Model Ensemble

Hanno 2 cose in comune:

* Costruiscono modelli predittivi multipli e diversi che vengono ri-ponderati o ricampionati
* Combinano le previsioni di questi modelli con una media o votazione

Bagging - Bootstrap Aggregating

Crea diversi modelli su diversi campioni del dataset originale. I campioni sono i **bootstrap samples**. Le differenze tra i campioni creeranno diversità tra i modelli dell’ensemble.

Idea base: **Sample with replacement**: Estrazione casuale dal training set e:

* Ogni elemento estratto può essere ri-estratto e appartenere a un altro campione
* Ogni campione ha la stessa DIM del training set originale
* Ogni campione è probabile che manchi circa ⅓ dei data points (Out of bag sample)

Algoritmo:

**Bagging(D,T, A)** - train an ensemble of models from bootstrap samples

*Input: dataset D, ensemble size T, learning algorithm A*

*Output: ensemble of models whose predictions are to be combined by voting or averaging*

*for t = 1 to T do:*

*build a bootstrap sample Di from D by sampling |D| data points with replacement*

*run A con Di ti produce a model Mi*

*return {Mi | 1 ≤ t ≤ T}*

Come previsione usa:

* Voting: Vinca la classe prevista dalla maggioranza dei modelli
* Averaging: Si prende la media dei valori

Bagging utile con modelli ad albero, molto sensibili alle variazioni dei training data. Dividendo i dati in modo casuale→gli alberi danno risultati diversi con alta varianza→utile per modelli soggetti a overfitting.

Quando viene applicato con substance sampling, ogni albero è costruito da un diverso sottoinsieme casuale di features → Random Forest. Nella RF, date F features, viene specificato un numero f ≪ F in modo che a ogni nodo le features siano selezionate casualmente.

**Random Forest**:

* Creo T bootstrap samples dal training set
* Seleziono f features random
* Costruzione del decision tree
* Costruzione dei decision trees per ogni bootstrap samples
  + Classificatore: Prende il majority vote
  + Regressore: Media dei risultati

Se vi è un predittore molto forte rispetto a tutti gli altri, Tutti gli alberi selezionano il predittore forte come root tree e verranno tutti simili. Quindi per la classificazione: f = F1/2, per la regressione: f = ⅓ F.

OOB Error Estimation

Tecnica di convalida per misurare l’errore di previsione di Random Forest. OOB sono gli esempi che non compaiono in ogni bootstrap sample. Permettono di convalidare un modello senza usare un validation set separato.

1. Training: Ogni albero è addestrato sul suo rispettivo bootstrap sample
2. Prediction with OOB: Per ogni istanza nel dataset originale, si identificano gli alberi che non hanno incluso questa istanza. Questi alberi si usano per prevedere il risultato di questa istanza
3. Aggregation: Si raccolgono tutte le previsioni e si usa o il voting o l’averaging.
4. Error Calculation: Confronto le previsioni OOB aggregate con i valori effettivi per calcolare l’errore.
   1. Classificazione: È la proporzione delle istanza classificate male
   2. Regressione: Si usa Mae/ Mse.

In entrambi i casi, bassi valori → Prestazioni migliori.

Boosting

Idea: Formare i classificatori in sequenza, e ognuno deve essere debole. Voting/ Averaging per combinare i risultati.

* La ponderazione è usata per dare più influenza ai più performanti
* Boosting è iterativo. Ogni nuovo modello è influenzato dalle prestazioni di quelli costruiti precedentemente.
* Ogni nuovo modello dovrebbe concentrarsi su quelle istanze in cui l’insieme ha un rendimento scarso.
* Ogni volta che l’errore sui dati supera o è uguale 1/2 , la procedura elimina il classificatore e non esegue iterazioni. Anche quando l’errore diventa 0, perché tutti i pesi diventano 0.

Regola pratica dell’apprendimento automatico: I modelli a basso bias hanno alta varianza e viceversa.

3 motivi per cui un modello può classificare male:

* Istanze con valori uguali, ma classe diversa.
* Manca l’espressività per un concetto target. Es: dati non linearmente separabili.
* Varianza: Un modello ha alta varianza se il suo confine decisionale dipende fortemente dai dati di formazione.

Il Bagging è una tecnica di riduzione della varianza, il Boosting per la riduzione del bias.