Yapay Öğrenme Ders Notları
Hazırlayan: Barış Baykant Alagöz
İnönü Üniversitesi, Bilgisayar Mühendisliği, 2023-2024 Bahar Dönemi Lisans Dersi
Sorularınız ve düzeltme önerileriniz için e-mail: baykant.alagoz@inonu.edu.tr

Öğrenme Felsefesi, Günümüzde Makine Öğrenmesi ve Temel Kavramları, Değerlendirme Metrikleri (Ölçütleri) II

Makine Öğrenmesinde Denetimli (Supervised) Öğrenme Problemleri:

1- Regresyon (Bağlanım) Problemleri: Regresyon analizi, iki veya daha fazla değişken arasındaki ilişkiyi belirleyen analiz yöntemidir. Örneğin, yıllık yağış miktarına göre yıllık kayısı üretimi arasındaki ilişki bir regresyon analizi ile incelenebilir. Regresyon analizinin en temel yöntemi olan en küçük kareler yöntemi 1805 yılında Adrien Marie Legendre tarafından ortaya konduğu bilinmektedir (Wikipedia, Regresyon analzi). Bu dönemde gök cisimlerinin güneş etrafında yörüngelerinin belirlenmesi ve modellenmesi için kullanılmıştır. Regresyon işleminde veri kümesindeki verileri en az hata ile temsil edebilen bir fonksiyon belirlenir. Günümüzde makine öğrenmesinde, regresyon problemi bir veri kümesini en az hata ile temsil edebilen bir matematiksel modelin elde edilmesi problemi olarak görülmektedir. Özünde verilere fonksiyon eğrilerinin uydurulması işlemidir. Elde edilen fonksiyonlar ile

(i) veri kümesi ihmal edilebilir düzeyde bir hata ile yeniden üretilebileceği (Verileri ezberlenmesi),

(ii) veri kümesinde olmayan verileri ise en az hata ile tahmin edilebileceği (Verilerden genelleme) varsayılır. Dolayısı ile, modelleme ve tahmin problemlerinin çözümü için makine öğrenmesinde regresyon analizi uygulanır.

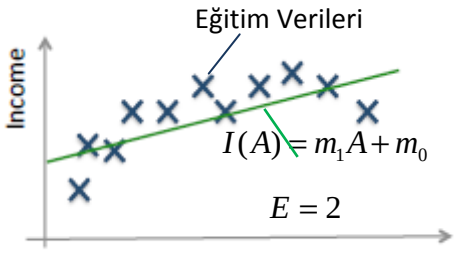
Regresyon problemlerinin çözümünde en sık kullanılan hata ifadesi karesel hatadır. Burada amaç, regresyon modelinin en az karesel hata ile sonuç üretebilmesidir. Karesel hata ifadesini sonlu bir veri kümesi için yazalım:

Elde edilen p adet veri noktaları $\{(x_1,d_1),(x_2,d_2),..,(x_p,d_p)\}$ olsun. Regresyon problemleri aşağıda verilen karesel hatayı minimize ederek çözülebilir.

$$E=(d_1-f(x_1))^2+(d_2-f(x_2))^2+...+(d_p-f(x_{Pi}))^2=\sum_{i=1}^{p}(\underbrace{d_i-f(x_i)}_{e_i})^2=\sum_{i=1}^{p}e_i^{\,2}$$

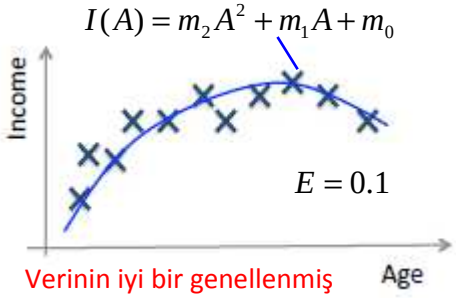
Burada, E değişkeni karesel hatayı, $f(x_i)$ ise veriyi modellemek için kullanılan fonksiyonu ifade eder. Ayrıca veri başına hata, $e_i=d_i-f(x_i)$, bir x_i verisi için istenen(ölçülen) değer d_i ile fonksiyon çıkışı $y_i=f(x_i)$ arasındaki farktır ve (x_i,d_i) verisi için bir öğrenme hatasını ifade eder. Regresyon probleminin nümerik çözümü, E karesel hatanın minimize edilmesi ($\min E$) ile sağlanır. Dolayısı ile regresyon problemi çözümü için "$\min E$" optimizasyon problemi çözülür.

Örnek: Aşağıda kişilerin yaşlarına (age) bağlı olarak gelir düzeyleri (Income) verileri grafiklerde çarpı (X) ile temsil edilmiştir. Bu örnekte seçilen regresyon modelinin karmaşıklığının veri temsildeki rolü incelenmiştir. Burada verileri en az hata ile temsil eden matematiksel modeli elde etmek istiyoruz. Böylece bir regresyon problemi çözümü elde ediyoruz. Model olarak doğru denklemi ve polinomları kullanıyoruz. Makine öğrenmesi açısından verilerin model ile temsili genellenme ve ezberleme durumlarını gündeme getirir.



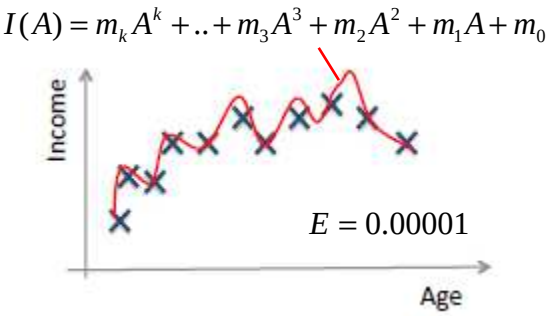
Verinin aşırı genellenmiş bir temsili (Under Fitting)

Veriler bir doğru denklemi ile temsil edilmek istenirse bir lineer regresyon problemi ifade eder. Yandaki grafikte bir doğru ifade eden $I(A)=m_1A+m_0$ fonksiyonun m_1 ve m_0 katsayılarının en az hata ile verileri temsil etmesi istenir. Bu katsayıların belirlenme süreci makine öğrenmesinde eğitim süreci olarak görülür. Yandaki grafik incelendiğinde model basit olmasına karşın veriye çok iyi uymadığı (under fitting) görülebilir. Bu duruma verinin aşırı genellenmiş bir temsili diyebiliriz. E karesel hata bu durum için büyük olabilmektedir ve aşırı genelleme durumunu ifade eder. Daha iyi bir genelleme için model karmaşıklığı artırılmalıdır.



Verinin iyi bir genellenmiş temsili (Good Fitting)

Veriler bir doğrusal olmayan fonksiyon ile temsil edilmek istenirse, örneğin bir polinom ile ifade edilebilir. Yandaki grafikteki veriler bir 2 derece polinom $I(A)=m_2A^2+m_1A+m_0$ ile ifade edilebilir. Burada m_2 , m_1 ve m_0 katsayıları en az hata ile belirlenir ise regresyon problemi çözülmüş olur. Bu süreç makine öğrenmesi için eğitim sürecidir. Böylece eğrinin verileri en az hata ile temsili sağlanır.
Yandaki grafikte verinin bu polinom ile iyi temsil edilebildiği görünüyor. Dolayısı ile $I(A)=m_2A^2+m_1A+m_0$ modeli iyi bir genelleme için uygundur çünkü verinin değişim trendi iyi karakterize edebilmiştir. İyi bir genelleme sağlanmış görünüyor. Karesel hata sıfır değildir ve yeterince düşüktür. BU MODEL BU VERİ İÇİN UYGUNDUR.



Verinin yetersiz genellemiş temsili (Over Fitting) yani veriler ezberlenmiştir. Bu durum aşırı öğrenme olarakda anılır.

Yukarıda doğrusal olmayan fonksiyon ile temsil edilmek edildiğinde iyi sonuç alındığını gördük. Peki, yüksek dereceli polinom ile iyi ifade edilebilir mi? Daha karmaşık bir model k. Derece polinom olan $I(A)=m_kA^k+..+m_3A^3+m_2A^2+m_1A+m_0$ denklemi (k>>2) ile ifade edilebilir. Burada fonksiyonun m_k, m_2 , m_1 ve m_0 katsayıları en az hata ile temsil etmesi için belirlenir. Fonksiyon bütün veri noktalarını ifade etmiş ve hata sıfıra yakın elde edilebilmiştir. Bu hali ile aşırı fit (over fitting) etmiş görünüyor. Bu model her veri noktalarını ifade edebilmiş (Ezberleme durumu) fakat verinin trendini ifade eden genelleme özelliği zayıf kalmıştır. Sahadan gelen gerçek veriler için ezberleme çoğu zaman modelin genelleme özelliğini azalttığı için istenmez. E karesel hata sıfır olsa bile bu model tercih edilmez. Bir önceki 2. derece polinom makine öğrenmesi açısından bu verileri daha iyi temsil edebilmektedir.

Örnek: Bir şehirde nüfusa karşılık günlük toplanan evsel katı atık miktarının (kg) değişimi polinomsal regresyon modelli $K(s) = 0.0001s^2 + 0.02s + 5$ elde edilmiştir. Bu model yardımı ile 1000 kişilik bir kasabanın katı atık miktarının kaç kg olabileceği şöyle tahmin edilebilir.

$$K(1000) = 0.0001(1000)^2 + 0.02(1000) + 5 = 100 + 20 + 5 = 125\,\text{kg}$$

Bu uygulamada, regresyon analizinin belli bir doğruluk ile tahmin yapabilmemizi sağladığı görüyoruz. Bu modelin elde edilmedi makine öğrenmesi açısından verilerden nüfus ile evsel katı atık miktarı arasındaki ilişkisinin öğrenilmesi anlamına gelir.

Örnek: Araçlarda bujilerin yakıtı yeterince yakamaması sonu aracın yakıt verimliliği düşer. Araçların yakıt sarfiyatı oranı (s) ve buji(ateşleyici) ekonomik ömrü K verileri kullanılarak oluşturulan lojistik regresyon modeli göre buji kullanım ömrü $(\%)\,K(s) = \frac{1}{1+e^{-(0.5s+2)}}$ ile ifade edilmiştir. Buna model yardımı ile

a) Bir aracın yakıt sarfiyatı oranı 1.2 ölçüldüğü durumda buji'lerinin ekonomik ömrünü $K(s)$ modeli yardımı ile tahmin ediniz.

$K(s) = \frac{1}{1+e^{-(0.5*1.2+2)}} = \%19.7$. Buji henüz yeni sayılır. Ekonomik ömrünün %19.7 si tamamlanmış.

b) Bir aracın yakıt sarfiyatı oranı 5.7 için buji ekonomik ömrünü model yardımı ile tahmin ediniz.

$K(s) = \frac{1}{1+e^{-(0.5*5.7+2)}} = \%70$. Buji ekonomik ömrünü tamamlamaya yaklaşmış ve yakıtın tam yanmasını sonucu aracın yakıt verimliliği düşürebilir.

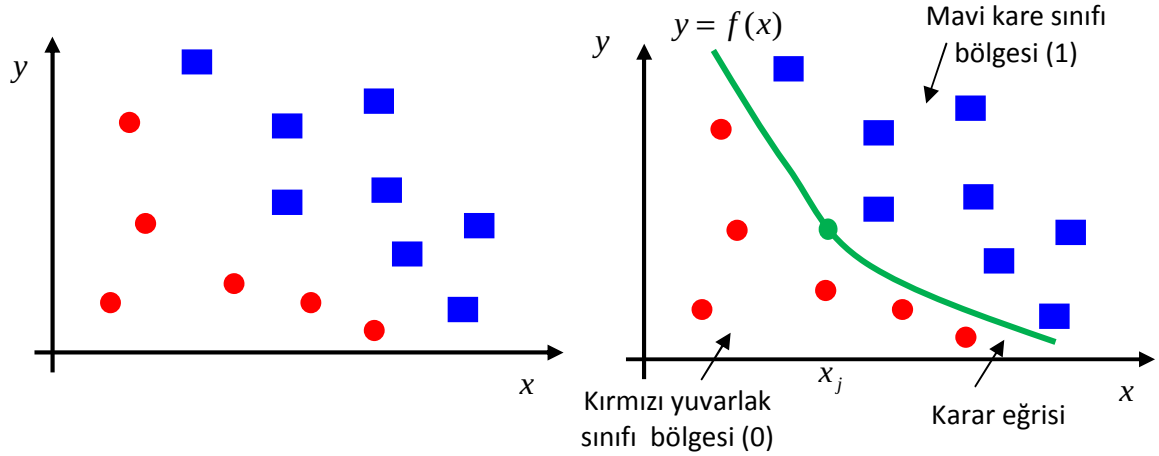
2- Sınıflama (Classification) Problemleri:

Sınıflama bir verinin hangi veri kümesine ait olduğunu bulma işlemidir (Phil Kim, 2017). Makine öğrenmesinde sınıflama probleminin çözümü için eğitim kümesinde etiketli veriler bulunur ve denetmen etiketlemesi ile verinin hangi sınıfa ait olduğu belirtilir. Sınıflama problemlerinde eğitim kümesinde toplanan verilere yaygın olarak "özellik" (feature) adı verilir. Sınıflama problemlerinde eğitim kümesi, özellikler ve bu özelliklere karşılık ait olduğu sınıfları ifade eden etiketlerle oluşturulur. Dolayısı ile öğrenen model'in girişi "özellikler" ve çıkışı ait olduğu sınıfı ifade eden "etiket" olur. Sınıflama problemleri iki sınıf için verilirse buna ikili sınıflama (binary classification) denir. İkiden fazla sınıf olursa buna çoklu sınıflama (multi-classification) problemi adı verilir. Aşağıda örnek üzerinden sınıflama probleminin çözümünün matematiksel olarak nasıl sağlandığını görmeye çalışalım.

Örnek: Kayısı kalitesi belirleme işlemi makine öğrenmesi ile yapılmak isteniyor. Sensörler yardımı ile ışık yansıtma değerleri (x) ve kayısı ağırlığı (y) ölçülmektedir. Bu iki veri kullanılarak kayısıları yüksek kalite ve düşük kalite olarak sınıflanması için bir eğitim veri seti oluşturunuz. Eğitim kümesinin sağlıklı oluşturulması makine öğrenmesi algoritmasının başarısı için önemlidir.

Girişler (Özellikler)		Çıkış (Sınıfı)
Işık Yansıtması (x)	**Ağırlığı (y)**	**Label (Denetmen Bilgisi)**
1.2	60 gr	1 (Yüksek Kalite)
0.5	45 gr	0 (Düşük Kalite)
0.4	50 gr	0 (Düşük Kalite)
1.3	63 gr	1 (Yüksek Kalite)
0.7	55 gr	0 (Düşük Kalite)
0.9	61 gr	1 (Yüksek Kalite)

Burada tablodaki her veri satırını 2 boyutlu (x,y) özellik uzayında bir nokta ile temsil edebiliriz. Her bir verinin etiketi mavi dikdörtgen (1- yüksek kalite) ve kırmızı yuvarlak (düşük kalite) ile gösterilmiştir. Veri tablosunda görülen bütün giriş ve çıkış değerleri grafikte temsil edilmiştir. Bu grafikte sınıfları birbirinden ayırmak için bir ayırıcı fonksiyona kullanılabileceği kolaylıkla görülür. Bu ayırıcı fonksiyonun eğrisi karar eğrisi olarak adlandırılır. Eğrinin altında kalan veriler bir sınıf, eğrinin üstünde kalan veriler diğer sınıfı ifade eder. Böylece matematiksel olarak sınıflama problemi çözülmüş olur. Karar eğrileri matematiksel olarak sınıflara ait veri bölgelerini birbirinden ayırır ve bu eğriyi temsil eden fonksiyon sınıflama problemi çözümü için öğrenen modeli oluşturur. Aşağıda verilen iki veri kümesini doğru olarak ayıran bir karar eğrisi çiziniz.



Yukarıda iki sınıfı ayıran yeşil renkli bir karar eğrisi belirlenmiştir. Karar eğrisinin, verileri özellik uzayında dağılımına göre sınıflara ayırmayı sağladığı görülebilir. Bu eğrinin fonksiyonu $f(x)$ olsun. Şekildeki karar eğrisine (ayırıcı eğri) göre ikili sınıflama şöyle yapılabilir.

* Bir (x_j, y_j) değerine sahip veri eğer $y_j < f(x_j)$ eğrisi altında kalıyorsa kırmız yuvarlak sınıfındır. Grafikte bu sınıfı 0 biti ile temsil ettik.

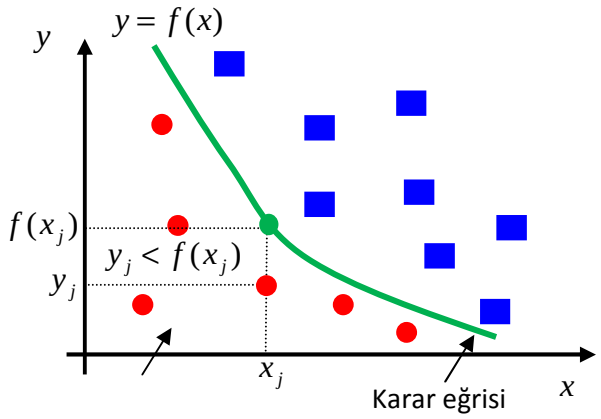
* Bir (x_j, y_j) değerine sahip veri eğer $y_j \geq f(x_j)$ eğrisi üstünde kalıyorsa mavi kare sınıfındadır. Grafikte bu sınıfı 1 biti ile temsil ettik.

Diğer bir ifade ile bu sınıflama işlemini matematiksel olarak şöyle ifade edebiliriz:

$$C = \begin{cases} 0 & y_j < f(x_j) \\ 1 & y_j \geq f(x_j) \end{cases}$$

Burada C hangi sınıfa ait olduğunu belirten koddur. Bu örnekte 0 kodu düşük kalite sınıfı (kırmızı yuvarlak), 1 kodu yüksek kalite sınıfı (mavi kare) için kullanılmıştır.

Bu sınıflama probleminin çözümü için makine öğrenmesi algoritmasının yapması gereken eğitim kümesi verileri yardımı ile en az sınıflama hatası ile uygun bir $y = f(x)$ karar eğrisini belirlenmesidir.

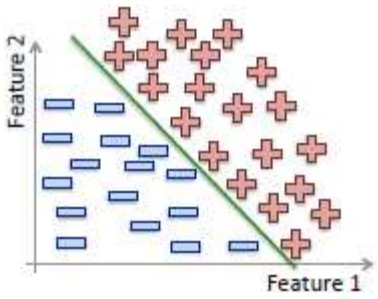


Burada verilerin doğarına uygun $y = f(x)$ ayırıcı fonksiyon seçimi önem kazanır. Sahadan toplanan gerçek verilerin özellik uzayındaki dağılımları sınıflama problemlerinin farklı zorluk derecesini belirler. Bazı dağılımlar lineer doğrular ile ayrıştırılabilir. Bazı dağılımları bir doğru ile ayrıştırmak mümkün olmayabilir ve eğriler ile ayırmak gerekir. Eğriler daha karmaşık ayırıcı fonksiyonlar kullanımını gerekli kılar. Bu durumda sınıflama problemi için veriler iki katagoride ele alınabilir.

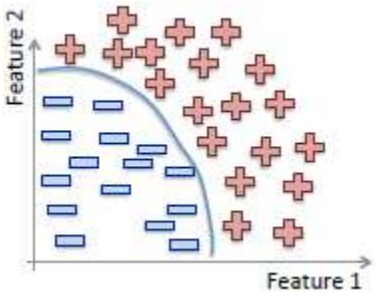
1- Lineer Ayrıştırılabilir Veriler: Bir doğru yardımı ile hatasız olarak ayrıştırılabilen verilere lineer ayrıştırılabilir veriler denir. Eğer veriler lineer ayrıştırılabilir ise ayrıcı fonksiyonun bir doğru denklemi (lineer fonksiyon, örneğin $y = m.x + b$) seçilmesi yeterli olur.

2- Lineer Ayrıştırılamaz Veriler: Bir doğru yardımı ile hatasız olarak ayrıştırışımı mümkün olmayan verilere lineer ayrıştırılamaz veriler denir. Eğer veriler lineer ayrıştırılamaz ise ayrıcı fonksiyonun bir doğrudan çok daha karmaşık eğriler üretebilen fonksiyonlardan seçilmesi

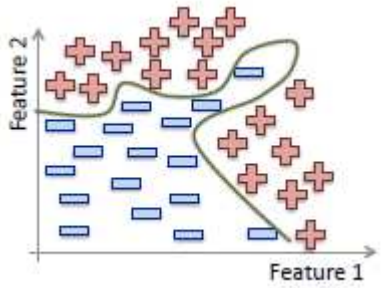
gerekir. (lineer olmayan fonksiyon, örneğin, bir polinom $y = m_2 x^2 + m_1 x + m_0$) Aşağıdaki örnekleri inceleyelim:



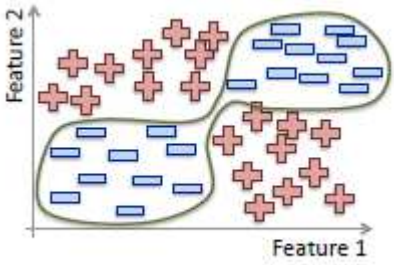
Lineer ayrıştırılabilir veriler için sınıflama bir doğru denklemi ile sağlanabilir.



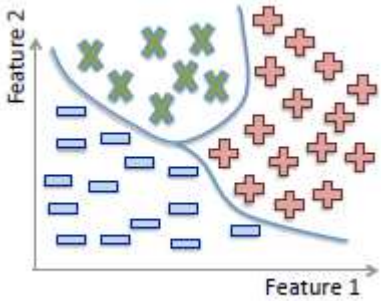
Lineer ayrıştırılamaz veriler. Karar eğrisi matematiksel olarak doğru denklemi ile ifade edilemez.



Lineer ayrıştırılamaz veriler. Bir eğri ile ayrıştırılabilir. Ancak, matematiksel olarak bu karar eğrisinin ifadesi nispeten daha zor.



Lineer ayrıştırılamaz veriler. Bir eğri ile ayrıştırılabilir. Matematiksel olarak karar eğrisinin ifadesi biraz daha zor.



Çok sınıflı lineer olarak ayrıştırılamaz veriler için ayırıcı eğrilir daha fazla sayıda olmalıdır. Dolayısı ile daha fazla ayırıcı fonksiyona ihtiyaç duyar.

Örnek: Bir lineer ayrıştırılabilir veri kümesi için $f(x) = -2x + 4$ formunda lineer bir karar eğrisi belirlenmiştir. Sınıflama işlemi ise şu eşikleyici kuralına göre yapılmış olsun.

$$C = \begin{cases} 1 & y_j < f(x_j) \\ 0 & y_j \geq f(x_j) \end{cases}$$

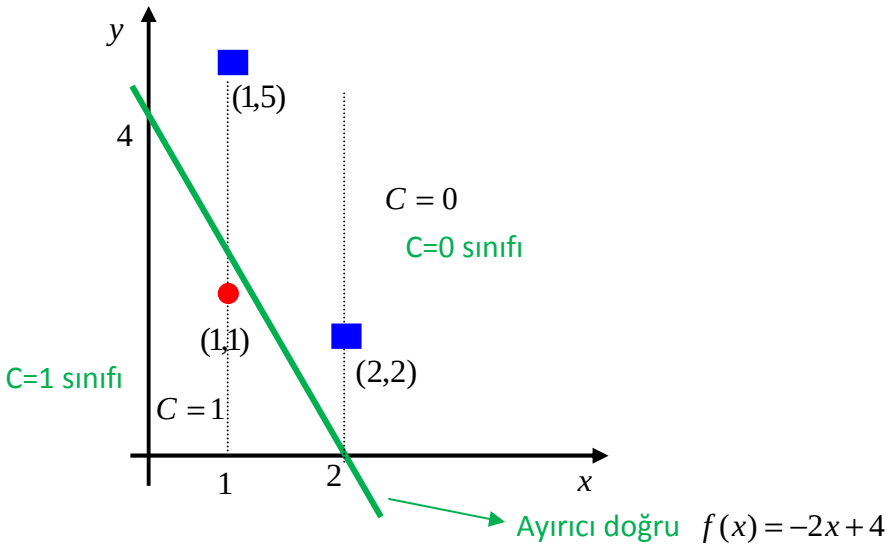
Buna göre ölçüm sonucu (x_j, y_j) formatında alınan $\{(1,1), (1,5), (2,2)\}$ verilerinin hangi sınıfa ait olduğunu bu fonksiyonlar yardımı ile bulunuz. İki boyutlu düzlemde verileri ve karar eğrisini çizerek sonuçların doğruluğunu kontrol ediniz.

$(x_j, y_j) = (1,1)$ noktası için karar eğrisi $f(1) = -2.1 + 4 = 2$ burada $1 < f(1)$ olduğu için $C = 1$ sınıfı

$(x_j, y_j) = (1,5)$ noktası için karar eğrisi $f(1) = -2.1 + 4 = 2$ burada $5 > f(1)$ için $C = 0$ sınıfı

$(x_j, y_j) = (2,2)$ noktası için karar eğrisi $f(2) = -2.2 + 4 = 0$ burada $2 > f(2)$ için $C = 0$ sınıfı

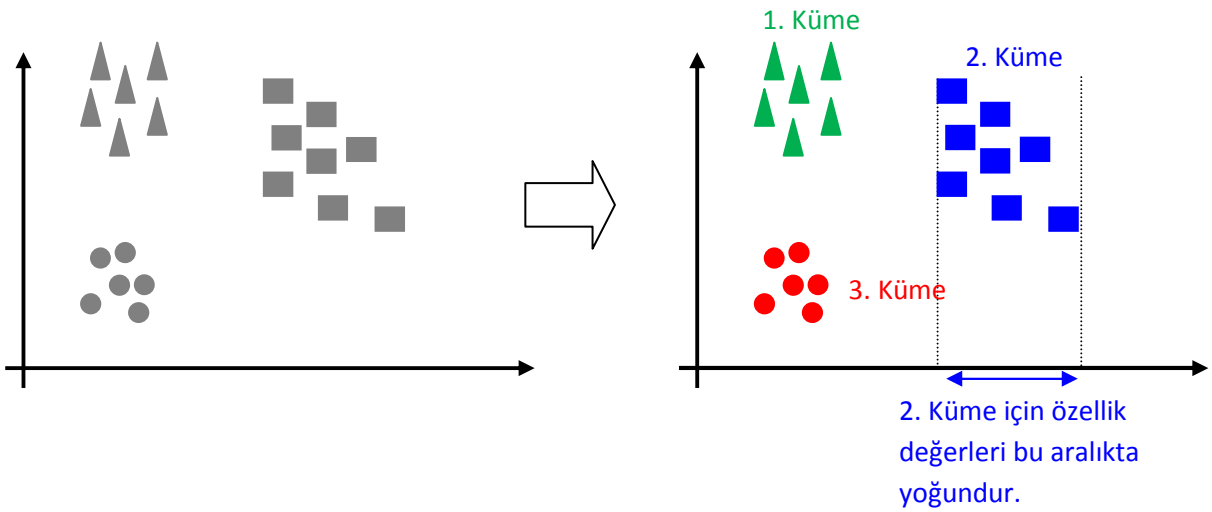
Bu sonuçların geçerliliği karar eğrisi grafiği üzerinde görebiliriz. Burada veriler nokta ile gösterilmiştir.



Makine Öğrenmesinde Denetimsiz (Unsupervised) Öğrenme Problemleri:

1- Kümeleme (Clustering) Problemleri: Etiketsiz verilerin benzerliğine göre sınıflanmasına kümeleme denir. Kümeleme problemlerinde veriler etiketli değildir. Makine öğrenmesi denetimsiz öğrenme tekniklerini uygulayarak gelen verilerin birbirine benzerlikleri göre kümeleme işlemini yürütür. Verilerin benzerlik çoğunlukla özellik ifade eden verilerin özellik

uzayında dağılımlarındaki yakınlık ve yoğunluğuna (yığılma bölgelerine) göre yapılabilmektedir. Burada metrik uzaylar ve veri benzerliği gibi kavramlar önem kazanır. Aşağıdaki dağılıma göre verilerin kümelenmesi gösterilmiştir. Veri noktaları birbirine yakınlığı ve kümelenmelerine göre yuvarlak, üçgen ve kare ile temsil edilen 3 küme ortaya çıkmıştır.



Makine Öğrenmesi Uygulamalarında Dikkat edilmesi Gereken Noktalar:

1- Problemin türü nedir? Regresyon, sınıflama, kümeleme vs?

2- Eğitim kümesi tasarımı: Eğitim kümeleri problemi ve çözümünü iyi ifade edebilir zenginlik ve dağılıma sahip olmalı. Çünkü, makine öğrenmesi problemi çözmeyi veri kümesinden öğrenir. "Veri kümesi neyi anlatır ise algoritma onu öğrenir." Doğal olarak makine öğrenmesinin başarısı, eğitim kümesinin problem çözümünü ifade edebilme derecesine bağımlıdır.

3- Probleme uygun makine öğrenmesi tekniğinin seçimi ve seçilen tekniğin uygulanan probleme dönük olarak performansının iyileştirilmesi ve parametrelerinin ayarlaması (meta parametre optimizasyonu).

Makine Öğrenmesinde Genelleme Özelliği ve Ezberleme (Aşırı Öğrenme):

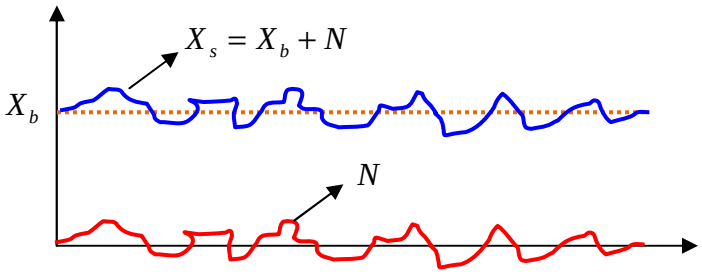
Sahadan elde edilen gerçek veriler ile çalışırken makine öğrenmesinde aşırı öğrenme (over-fitting) yani verilerin ezberlenmesi çoğunlukla istenmeyen bir durumdur. Regresyon konusunda ezberlemenin etkisi görülmüştü. Ezberleme durumu genelleme özelliğinin zayıflamasına yol açıyordu. Bu kısımda biraz daha detaylı olarak ezberleme ve genelleme kavramlarına eğileceğiz.

Öncelikle gerçek dünya verilerinin bir miktar gürültü (istenmeyen veya yanıltıcı) bilgi içerdiğini unutmayalım. Veriler, ölçen sisteme, ölçüm ortamına, ölçüm tekniğine, ölçüm ortamının kontrol edilememiş şartlarına bağlı olarak bir miktar belirsizlik içerir. Genelde

ölçümün gerçek sonucunu biraz değiştiren istenmeyen bilgiye belirsizlik veya gürültü adı verilir. Haberleşmede, gürültülü kanalda iletilen bilgi işareti (X_s) , gönderilen orijinal bilgi işareti (X_b) ve bir rastgele gürültü işareti (N)'nin toplamı ile basit bir şekilde ifade edilir. Buna göre bir haberleşme kanalından alınan bilgi iki bileşen içerir.

$$X_s = X_b + N$$

Kanaldan alınan bilgi işareti orijinal bilgi ve gürültü bileşenleri ile aşağıda temsili olarak çizilmiştir. Burada orijinal bilgi işareti sabit değerli X_b (Turuncu nokta) iken üzerine toplamsal olarak binen N rastgele karakterdeki gürültü (Kırmızıeğri) nedeni ile ölçülen bilgi işareti X_s (mavi) bir miktar orijinal bilgi işareti X_b den sapma gösterir (farklılaşır).



Dolayısı ile sahadan toplanan gerçek verilerinde yukarda ki gibi bir miktar belirsizlik ve gürültü içermesi beklenir ve oldukça doğaldır. Gürültünün ezberlenmesi faydalı değildir ve orijinal bilginin öğrenilmesinde yanıltıcı olabilir. Bu sorunu aşmak için genelleme istenir, diğer bir ifade ile verinin almış olduğu değerleri değil de, verinin trendinin veya yönelimlerinin öğrenilmesi tercih edilir. (Bu örnekte X_s toplanan veri için, X_b yi ifade eden sabit trendin (turuncu nokta) öğrenilmesi istenir.) Çünkü, verinin trendi orijinal veriyi X_s üreten sistemi daha doğru temsil edebilir. Burada ezberleme (fazla öğrenme) gürültününde öğrenilmesine yol açar ve bu çoğu durumda istenmez. Uygulamada gürültünün ezberlenmesi makine öğrenmesi algoritmasını yanıltabilir.

Genelleme (generalization) aynı zamanda veriyi temsil edebilen en basit modelin kurulması ile sağlanabilir (Phil Kim,2017). Dolayısı ile ezberleme, aslında model karmaşıklığının veriye göre çok yüksek olması yani veri değerlerini ezberleyecek düzeyde olmasından kaynaklanır. Model basitleştirilerek genelleme özelliği artırılabilir. Ayrıca düzleştirme (regularization) adı verilen modelin öğrenme esnekliğini kontrol ederek genelme sağlayabilen yaklaşımlar vardır.