**Homework 1**

1) How would you define Machine Learning?

Makine Öğrenimi, bilgisayarların verilerden öğrenebilmelerini sağlayan bir programlama bilimidir.

Arthur Samuel, 1959, makine öğrenimini, bilgisayarlara açıkça programlanmadan öğrenme yeteneği veren çalışma alanı olarak tanımlamıştır.

2) What are the differences between Supervised and Unsupervised Learning? Specify example 3 algorithms for each of these.

Supervised Learning: Tipik bir denetimli öğrenmenin görevi sınıflandırmadır. İstenmeyen posta filtresi bunun güzel bir örneğidir: Birçok spam ve ham örnekleriyle eğitim gerçekleştiği zaman yeni e-postaları nasıl sınıflandıracağını da öğrenmesi gerekir.

Ek olarak tahminler adı verilen bir dizi özellik (kilometre, yaş, marka vb.) verildiğinde, bir otomobilin fiyatı gibi bir sayısal değeri tahmin etmek de supervised öğrenmeye örnektir. Bu tür bir göreve regresyon denir. Sistemi eğitmek için, ona hem tahmin ediciler hem de etiketler (yani fiyatları) dahil olmak üzere birçok araba örneği vermeniz gerekir.

* k-Nearest Neighbors • Linear Regression • Logistic Regression algoritma örnekleridir.

Unsupervised Learning: eğitim verileri etiketlenmez. Sistem öğretmensiz öğrenmeye çalışır. Örneğin, blogunuzun ziyaretçileri hakkında çok fazla veriye sahip olduğunuzu varsayalım. Benzer ziyaretçi gruplarını tespit etmeye çalışmak için bir kümeleme algoritması çalıştırmak isteyebilirsiniz. Hiçbir noktada, bir ziyaretçinin hangi gruba ait olduğunu algoritmaya söylemezsiniz: Bu bağlantıları sizin yardımınız olmadan bulur. Örneğin, ziyaretçilerinizin% 40'ının çizgi romanları seven ve genellikle blogunuzu akşamları okuyan erkekler, % 20'sinin ise hafta sonları ziyaret eden genç bilimkurgu severler olduğunu fark edebilir. Hiyerarşik bir kümeleme algoritması kullanırsanız, bu aynı zamanda her grubu daha küçük gruplara ayırabilir. Bu, her grup için gönderilerinizi hedeflemenize yardımcı olabilir.

• Clustering — K-Means

• Anomaly detection and novelty detection — One-class SVM

* Visualization and dimensionality reduction — Principal Component Analysis (PCA) algoritma örnekleridir.

3) What are the test and validation set, and why would you want to use them?

Validation set, bir sınıflandırıcının hiper parametrelerini (yani mimarisini) ayarlamak için kullanılan örneklerin bir veri kümesidir. Test seti gibi eğitim veri setiyle aynı olasılık dağılımını takip etmelidir.

test set, eğitim veri kümesinden bağımsız olan, ancak eğitim veri kümesiyle aynı olasılık dağılımını izleyen bir veri kümesidir.

Aşırı uyumu önlemek için, herhangi bir sınıflandırma parametresinin ayarlanması gerektiğinde, eğitim ve test veri setlerine ek olarak bir doğrulama veri setine sahip olmak gerekir.

Test setindeki örneklerin sınıflandırmalarını tahmin etmek için son model kullanılır. Bu tahminler, modelin doğruluğunu değerlendirmek için örneklerin gerçek sınıflandırmalarıyla karşılaştırılır.

4) What are the main preprocessing steps? Explain them in detail. Why we need to prepare our data?

Makine Öğreniminde veri ön işleme, verilerden anlamlı içgörülerin çıkarılmasını teşvik etmek için verilerin kalitesini artırmaya yardımcı olan çok önemli bir adımdır. Makine Öğreniminde veri ön işleme, Makine Öğrenimi modellerini oluşturmaya ve eğitmeye uygun hale getirmek için ham verileri hazırlama (temizleme ve düzenleme) tekniğini ifade eder. Basit bir deyişle, Makine Öğreniminde veri ön işleme, ham verileri anlaşılır ve okunabilir bir biçime dönüştüren bir veri madenciliği tekniğidir.

Makine Öğreniminde Veri Ön İşleme Adımları

1. Veri kümesini edinin

2. Tüm önemli kütüphaneleri içe aktarın

3. Veri kümesini içe aktarın

4. Eksik değerlerin belirlenmesi ve ele alınması

5. Kategorik verilerin kodlanması

6. Veri kümesini bölme

7. Özellik ölçekleme

5) How you can explore and analyse countionus and discrete variables?

Countionus: Ölçülebilir değerlerdir, zaman gibi.

discrete variables: sayılabilir değerlerdir, haftanın günleri gibi.

6) Analyse the plot given below. (What is the plot and variable type, check the distribution and make comment about how you can preproccess it.)

Plot: barplot

Variable type: numerical data (Countionus)

* Öncelikle verilerdeki veri içerisinde bulunan her bir sütun içerisinde kaç adet örnek var, bu verilerin veri tipleri nedir ve veri tiplerinin tüm veri içerisinde dağılımı nedir gibi veriyi tanımak için gerekli temel soruların cevaplanması gerekir. Eksik veriler mean veya medyana göre doldurulur. Daha sonra veriler kodlanır.
* Bu grafikte tek değişkenli bir verimiz var ve verinin sıklığı histogram grafiği ile verilmiş. Grafiği linear regresyona göre yorumlamak gerekirse tahminleme yapılacak bir çizgi belirlenmelidir ve taç yapraklarının genişliğinin belirlediğimiz çizgiye ne kadar uzak veya yakın olduğuna bakılır. Buna göre modelimizi verimize göre tahminlenebilen bir hale getirmiş oluruz.