Uçtan Uca Makine Öğrenmesi Sistemi

Uçtan uca bir makine öğrenmesi projesinin temel adımları:

1- Projenin genel çerçevesinin çizilmesi.
2- Veriyi elde etme
3- Veriyi inceleme ve görselleştirme.
4- Veriyi makine öğrenmesi için uygun hale getirme
5- Model seçimi ve eğitimi
6- Modeli optimize etme
7- Çözümü sunma
8- Çözümün canlıya alınması, izlenmesi, bakım yapılması.

Projenin Çerçevesinin Çizilmesi

Projenin çerçevesi çizilirken ilk dikkat edilmesi gereken projenin amacının ne olduğudur. Projenin amacı kullanılacak modelleri, kullanılacak başarı metriğini belirlemek açısından önemlidir.

İkinci önemli nokta problemin hali hazırda nasıl bir yöntem ile çözüldüğüdür. Bu bilgi sayesinde çözüm ile alakalı pek çok bilgi elde edilir.

Proje ile ilgili bilgiler elde edildikten sonra performans metriği belirlenir. Regresyon problemlerinde genellikle Root Mean Square Error ya da Mean Absolute Error metrik olarak kullanılırken sınıflandırma projelerinde Accuracy Score ya da F1 Score kullanılır.

Veriyi Elde Etme

Veriyi elde etme sürecinde öncelikle problemin çözümü için gerekli olabilecek verilerin listesi çıkarılır. Verinin nerede olduğu, boyutunun büyüklüğü, yasal bir sakınca içerip içermediği ve erişim yetkileri belirlenir.

Veri ilgili yerden çekilir, manipüle etmeye uygun bir hale getirilir, hassas bilgiler silinir ya da maskelenir. Veri eğitim ve test seti şeklinde iki parçaya ayrılır.

Veriyi inceleme

Veri inceleme aşamasında öncelikle verinin bir kopyasını alıp onun üzerinde çalışmak gerekir. Böylece orijinal veri korunur ve bir yanlışlık olduğunda tüm veriyi baştan çekmeye ya da okumaya gerek kalmaz.

Verideki featureların tipleri, aykırı değerleri, kayıp verileri, gürültüleri, dağılımları, kullanışlılığı incelenir. Denetimli öğrenme projelerinde label/target belirlenir.

Veri görselleştirilerek featureların korelasyonu incelenir böylece yüksek korelasyonlu featurelardan bazıları veriden çıkarılarak boyut indirgenir.

Bazı gerekli görülen özellikle döngüsel verilerde (yılın günü, haftanın günü... gibi) dönüşüm uygulanır. Sinus, Cosinus... vb. Problemin çözümü için dışarıdan elde edilebilecek ekstra veriler belirlenir.

Veriyi Makine Öğrenmesine Hazırlama

Verinin bir kopyası üzerinde çalışılmalı ve yapılan tüm işlemler fonksiyonlaştırılmalıdır. Bu sayede test seti oluşturma, sonraki projelerde kullanma açısından kolaylık sağlanır.

Verideki ayrık değerler temizlenir. Ya tamamen veriden kaldırılır ya da ortalama vs ile doldurulur.

Veriden gereksiz, modele katkı sağlamayan featurelar çıkarılır.

Bazı featurelara dönüşüm uygulanır, bazı featurelar birleştirilerek yeni featurelar oluşturulur. Bazı featurelar normalize edilir.

Model Seçimi Ve Eğitimi

Model seçimi için pek çok tipte model hızlıca eğitilir. (Lineer, Naive Bayes, SVM, Random Forest, Neural Networks... vb.)

Bu modellerin performansları ölçülür ve karşılaştırılır. Bu sırada her modele K Fold CV uygulanır.

Her algoritma için en etkili çıkan featurelar ve modellerin karşılaştığı hatalar belirlenir.

Özellik mühendisliği ve seçimi uygulanır.

Yukarıdaki adımlar bir kaç defa tekrarlandıktan sonra en iyi 3 model belirlenir.

Sistemi Optimize Etme

Hiperparametre optimizasyonu için Cross-validation yöntemi kullanılır. Çok az sayıda hiperparametre varsa grid search yerine random search kullanılır. Hiperparametre sayısı fazlaysa Bayesian optimizasyon yaklaşımı kullanılır.

Çeşitli yöntemlerle modelleri kombine halk getirmek genelde daha iyi sonuç verir. Bunun için kullanılan yöntemlere ensemble yöntemleri denir.

Modelden tam anlamıyla emin olunca test set üzerinde genelleştirme hatası hesaplanır.