Preprilagodjavanje modela (eng. Overfitting)

Kada model daje odlične rezultate na skupu na kome je izgradjen, a na test skupu pravi znatno veće greške, kažemo da je došlo do preprilagodjavanja. Ovo se može ilustrovati na primeru učenika koji odredjenu lekciju nauči napamet. Dakle, on tu lekciju zna potpuno precizno (trening skup), ali nije naučio i razumeo principe koji se odnose na tu oblast, pa će njegovo znanje na bilo kom sličnom gradivu (test skup) biti jako loše.

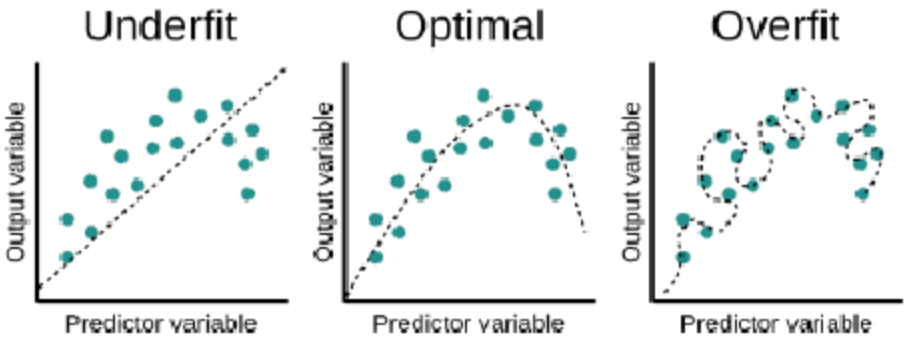
Do ove pojave može doći zbog neadekvatnog izbora modela, prevelikog broja parametara, nereprezentativnog skupa za obučavanje itd. Dakle, da bismo uočili pojavu preprilagodjavanja neophodno je uporediti preciznost na trening i test skupu. Ako je preciznost na trening skupu mnogo veća to je znak da je došlo do preprilagodjavanja.

Zašto npr. u regresiji ne koristimo Lagranžov interpolacioni polinom, kada njime možemo dobiti tačnu predikciju svake tačke iz trening skupa? Zato što takvo savršeno predvidjanje na trening skupu ne garantuje sjajna predvidjanja i u budućnosti. Štaviše, uglavnom rezultira velikim oscilacijama u predikcijama novih tačaka. Ovo je klasičan primer u kome se naš model preprilagodio podacima. Zato je korisnije dozvoliti modelu da pravi male greške na skupu na kome se izgradjuje, tako da se može i na novim uzorcima uspešno koristiti. Za više detalja potražiti: **bias and variance tradeoff.**

Ako je preciznost modela mala i na test i na trening skupu onda kažemo da je došlo do tzv. potprilagodjavanja (eng. underfitting).

Da bismo izbegli preprilagodjavanje koristimo validacioni skup, pomoću kojeg biramo optimalne vrednosti hiperparametara modela. Dakle, trening skup koristimo za nalaženje parametara modela, a validacioni skup za nalaženje hiperparametara modela.

Da li još nešto možemo da učinimo kako bismo se bar delimično zaštitili od preprilagodjavanja? Odgovor je regularizacija.



Regularizacija

Kod parametarskih modela kao što su linearna i logistička regresija ocene parametara dobili smo traženjem ekstremuma nekih funkcija. U linearnoj regresiji tražili smo minimum zbira kvadrata reziduala, a u logističkoj maksimum funkcije verodostojnosti, ili još češće maksimum logaritma funkcije verodostojnosti. Kako je većina optimizacionih algoritama konstruisana za problem minimizacije, ocene parametara logističke regresije su dobijene minimizovanjem logaritma funkcije verodostojnosti pomnoženog sa -1. U tom slučaju uz praćenje oznaka sa prethodnog časa možemo reći da su ocene parametara nekog od ovih modela dobijene na sledeći način:

$$(\beta_0, \beta) = \underset{(\beta_0,\beta)}{argmin} \sum_{i=1}^{n} L(y_i, \beta_0 + \beta^T x_i)$$

gde je $\beta = (\beta_1, \ldots, \beta_p)^T$, $x_i = (x_{i1}, \ldots, x_{ip})^T$ i:
$L(y_i, \beta_0 + \beta^T x_i) = (y_i - \beta_0 - \beta^T x_i)^2$ u slučaju linearne regresije
$L(y_i, \beta_0 + \beta^T x_i) = -(y_i(\beta_0 + \beta^T x_i) - ln(1 + e^{\beta_0 + \beta^T x_i}))$ u slučaju logističke regresije
Od preprilagodjavanja se možemo zaštititi uvodjenjem nenegativnog hiperparametra λ i traženjem parametara modela minimizacijom sledeće funkcije:

$$\sum_{i=1}^{n} L(y_i, \beta_0 + \beta^T x_i) + \lambda \|\beta\|$$

gde je norma koja se primenjuje na β najčešće l_1 ili l_2 norma. Kakva je uloga hiperparametra λ? λ ne dozvoljava pojedinačnim koeficijentima da postanu previše veliki, odnosno ne dozvoljava da se nekom prediktoru dodeli prevelika težina, a samim tim se sprečava preprilagodjavanje trening podacima. Pitanje je kako odabrati λ tako da ne bude ni premalo ni preveliko. Za premalo λ regularizacioni izraz postaje zanemarljiv u odnosu na $\sum_{i=1}^{n} L(y_i, \beta_0 + \beta^T x_i)$, pa će doći do overfitting-a, a za preveliko λ izraz $\sum_{i=1}^{n} L(y_i, \beta_0 + \beta^T x_i)$ postaje zanemarljiv, pa će doći do underfitting-a.
Dakle, λ je hiperparametar modela i njegovu vrednost izračunavamo na validacionom skupu. To radimo tako što uzimamo vrednosti za λ iz nekog fiksiranog skupa, zatim pravimo model na trening skupu i računamo preciznost na trening i validacionom skupu. Taj postupak ponavljamo za svaku vrednost λ iz tog fiksiranog skupa. Optimalno λ će biti ono za koje je preciznost na validacionom skupu najveća.
Postavlja se i pitanje koju normu koristiti? Poznat teorijski rezultat kaže da se pri korišćenju l_1 norme može dogoditi da neki parametri modela budu jednaki 0. Dakle, korišćenje l_1 norme, poznatije kao **LASSO** regularizacija, može poslužiti i za selekciju prediktora. Pri korišćenju l_2 norme, poznatije kao **RIDGE** regularizacija, nijedan od parametara ne može biti jednak nuli. Iako l_1 ima tu prednost, nekada l_2 daje bolje rezultate.
Važno je istaći da regularizaciju ne treba nužno koristiti kada je baza velikog obima, jer tada su šanse da dodje do preprilagodjavanja manje i može doći do bespotrebnog smanjivanja preciznosti modela.

Evaluacija modela

Pomenuli smo da je nakon izgradnje modela potrebno izvršiti proveru njegovog kvaliteta. Kako to možemo učiniti? Pozabavimo se problemom klasifikacije. Neretko se dešava da za modele koji daju odličan procenat tačnih predvidjanja pomislimo da su zaista korisni. Medjutim, takva mera kvaliteta klasifikatora nije uvek najinformativnija. Kada se jedna kategorija daleko češće pojavljuje u skupu za obučavanje od ostalih, većina modela teži ka tome da skoro sve tačke klasifikuje kao tu dominantnu kategoriju. Dakle, u takvoj situaciji modeli nisu naučili dobro da razlikuju kategorije, već ih sve vide kao tu dominantnu. Ako je dominantna kategorija dominantna i u skupu za testiranje, model će sigurno imati visoku uspesnost u terminu procenta tačnih predvidjanja, čak iako sve ostale kategorije ne klasifikuje dobro. Jednostavno, drugih kategorija nema dovoljno da bismo kroz ovu meru kvaliteta videli da model ipak nije toliko sjajan.
Opišimo jedan očigledan primer kada je bitno koju meru kvaliteta koristimo. Klasifikujemo tumor u dve grupe: maligni i benigni. Da li je greška koju pravimo kada kažemo za benigni tumor da je maligni ista kao kada za maligni kažemo da je benigni? Greška u drugom slučaju je jako opasnija.
Postoji više mera kvaliteta klasifikatora i skoro sve ih možemo pročitati iz tzv. matrice konfuzije.

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

Najčešće korišćene mere kvaliteta su:

accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$

precision = $\frac{TP}{TP+FP}$

tpr (true positive rate) = $\frac{TP}{TP+FN}$

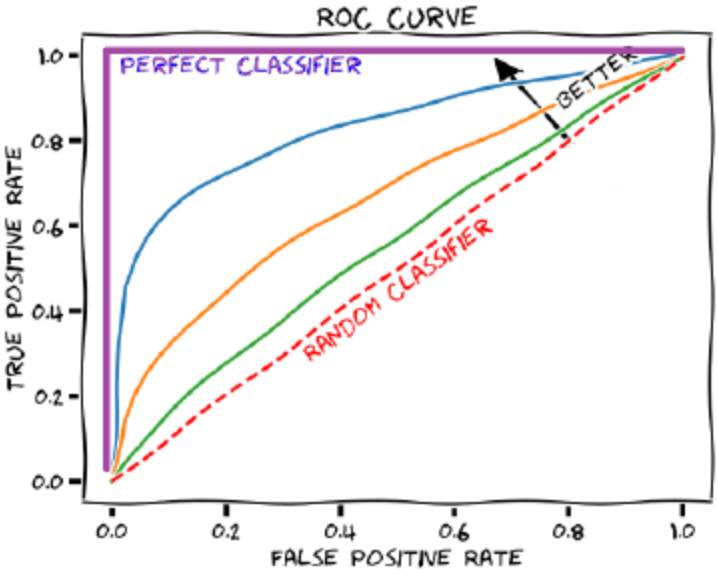
tnr (true negative rate) = $\frac{TN}{TN+FP}$

Svaka od ovih mera je osetljiva na nebalansiranost kategorija. Zato uvodimo meru koja se zove F score i jednaka je harmonijskoj sredini precision-a i tpr-a, odnosno

F score = $\frac{2 \cdot precision \cdot tpr}{precision+tpr}$

Još jedna mera koja je otporna na nebalansiranost kategorija je površina ispod ROC krive.

ROC kriva je skup tačaka $(1 - tnr(r), tpr(r))$ ucrtanih za različite vrednosti praga $r \in [0, 1]$. Ta kriva počinje u $(0, 0)$, a završava se u $(1, 1)$. Kako varira vrednost praga, proveravamo da li kriva pravi "grbu" ka gornjem levom uglu, jer je to znak da se $tnr(r)$ i $tpr(r)$ približavaju ka 1.



Dakle, što je veća izbočina, klasifikator je bolji u smislu prepoznavanja obe kategorije. Zbog toga se često posmatra veličina AUC (area under the curve), koja meri površinu ispod ROC krive. Što je AUC bliže jedinici, to je klasifikacija uspešnija. Ako je AUC blizu 0.5 to je znak da je model loš.

Jedna lepa karakterizacija vrednosti AUC: AUC je jednak verovatnoći da slučajno odabrani element iz pozitivne kategorije ima veći skor (skor se izračunava modelom) od slučajno odabranog elementa iz negativne kategorije. Dakle, savršeni klasifikator svim pozitivnim elementima dodeljuje veće skorove nego negativnim, pa će njegov AUC biti jednak 1.