Veri madenciliği

Malik SARI

Bilgisayar MÜH.

Boostting:

Yükseltme *sıralı bir öğrenme* tekniğidir. Algoritma, tüm eğitim seti ile birlikte bir modelin eğitimi ile çalışır ve sonraki modeller, ilk modelin kalıntı hata değerlerinin yerleştirilmesiyle oluşturulur. Bu şekilde, Yükseltme önceki model tarafından kötü bir şekilde tahmin edilen gözlemlere daha fazla ağırlık vermeye çalışır. Modellerin sırası oluşturulduktan sonra, modellerin yaptığı tahminler doğruluk puanlarına göre ağırlıklandırılır ve sonuçlar nihai bir tahmin oluşturmak için birleştirilir. Genellikle Boostting tekniğinde kullanılan modeller XGBoost (Extreme Gradient Boost), GBM (Gradient Boost Machine), ADABoost (Adaptive Boost) vb.

Destek Vektör Makineleri:

Destek Vektör Makineleri, temel olarak iki sınıfa ait verileri birbirinden en uygun şekilde ayırmak için kullanılır. Bunun için karar sınırları yada diğer bir ifadeyle hiper düzlemler belirlenir. DVM'ler günümüzde yüz tanıma sistemlerinden, ses analizine kadar birçok sınıflandırma probleminde kullanılmaktadırlar.

KNN (K nearest neighborhood, en yakın k komşu):

Sınıflandırmada (classification) kullanılan bu algoritmaya göre sınıflandırma sırasında çıkarılan özelliklerden (feature extraction), sınıflandırılmak istenen yeni bireyin daha önceki bireylerden k tanesine yakınlığına bakılmasıdır. Örneğin k = 3 için yeni bir eleman sınıflandırılmak istensin. bu durumda eski sınıflandırılmış elemanlardan en yakın 3 tanesi alınır. Bu elamanlar hangi sınıfa dahilse, yeni eleman da o sınıfa dahil edilir. Mesafe hesabından genelde öklit mesafesi (euclid distance) kullanılabilir.

Yerine koyarak örnekleme (Bagging):

Bagging (Bootstrapping Aggregating) metodu, var olan bir eğitim setinden yeni eğitim setleri türeterek temel öğreniciyi yeniden eğiten bir yöntemdir. Her bir veri seti için eğitilen temel öğrenicinin belirli bir test örneği üzerindeki kararları ortalama alınarak hesaplanır. Bagging'de amaç yeni veri setleri türeterek farklılıkları oluşturmak ve bu sayede toplam sınıflandırma başarısını artırmaktır.

Bu yöntemde eğitim setinden yaklaşık olarak %63,2 kadar orijinal örnek rastgele alınır ve alınan örneklerden bazıları çoğaltılarak (resampling) eğitim seti %100'e tamamlanır. Bu yöntemle birbirinden farklı bir miktar eğitim seti elde edilir. Her eğitim seti aynı temel öğreniciye

uygulanır ve alınan kararlar ağırlıklı oylama yöntemiyle birleştirilir [15]. Eğitim setinden %63,2 kadar örnek seçilmesindeki neden formülle açıklanır:

$$-(1 - \frac{1}{n})^{n-1} = 0{,}368 \approx e^{-1}$$

Burada n işlem sayısını veya eleman sayısını gösterir. n değeri sonsuza giderken doğal logaritma tabanı olan e sayısının tersi elde edilmiş olur. n sonsuza giderken Bu yöntemle orijinal eğitim kümesinden yaklaşık olarak %36'sı büyük ihtimalle hiçbir zaman seçilmemiş olacaktır. Bagging yönteminde seçilmeyenlerin seçilmesini sağlamak için eğitim setinden %63,2 kadar örnek $(1 - 0{,}368 = 0.632)$ rastgele seçilir ve yeni bir eğitim seti oluşturulur.

Bu ödev için sklearn kütüphanesi kullandım sklearn kütüphanesini iyi derecede kullanabiliyorum artık **Kaynaklar:**

1-https://www.datacamp.com/community/tutorials/k-means-clustering-python

2-https://www.datacamp.com/community/tutorials/svm-classification-scikit-learn-python

3-http://bilgisayarkavramlari.sadievrenseker.com/2008/11/17/knn-k-nearest-neighborhood-enyakin-k-komsu/

4-https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.KFold.html

5-https://medium.com/@k.ulgen90/makine-%C3%B6%C4%9Frenimib%C3%B6l%C3%BCm-4-destek-vekt%C3%B6r-makineleri-2f8010824054 6-https://www.datacamp.com/community/tutorials/adaboost-classifier-python