Agregacja bootstrapowa (bagging)/ Lasy losowe

Cel: próba rozwiązania problemu niestabilności drzew decyzyjnych.

Jak?: Tworzymy kilka podzbiorów poprzez losowanie ze zwracaniem. Na otrzymanych grupach tworzymy drzewa, na koniec uśredniając wyniki (np. poprzez głosowanie).

Zalety: Nie musimy dzielić na zbiór uczący i walidacyjny. Ograniczamy overfitting w porównaniu ze zwykłą metodą pojedynczego drzewa, dodatkowo zwiększając uniwersalność wyników.

Wady: Wybór elementów do podzbiorów jest losowy, przez co część obserwacji nie będzie należeć do grupy uczącej (krosswalidacja OOB). Przez losowość klasyfikatory mogą być podobne przez co nie uzyskamy satysfakcjonującego uogólnienia. Dodatkowo jeśli jakaś zmienna jest silnie skorelowana z badanym zjawiskiem to będzie wybierana jako pierwszy test, zaniedbując te słabiej skorelowane zmienne.

Różnica między nimi: „Zasada działania algorytmu Lasu losowego jest bardzo podobna do agregacji bootstrapowej, z tą różnicą, że do poszczególnych modeli drzew decyzyjnych wykorzystywane są nie wszystkie dostępne predyktory, a jedynie ich część.”

Adaptive Boosting ADABOOST

Cel: łączenie pojedynczych modeli w zespoły.

Jak?: Drzewa nie tworzymy równolegle, lecz są one ze sobą związane. Na początku mamy bazowy klasyfikator, po czym skupiamy się na tych obiektach, które zostały źle zaklasyfikowane i dostosowujemy wagi w kolejnej iteracji. Te błędne będą zwiększone, a te poprawne zmniejszone. Itd. przez określoną liczbę iteracji.

Zalety: Eliminuje się błędy poprzedniego modelu.

Wady: Kiedy chcemy dopasować wagi używając metody spadku gradientu model staje się przetrenowany, występuje overfitting, brak zdolności uogólniania wyników.

Extreme Gradient Boosting XGBOOST

Cel: ulepszenie powyższych metod przez algorytm wzmacniania gradientowego.

Jak?: Bazując na Adaboost wprowadza składnik regulacji, czyli kary za zbyt dużą liczbę liści. Buduje się go z dwóch części: pierwszy to funkcja kosztu (odpowiada za minimalizację błędu), drugi to regularyzacja (zapobiega przetrenowaniu).

Zaleta: Zwiększa dokładność przewidywania.