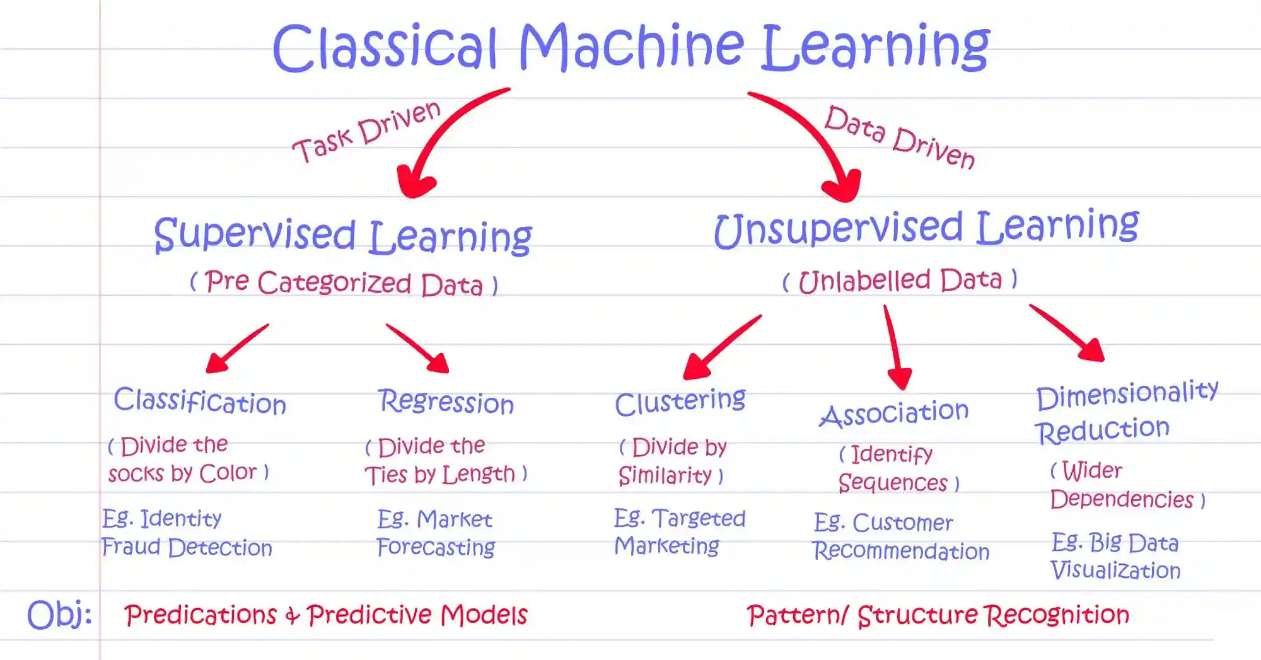
**ML SESSION -1**



**independent ile dependent farkı:**

modelde prediction için kullanılan özellikler independent , bunlara istinaden bağıl değişkende dependent oluyor.

featurelar  birbirinden etkilenmez, ama target bu independentlarda yapılan değişiklerden etkilenir.

biz bu featureları kullanalarak o target a ulaşacak rule(kuralı) modeller kullanarak buluyoruz.

Arabaların markası, modeli, motor hacmi, [km.si](http://km.si/), yaşı vs. gibi birbirlerinden **bağımsız**olan değişkenlere *independent değişken* deniyor.  
Bu independent değişkenlere **bağlı**olarak değeri değişen **price**değişkenine *dependent değişken* deniyor.

**Examples of supervised machine learning include:**

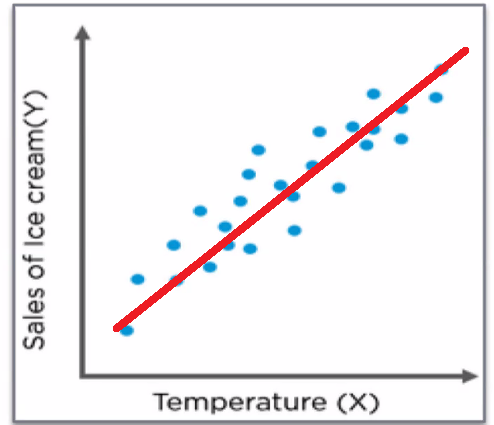
* Classification, identifying input data as part of a learned group.
* Regression, predicting outcomes from continuously changing data.

**Examples of unsupervised machine learning include:**

* Clustering, grouping together data points with similar data.
* Association, understanding how certain data features connect with other features.

(Linear regression is a machine learning algorithm that tries to capture the linear relationship between features and the continuous value that we want to predict. Let’s say you want to predict the price that you need to pay for a taxi. You can do so by first collecting data that shows the taxi price with respect to the corresponding travel distance.

**Linear regression:** featurelar ve predict etmek istediğimiz continuous valuelar arasında **linear bir ilişki** yakalamaya çalışan bir ML algoritmasıdır.



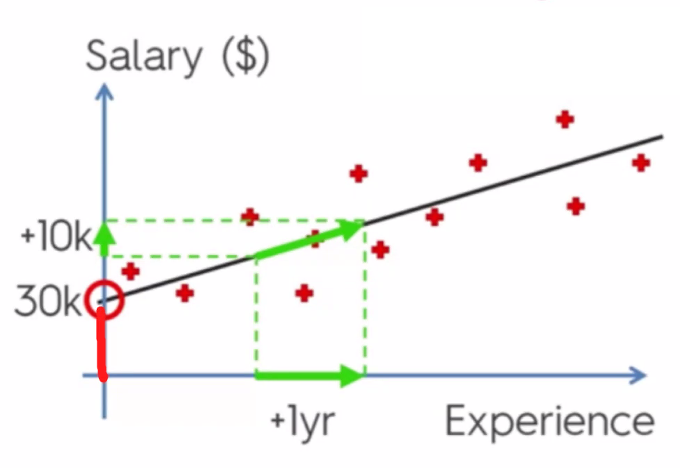
Şu kırmızı lineer çizgiyi veren fonksiyon, (gerçek) observation noktalarını en iyi şekilde tanımlıyor.

**LINEAR REGRESSION**

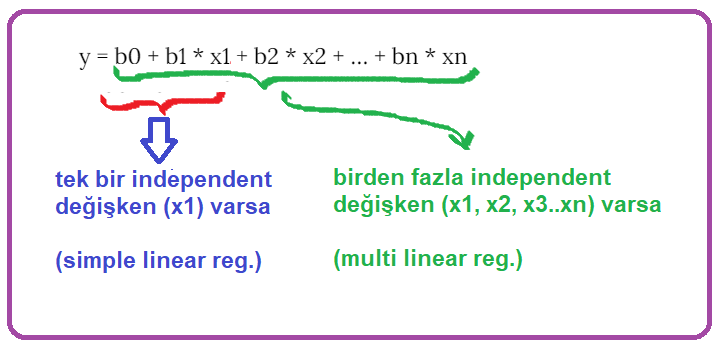
Doğrusal regresyon modelleri tipik olarak, bağımlı değişkenin tahmin edilen ve gözlemlenen değerleri arasındaki kare farklarının toplamını en aza indiren bir optimizasyon algoritması kullanılarak uygundur.

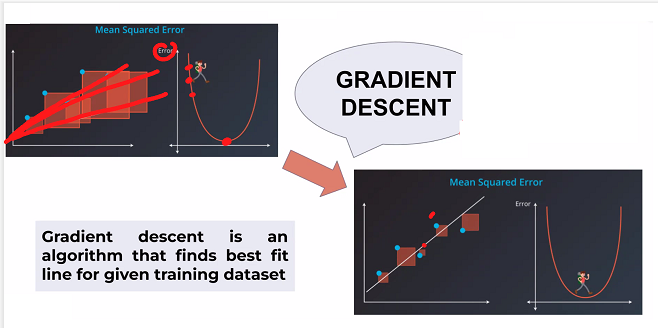
Ortaya çıkan **model**, değişkenler arasındaki ilişkiyi açıklayan **matematiksel bir denklemdir.**

Bu denklemi kullanan modelleme ile ne yapıyoruz? Bağımsız değişkenlerin değerlerine dayalı olarak bağımlı değişkenin değeri hakkında tahminlerde bulunmak için kullanıyoruz.



**Line**üzerindeki değerler **tahmin edilen değerler oluyor.**  
Etrafındaki noktalar ise**gerçek observation'lar oluyor.**

****

****

**Doğru Line’ı bulurken;**

Tek bir independent feature varsa, yani simple linear regresyonda en yakın kareler toplamı yöntemi kullanılıyor.

Fakat birden fazla independent feature varsa, yani multi linear regresyonda gradient descent yöntemi kullanılıyor.

**Gradient Descent:**

**Optimizasyon** ; bir fonksiyonu minimize ya da maksimize etmek amacı ile gerçek veya tamsayı değerlerini bir fonksiyona yerleştirerek sistematik olarak bir problemi incelemek veya çözmek anlamına gelmektedir.

***Gradient Descent****algoritması**en populer optimizasyon algoritmalarından birisidir.****Bu algoritmanın amacı****, fonksiyona ait parametreleri devamlı güncelleyerek fonksiyonun en minimum değerine ulaşmaktır.*

Dereceli azalma , bir fonksiyon üzerinde rastgele bir noktadan başlayan ve o işlevin en düşük noktasına ulaşıncaya kadar eğiminde adım adım ilerleyen yinelemeli bir algoritmadır.

**Gradient Descent’e Örnek;**

Gradient descent’i gerçek bir problemmiş gibi düşünecek olursak;

Bir dağın zirvesinden aşağı deniz seviyesine inmeniz gerektiğini hayal edin fakat görme yeteneğiniz çok az ve etraf da karanlık bu durumda aşağı nasıl inerdiniz?

Öncelikle küçük adımlarla ilerlemeye gayret gösteririz, çok hızlı inmek tehlikeli olabilir ayağınız kayıp düşebilirsiniz amacınız genellikle küçük adımlarla inmek olur.

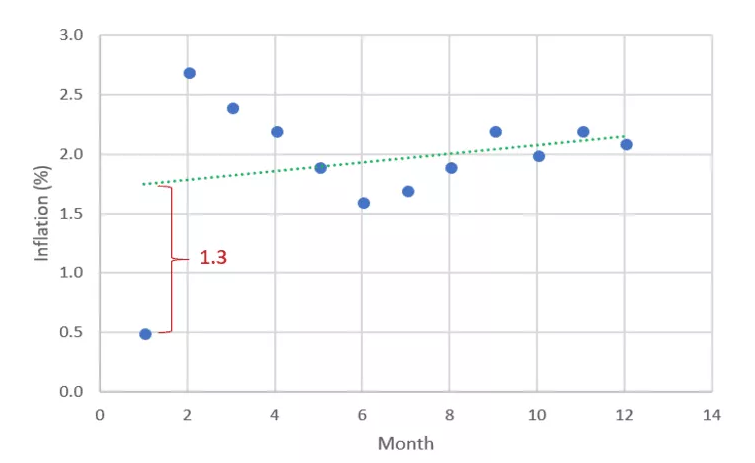
Gradient Descent’te ise durum buna çok benzerdir, bir problemin çözümü için en uygun noktayı bulmak bunun içinse küçük adımlarla ilerleyerek hedefe varmaktır. Bu örnekte attığınız adım sayısı iterasyonun sayısına denk gelir. Aşağı inmek için kaç adım attınız? Hedefe ulaşmak için kaç iterasyona ihtiyacınız oldu ?

Adımların büyüklüğü gradient descentte ki learning rate’e yani öğrenme oranına denk gelmektedir. Dağdan aşağıya doğru indiğiniz sırada belki 10-15 adım sonrasında daha rahat adım atmaya başladınız bunun nedeni artık bir sonraki adımınız için bir tahminleme yeteneği kazandınız 10 adımda az çok öğrenme gerçekleşti aslında.

İşte gradient descentte de öğrenme işlemi bu şekilde gerçekleşmektedir. Birkaç adım iterasyona girdikten sonra öğrenme ve optimizasyon işlemleri yapılmaya başlanmaktadır.

Bu örnek , Gradient Descent algoritmasını somut olarak düşünüp makine öğrenmesinde kullanılan bu algoritma ile optimize edilmeye çalışılan modelde bize en doğru sonucu verecek olan parametreleri bulmamız konusunda daha anlaşılırlık sağlıyor.

**Residuals**



**Residuals**in a statistical or machine learning model are the **differences between observed and predicted values of data**. They are a diagnostic measure used when assessing the quality of a model. They are also known as **errors**

**Tüm degerler simetrik olmamasina ragmen toplam nasil sifir oluyor?**

Simetrik olmasına gerek yok. Train datasında hata toplamı sıfır olacak şekilde best fit line'ı çiziyor. Bu nedenle de hata toplamı sıfır oluyor.

10 tane +1 birimlik hataya karşılık 1 tane -10 birimlik hata toplam hatayı sıfırlar ama bu simetrik değildir.

**ML SESSION -2**

**Hangi lineer regresyonu ne zaman kullanırız?**

* Amaç bağımsız değişkenlerin değerlerine bağlı olarak bağımlı değişkenin (target) değeri hakkında tahmin yapmaksa, simple (basit) veya multi linear regresyon uygun olabilir.
* Fakat dikkate alınması gereken bir diğer faktör; bağımlı ve bağımsız değişkenler arasındaki ilişkinin gücüdür. İlişki güçlü ve lineer (doğrusal) ise, simple veya multi linear regresyon uygun olabilir. Fakat ilişki daha karmaşıksa, **polinom regresyon**daha iyi bir seçim olacaktır.
* Ayrıca; verileriniz tek bağımsız değişkenli bir veri ise simple linear regresyon, birden çok bağımsız değişkenli veriler için multi linear regresyon kullanılır.
* Bunların dışında; iki regresyon daha var. **Polinom regresyon**, düz bir çizgiden daha karmaşık bir ilişkiyi (eğri büğrü) izleyen veriler için uygundur ve **lojistik regresyon**, *ikili sonucu*olan veriler için uygundur.

**Hata Metrikleri:**

**MAE** modelin genel performansında yapılan büyük prediction hatalarını öne çıkarmaz

**MSE** = hataların varyansı

**RMSE** = hataların Std sapması

MSE genelde iki regression modelini karşılaştırmada kullanılır hataları insanın gözüne soktuğu için

RMSE’de önce karesini alıp sonra karekökünü aldığı için seviyeyi MAE ile eşitlemiş oluyor. iki modeli birbirinden ayırmaksa amacımız, MSE bunun için biçilmiş kaftan.

MAE ve RMSE değerleri birbirine yakın olmalı. Arasında fark büyük ise modelinizde harbi büyük hatalar vardır.

mae, mse, rmse bu üç metriğede aynı anda bakmak en sağlıklısı.

**Random state:**

Datayı split ederken aynı dataya sahip herkesin farklı bilgisayarda aynı satırların train ve test taraflarını aynı şekilde gitmesini sağlayan bir anahtar diyelim. bir nevi random durumu sabitliyorsunuz.

her çalıştırdığınızda aynı dataları train ve test datasına ayırsın ve böylece aynı model sonuçlarını görelim diye kullanıyoruz. pandas taki seed metodu ile aynı işi yapıyor gibi düşünebilirsiniz.

**Correlation:**

0.7 - 0.8 den yüksek korelasyona yüksek koralasyon diyebiliriz. Bu, iki değişken arasında güçlü bir ilişki olduğunu gösterir.

**LinearRegression() sınıfı:**

Model eğitim işlemi sırasında, verilerinizi kullanarak nesneyi "**fit**" etmeniz gerekiyor, yani **verilerinizi modele uydurmanız** gerekiyor. Bu nedenle, **LinearRegression() sınıfı**, modeli oluşturmak ve eğitmek için kullanılan bir araçtır. Ancak bu işlemleri gerçekleştirmek için **bir nesne üzerinden çağrılması**gerekmektedir.

Bu nedenle LinearRegression() sınıfını bir değişkene atamak suretiyle bir nesne oluşturuyoruz. Ve işlemleri hep o nesne üzerinden yürütüyoruz.

**Bir modelde yuzde kac hataya kadar iyidir diyebiliyoruz?**

Datadan dataya değişir. Problemden probleme değişir. Standart bir değeri yoktur. problem vardır 1 birim büyük hatadır, problem vardır 1000 normaldir.

**R2 score:**

* Regresyon **modelinin verileri ne kadar iyi açıkladığını**ölçmek için kullanılan bir istatistiksel ölçüttür.
* R2 score, gerçek değerlerin değişkenliği ile tahmin edilen değerlerin değişkenliği arasındaki oranı ifade eder.
* 0 ile 1 arasında değer alır.
* R2 score 1'e yaklaştıkça, model gerçek verileri daha iyi açıklıyor demektir.
* Eğer R2 score değeri 1 ise, model tüm varyasyonu açıklar, yani gerçek veriler ile tahmini veriler arasında tam bir uyum vardır. Ancak, R2 score değeri 0'a yaklaştıkça, model gerçek verileri açıklamada başarısız olur.

R2 score, bir regresyon modelinin performansını ölçmek için kullanılan en yaygın ölçümlerden biridir.

R2 score değeri, modelin doğru şekilde yapılandırılıp yapılandırılmadığını, kullanılan değişkenlerin uygun olup olmadığını ve outlierların varlığı gibi faktörleri değerlendirmeye yardımcı olur.

R2 score değeri, farklı regresyon modelleri arasındaki performans farklarını da karşılaştırmak için kullanılabilir. Ancak, R2 score'nun yalnızca bir regresyon modelinin performansını ölçmek için kullanılabileceği unutulmamalıdır ve modelin diğer özellikleri de göz önünde bulundurulmalıdır.

**R2 Score negatif çıkar mı?**

R2 skoru, bir regresyon modelinin ne kadar iyi uyduğunu gösteren bir ölçüttür. R2 skoru, 1'e yaklaştıkça modelin iyiliği artar ve 0'a yaklaştıkça modelin kötüleşir. Negatif bir R2 skoru, modelin gerçek değerlerden daha kötü tahminler yaptığını gösterir ve bu, gerçek hayatta nadiren kullanışlı bir durumdur.

Bir örnek vermek gerekirse; bir ev fiyatlarını tahmin etmek için bir regresyon modeli oluşturduğunuzu varsayalım. Modeliniz, ev fiyatlarını tahmin etmek için çeşitli faktörleri kullanır, örneğin evin konumu, büyüklüğü ve özellikleri gibi. Ancak modeliniz yanlışlıkla tüm girdileri tersine çevirir ve tahminler yapar. Yani, modeliniz büyük evlerin ucuz olduğunu, küçük evlerin pahalı olduğunu ve iyi bir konumdaki evlerin kötü olduğunu tahmin eder. Bu durumda, R2 skoru negatif olacaktır çünkü modeliniz gerçek değerlerden çok daha kötü tahminler yapıyor.

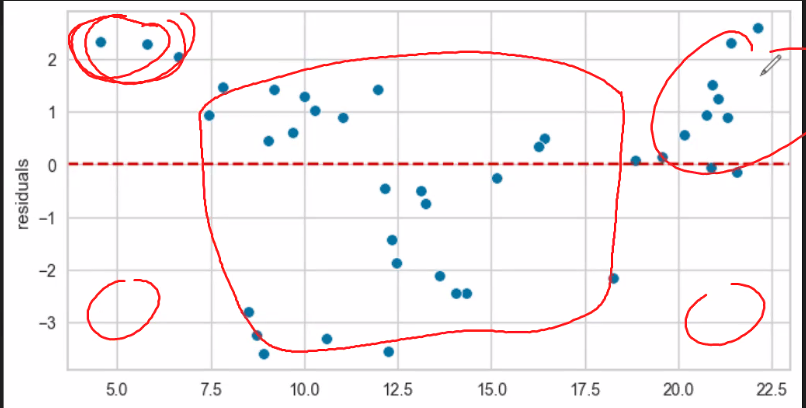
Tabii ki, bu örnek gerçek hayatta nadiren karşılaşılacak bir durumdur ve genellikle negatif R2 skorları, modelin doğru olmayan bir şekilde uygulanmasından veya veri setinin yanlış şekilde işlenmesinden kaynaklanır. Bu nedenle, modelinizin sonuçlarını yorumlamadan önce, özellikle negatif bir R2 skoru gibi anormal sonuçları dikkate almanız gerektiğini unutmayın

**Regresyon modelindeki tv radio newspaper in katsayilarinin istatistiksel anlamliligina bakmamiz gerekiyor mu?**

İstatistik ile ML farklı. Regresyon aynı ama biz datayı açıklamaya çalışmıyoruz. ML için prediction önemli

**İki uctaki degerlerde karsilik olmadigindan outliera yakin degerlerde hata var diyebilir miyiz?**

Evet diyebiliriz.



Sol ve sağ üstteki uç değerlerin line’ın altında karşılıkları olmadığından o değerlerin outlier olduğunu söyleyebiliriz.

**Residualler uzerinde shapiro testi vb uygulanabilir mi yani gerek var mı?**

Gerekli değil. Siz predictionlar ve metriclere yoğunlaşın

**regresyona ne demiştik x in y e etkisini açıklar demiştik değil mi? x bir birim artartsa y coef kadar artar..**