Makine Öğrenmesinde Model Seçimi :

Makine öğrenmesi bir örüntü bulma işidir ve bu görev istatistiksel ve matematiksel yöntemlerle gerçekleştirilir. Örneğin elimizde modellememiz gereken veriler var ve bu veriler için en uygun modeli seçmemiz gerekiyor. Bu konuda ne yapmalıyız? En yüksek puanlı metriğe sahip modeli mi yoksa dağılıma en uygun olanı mı seçmeliyiz?

Burada yapılacak işe özgü ya da kullanılmaya devam edilmesi gereken bir model varsa model seçimi kısmı önemini yitiriyor. Doğadaki tüm veriler mutlaka bir dağılımı takip eder. Buna örnek olarak, bir kuşun cıvıl cıvıl aralığı veya kullanıcıların bir reklamı tıklama sıklığı verilebilir. Bizim için önemli olan dağıtıma uygun bir model seçmek ve seçtiğimiz modeli optimize etmektir.

Model seçiminin otomatik bir süreç olmadığını ve genellikle içgüdüsel olarak ele alınması gerekir.

Model seçerken kullandığımız metrikler dağılımı yansıtmalıdır. Bununla birlikte, denetimli yöntemlerde, verileri genellikle belirli bir oranda test ve eğitim verisine böleriz. Peki ya tren bölümünün dağılımı ile test bölümünün dağılımı farklıysa? Metrikler bizi yanıltabilir mi? Cevap elbette evet. Ölçümlerin verilere bütünsel olarak bakmasını istiyorsanız tüm verileri doğrulamanız gerekir. Bir saniyede 100 megabayt veriyi doğrulamak kolaydır, peki 100 terabayt veriyi nasıl doğrularız? Bir örnek seçerek. Bir veriden rastgele seçilen 1000 satırlık tren verisi ile bu verileri özetleyebilen 1000 satırlık bir örnek arasında büyük bir fark vardır. 1000 satır tren verisi, tüm verinin dağılımı hakkında bir fikir vermeyebilir. Bununla birlikte, doğru seçilmiş bir örnek, verilerin dağılımı hakkında doğru bilgiler içerebilir.