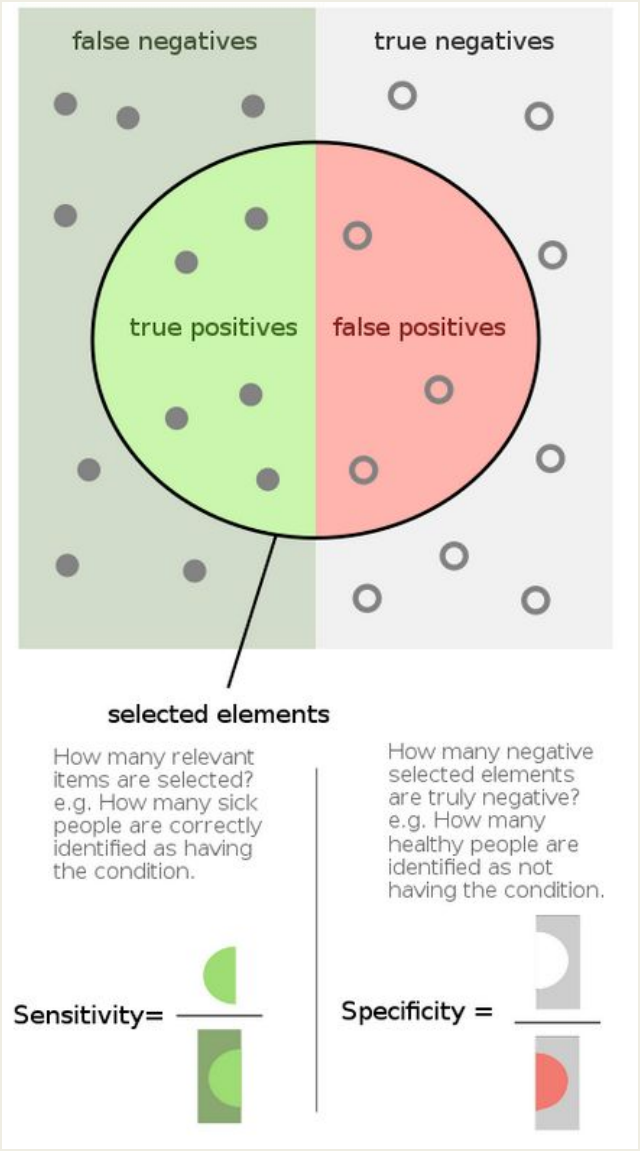
Confusion matrix

- Neka smo izvršili predviđanje klasa za neki skup podataka. Tada definišemo:
  - *TP (True positives) – broj tačnih klasifikacija da je podatak iz klase 1 (1->1)*
  - *TN (True negatives) – broj tačnih klasifikacija da podatak nije iz klase 1 (0->0)*
  - *FP (False positives) – broj netačnih klasifikacija da je podatak iz klase 1 (0->1)*
  - *FN (False negatives) – broj netačnih klasifikacija da podatak nije u kalsi 1 (1->0)*
  - *P (Positives, P=TP+FN) – broj elemenata klase 1*
  - *N (negatives, N=TN+FP) - broj elemenata koji nisu u klasi 1*

Actual class \ Predicted class	P	N
P	TP	FN
N	FP	TN

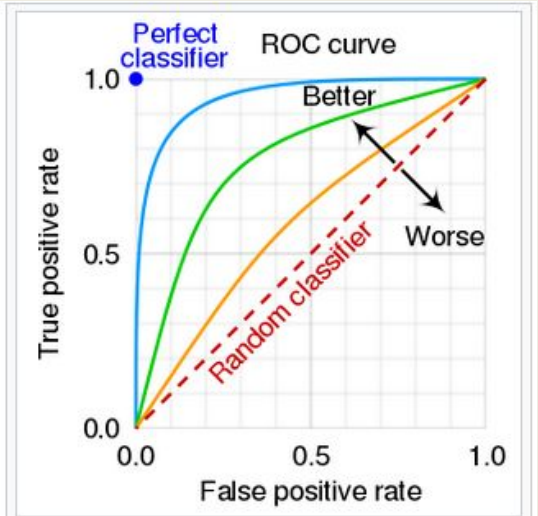
Ocena kvaliteta klasifikacije

- Kvalitet klasifikacije ocenjujemo pomoću:
  - ***accuracy*** *(preciznost) = (TP+TN)/(TP+TN+FP+FN) = (TP+TN)/(P+N), preciznost (kvalitet) klasifikacije*
  - ***sensitivity*** *(osetljivost, True positive rates) = TP/(TP+FN) = TP/P , procenat tačno klasifikovanih podataka klase 1*
  - ***specificity*** *(specifičnost, True negative rates) = TN/(TN+FP) = TN/N, procenat tačno klasifikovanih podataka klase 0*
- Ove vrednosti treba da su što veće, tj. što bliže vrednosti 1. Potrebno je da su veće od 0.5, a poželjno je da su veće od 0.7 ili čak 0.9
- Ukoliko su osetljivost i specificnost približne vrednosti, tada je klasifikator dobar za klasifikaciju obe klase, u suprotnom nije dobar za obe klase

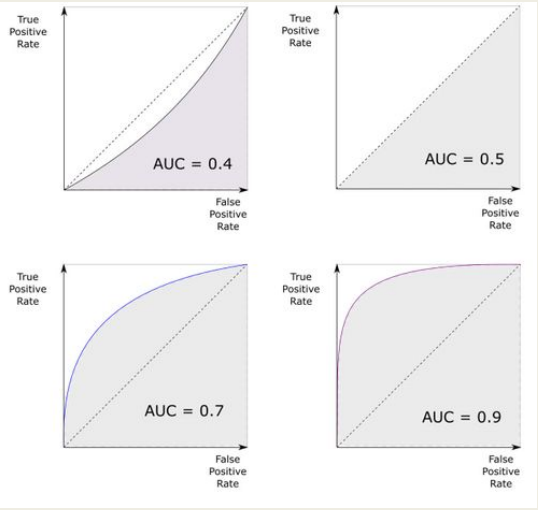


ROC kriva

- ROC (Receiver operating characteristic) kriva je kriva na garfikonu koja pokazuje mogućnosti binarnog kalsifikatora za različita podešavanja sistema
- U suštini, predstavlja odnos između **osetljivosti** (TPR, True positive rate, TP/P) i **greške** (FPR, False positive rate, FP/N = **1-sensitivity**)
- AUC vrednost kalsifikatora predstavlja površinu ispod krive. Ova provršina treba da je što bliža vrendosti 1
- AUC predstavlja objektivnu meru kvaliteta klasifikatora



The ROC space for a "better" and "worse" classifier.



Primer 1: Klasifikacija tačaka

- Primer 1. Neka je dat skup tačaka u ravni svojim X i Y koordinarama, kao i svojom bojom koja može biti 0 ili 1. Napraviti klasifikator koji može da klasifikuje tačke u jednu od dve grupe, određene ovim bojama. Pri tome:
  - *Podatke učitati iz fajla tacke1000.csv*
  - *Koristiti podelu skupa za obuku i trening u razmeri 80:20*
  - *Izračunati i ispisati precisnost, osetljivoist i specifičnost rezultata*
  - *Nacrtati ROC krivu*
  - *Izračunati i ispisati površinu ispod ROC krive i odrediti kvalitet klasifikatora*

Primer 2: Klasifikacija pacijenata

- Primer 2. Neka je dat skup podataka o pacijentima sa sledećim podacima
  - *Podaci:*
    - visina – visina pacijenta u cm
    - težina – težina pacijenta u kg
    - dbp – donji krvni pritisak (diastolic blood pressure) mmHg
    - sbp - gornji krvni pritisak (systolic blood pressure) mmHg
    - bolest – da li pacijent boluje ili ne od dijabetesa (0-ne, 1-da)
  - *Zadatak:*
    - Podatke učitati iz fajla pacijenti1000.csv
    - Koristiti podelu skupa za obuku i trening u razmeri 80:20
    - Izračunati koliko pacijenata ima, tj. nema dijabetes