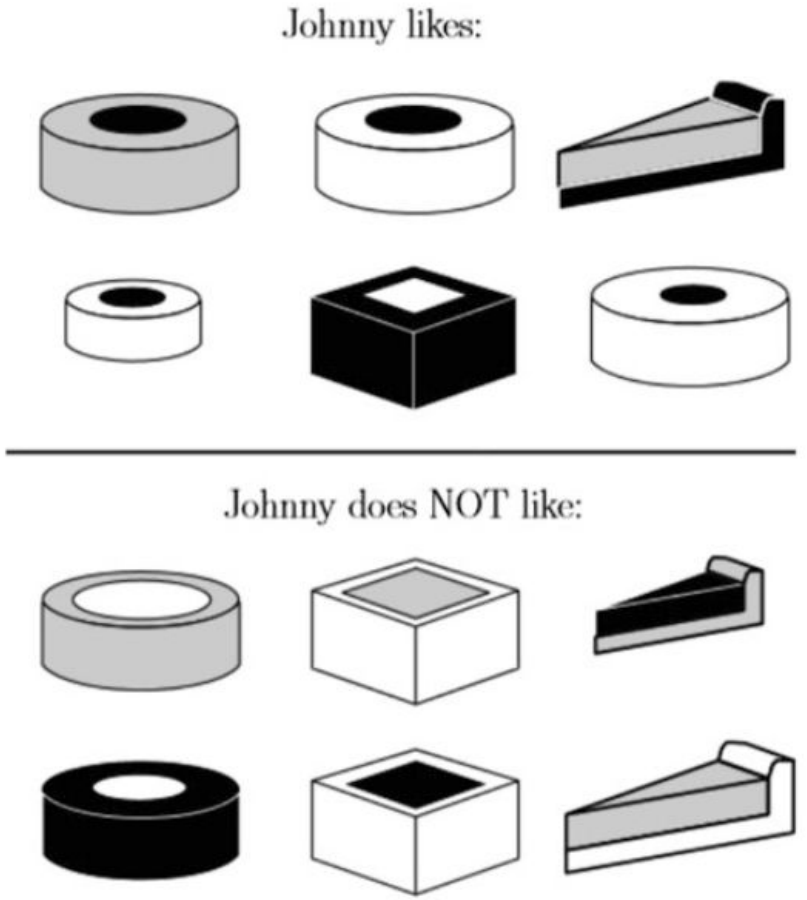
Makine Öğrenmesi

SD413

Sınıflandırma Problemi

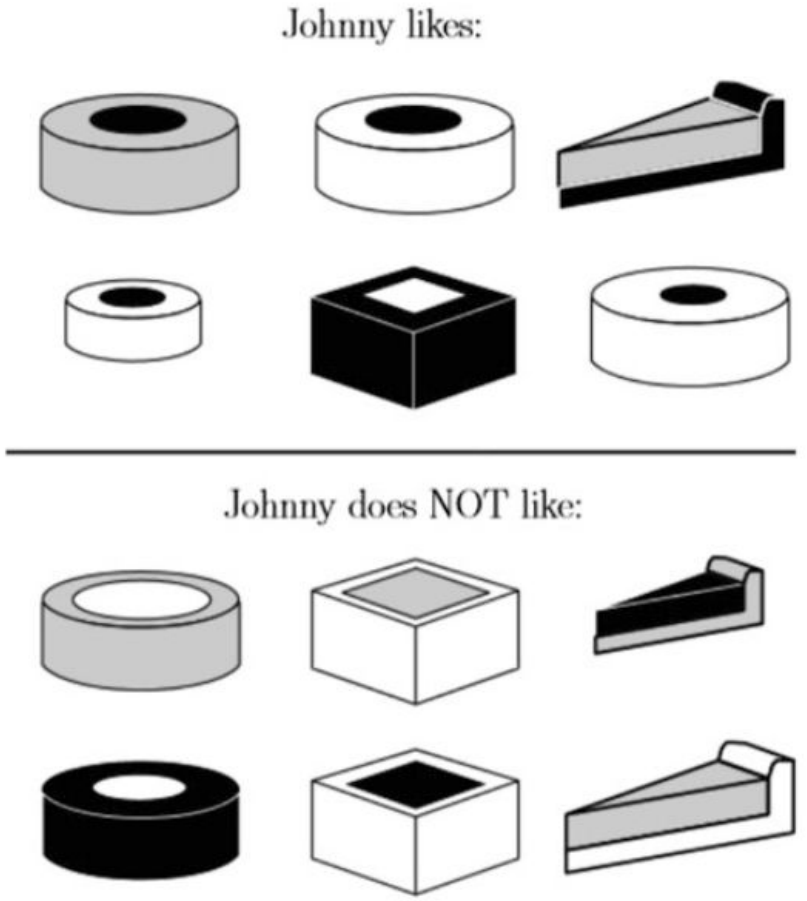
- Burada Johnny'nin sevdiği ve sevmediği pastalar görülmektedir.
- Bu pasta gruplarını ayırt etmek için **hangi özellikler** kullanılabilir?



Sınıflandırma Problemi

Öznitelikler:

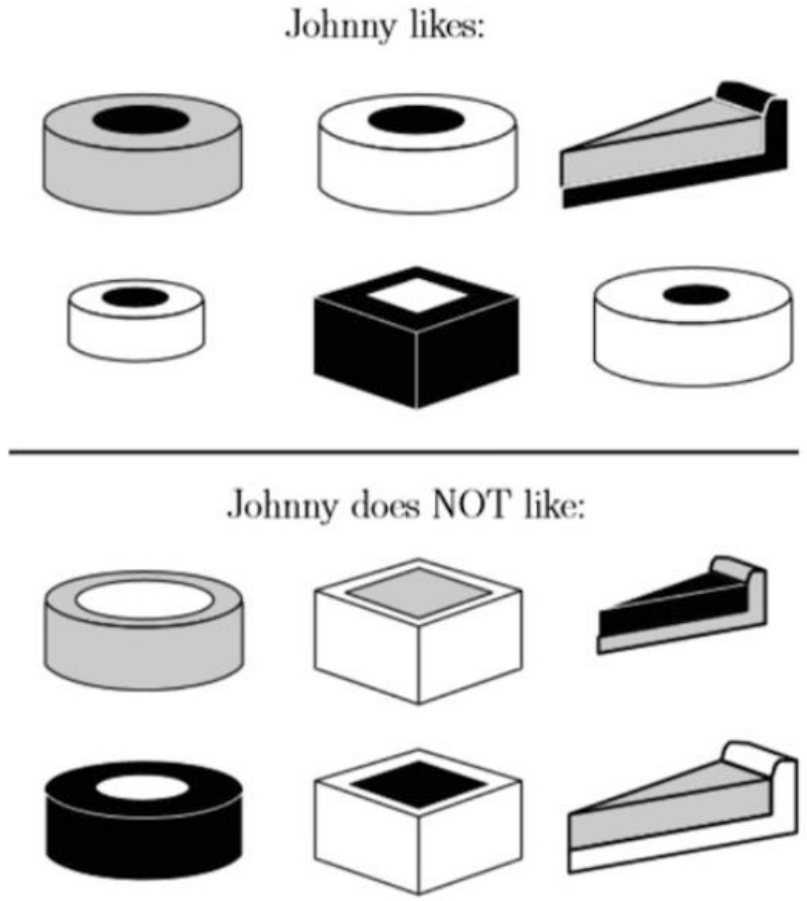
- Shape (circle, triangle, square)
- Crust-size (thin, thick)
- Crust-shade (white, gray, dark)
- Filling-size (thin, thick)
- Filling-shade (white, gray, dark)



Sınıflandırma Problemi

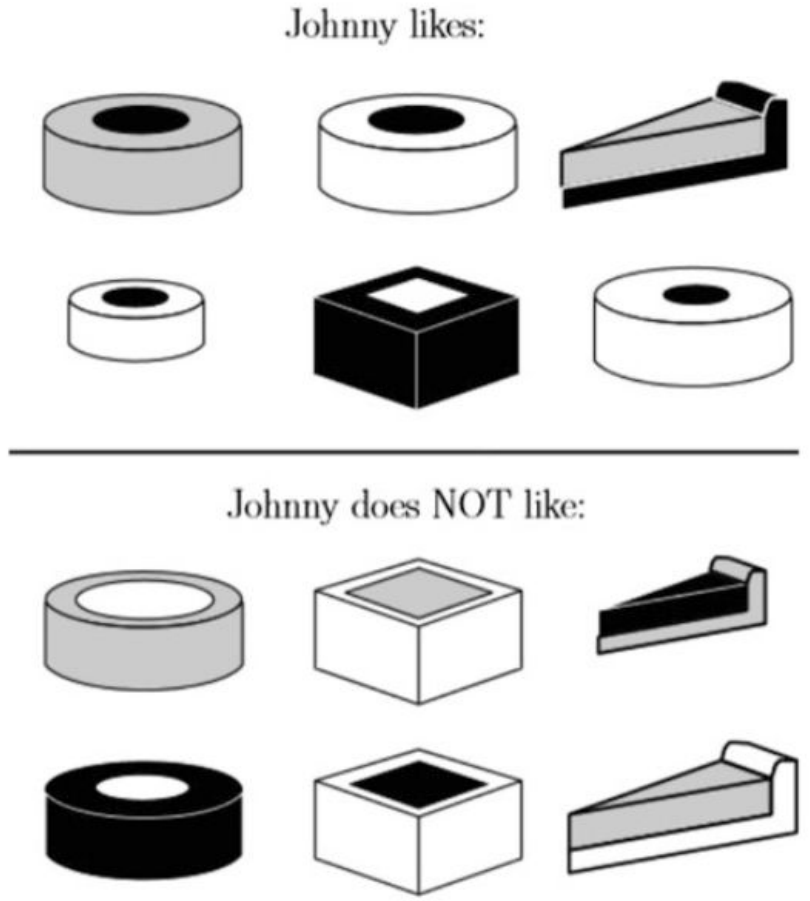
Veri tablosu:

Example	Shape	Crust		Filling		Class
		Size	Shade	Size	Shade	
ex1	Circle	Thick	Gray	Thick	Dark	pos
ex2	Circle	Thick	White	Thick	Dark	pos
ex3	Triangle	Thick	Dark	Thick	Gray	pos
ex4	Circle	Thin	White	Thin	Dark	pos
ex5	Square	Thick	Dark	Thin	White	pos
ex6	Circle	Thick	White	Thin	Dark	pos
ex7	Circle	Thick	Gray	Thick	White	neg
ex8	Square	Thick	White	Thick	Gray	neg
ex9	Triangle	Thin	Gray	Thin	Dark	neg
ex10	Circle	Thick	Dark	Thick	White	neg
ex11	Square	Thick	White	Thick	Dark	neg
ex12	Triangle	Thick	White	Thick	Gray	neg



Sınıflandırma Problemi

Example	Shape	Crust		Filling		Class
		Size	Shade	Size	Shade	
ex1	Circle	Thick	Gray	Thick	Dark	pos
ex2	Circle	Thick	White	Thick	Dark	pos
ex3	Triangle	Thick	Dark	Thick	Gray	pos
ex4	Circle	Thin	White	Thin	Dark	pos
ex5	Square	Thick	Dark	Thin	White	pos
ex6	Circle	Thick	White	Thin	Dark	pos
ex7	Circle	Thick	Gray	Thick	White	neg
ex8	Square	Thick	White	Thick	Gray	neg
ex9	Triangle	Thin	Gray	Thin	Dark	neg
ex10	Circle	Thick	Dark	Thick	White	neg
ex11	Square	Thick	White	Thick	Dark	neg
ex12	Triangle	Thick	White	Thick	Gray	neg




```
[ (shape=circle) AND (filling-shade=dark) ] OR [ NOT(shape=circle) AND (crust-shade=dark) ]
```

Sınıflandırma Problemi

- Amacımız sınıflandırıcıyı kaba kuvvet (brute force) yaklaşımı ile bulmak değildir.
- $3 \times 2 \times 3 \times 2 \times 3 = 108$ farklı örnek
- Bu örneklerle oluşturulabilecek 2^{108} farklı alt küme vardır!

Sınıflandırma Problemi

- Şimdiye kadar, eğitim örneklerinin bilinen sınıfları ile sınıflandırıcı tarafından önerilen sınıfları karşılaştırarak hata oranını ölçtük.
- Pratik olarak konuşursak, amacımız sınıflarını zaten bildiğimiz nesneleri yeniden sınıflandırmak değildir.
- **Nihai amacımız hangi sınıfa ait olduğunu bilmediğimiz gelecekteki örnekleri etiketlemektir.**
- Bu sebeple verilerin bir kısmını eğitim, bir kısmını test olarak iki gruba ayırırız.

- **Problem**: Verileri bu şekilde bölmenin dezavantajları neler olabilir?

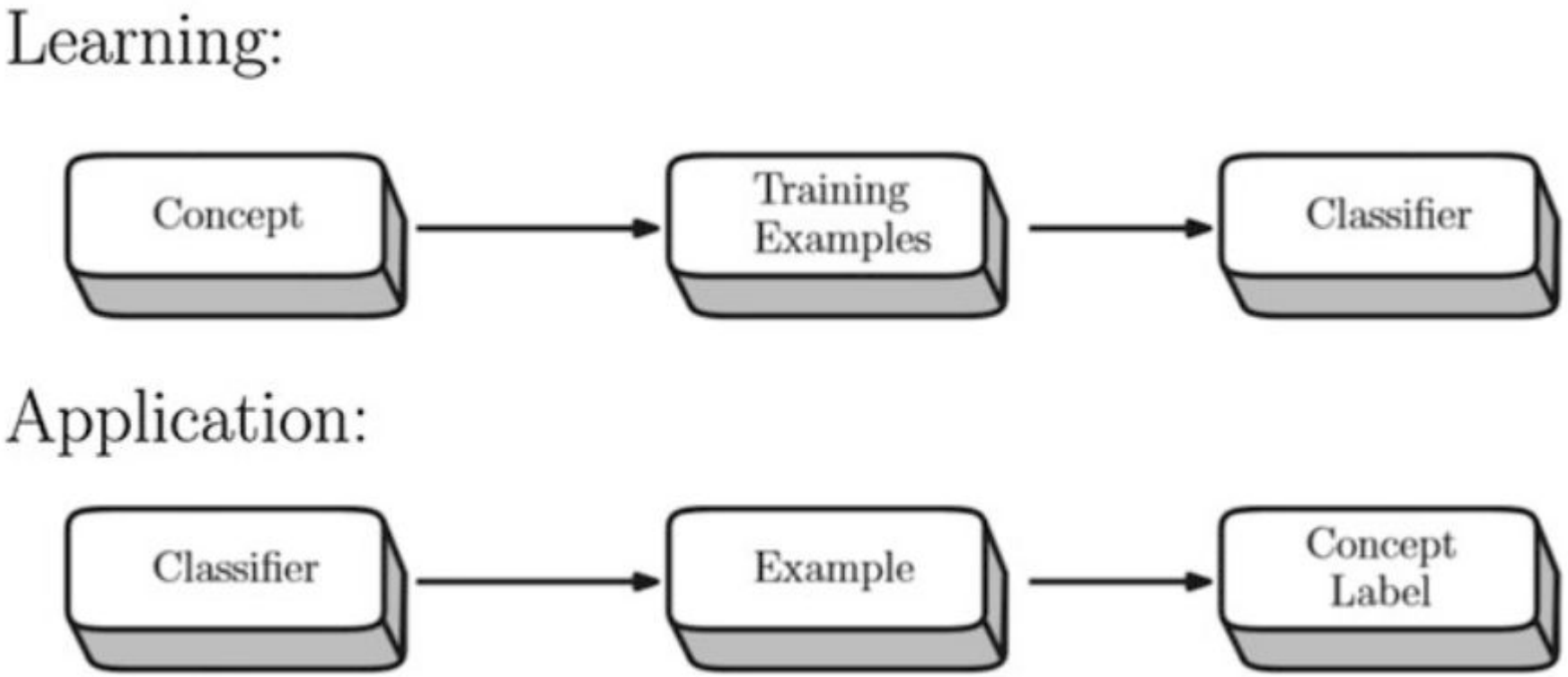
Sınıflandırma Problemi

- "Pasta" örneğinde 12 eğitim verisi bulunmaktaydı ve kalan 96 örneğin sınıfları bilinmiyordu.
- Mantıksal ifade olarak yazdığımız sınıflandırıcı, görünen örnekleri doğru şekilde ayırt etse de **bilinmeyen örnekler üzerinde** farklı işleyebilir.
- Johnny belki de çok daha kompleks pasta zevkine sahiptir.
- Eldeki sınırlı sayıda örnek kendisine sorularak evet hayır şeklinde veri toplanmış olabilir.

Sınıflandırma Problemi

- Öğrenme sırasında görülmeyen örneklerdeki hata oranını nasıl tahmin edebiliriz?
- Rastgele alt örnekleme nedir?
- **Neden test setindeki hata oranı eğitim setindekinden genellikle daha yüksektir?**
- Sınıflandırıcının eylemini de açıklamak zorunda olduğu bir etki alanı (domain) ve bunun gereksiz olduğu bir etki alanı örneği verin.
- "Bütün eğitim örneklerini doğru bir şekilde sınıflandıran kombinatoryal sayıda sınıflandırıcı vardır" derken ne demek istiyoruz?

Sınıflandırma Problemi



Sınıflandırma Problemi

- Önceki şekilde gösterilen sınıf tanıma görevi, makine öğrenmesi disiplininin en popüler görevidir.
- Bu çerçevede birçok somut mühendislik problemi ortaya konulabilir: görsel nesnelerin tanınması, doğal dili anlama, tıbbi teşhis ve bilimsel verilerdeki gizli kalıpların belirlenmesi.
- Bu alanların her biri, bu nesneleri karakterize eden özelliklere, niteliklere ve niteliklere dayalı olarak nesneleri doğru sınıflarla etiketleyebilen sınıflandırıcılarla çalışır.

Sınıflandırma Problemi

- Bazı uygulamalarda eğitim seti manuel olarak oluşturulur: bir uzman örnekleri hazırlar, onları sınıf etiketleriyle etiketler, nitelikleri seçer ve her örnekte her bir özelliğin değerini belirtir.
- Diğer alanlarda, süreç bilgisayarlıdır. Örneğin, bir şirket, bir çalışanın ayrılma niyetini tahmin edebilmek isteyebilir. Veritabanları, her bir kişi için adres, cinsiyet, medeni durum, görev, maaş artışları, terfilerin yanı sıra kişinin hala şirkette olup olmadığı veya değilse ayrıldığı gün hakkında bilgileri içerir. Bir program, ilgili kişinin veri tabanı kaydının son güncellemesinden bu yana bir yıl içinde ayrılması durumunda pozitif olarak etiketlenen öznitelik vektörlerini elde edebilir.
- Bazen öznitelik vektörleri bir veri tabanından otomatik olarak çıkarılır ve bir uzman tarafından etiketlenir. Alternatif olarak, bazı örnekler bir veritabanından alınabilir ve diğerleri manuel olarak eklenebilir. Çoğu zaman, iki veya daha fazla veritabanı birleştirilir.

Farklı Nitelik Türleri

- Pasta örneğinde, niteliklerden herhangi biri iki veya üç farklı değerden yalnızca birini alabilir. Bu tür niteliklere "ayrık" (discrete) değişken denir.
- Yaş gibi diğer nitelikler "sayısal" (numeric) olarak adlandırılacaktır, çünkü değerleri sayıdır, örneğin yaş = 23.
- Bazen, sayısal değer sürekli bir etki alanından gelir. Örneğin ağırlık = 73.5. Bu durumda özniteliğin "sürekli değerli" (continuous) olduğunu söyleyeceğiz.

Verilerle İlgili Sorunlar

Alakasız Nitelikler (Irrelevant Attributes)

- Bazı nitelikler önemlidir, diğerleri değildir. Mesela, Johnny kakaolu pastayı seviyor olsa da, pasta tercihi aşçının ayakkabı numarasından pek etkilenmeyecektir.
- **Alakasız nitelikler hesaplama maliyetlerini artırır; hatta öğreniciyi yanıltabilirler.**
- Örnekler bir veri tabanından otomatik olarak ayıklandığında, daha sık olarak bu tip sorunlar ortaya çıkacaktır. Veritabanları, öncelikle çok sayıda bilgiye erişim sağlamak amacıyla geliştirilir ve bunların genellikle yalnızca küçük bir kısmı öğrenme göreviyle ilgilidir. Bunun hangi kısım olduğu konusunda genellikle hiçbir fikrimiz olmaz.

Verilerle İlgili Sorunlar

Eksik Nitelikler (Missing Attributes)

- Bazı kritik özellikler eksik olabilir.
- Ailesinin mali durumunu göz önünde bulunduran Johnny, pahalı pastalara karşı önyargılı olabilir. Fiyat niteliğinin olmaması, iyi bir sınıflandırıcıyı inşa etmeyi imkansız hale getirecektir: mevcut nitelikler açısından özdeş olan iki örnek, hayati "eksik" özelliğin değerlerinde farklılık gösterebilir.
- Aynı şekilde tanımlansa da, bir örnek olumlu, diğeri olumsuz olabilir. **Bu durumda, eğitim setinin tutarsız olduğunu söylüyoruz.**
- Bazı durumlarda bundan kaçınılması zordur: Uzman yalnızca fiyatın uygunluğu konusunda bilgisiz olmakla kalmaz; bu niteliğin değerlerini sağlamak imkansız olabilir ve bu nedenle nitelik hiçbir şekilde kullanılamaz.

Verilerle İlgili Sorunlar

Gereksiz Nitelikler (Redundant Atrributes)

- Değerleri diğer niteliklerden elde edilebilen nitelikler biraz daha az zarar vericidir.
- Veritabanı bir hastanın doğum tarihini ve yaşını içeriyorsa, değeri bugünün tarihinden doğum tarihi çıkarılarak hesaplanabileceğinden, ikincisi gereksizdir.
- Neyse ki, gereksiz nitelikler, alakasız veya eksik olanlardan daha az tehlikelidir.

Verilerle İlgili Sorunlar

Eksik Öznitelik Değerleri (Missing Attribute Values)

- Bazı uygulamalarda, bazı niteliklerin değerleri bilinmeyebilir.
- Örneğin, bir şirketin veri tabanı, yalnızca bazı çalışanlar için çocuk sayısı hakkında bilgi içerebilir, diğerleri için olmayabilir. Veya bir hastanenin hasta dosyasının bir veri tabanında, her hasta yalnızca bazı laboratuvar testlerinden geçmiştir. Bunlar için değerler bilinmektedir; ancak her hastayı mevcut tüm testlere tabi tutmak imkansız (ve mantıksız) olduğundan, çoğu test sonucu eksik olacaktır.

Verilerle İlgili Sorunlar

Nitelik-Değer Gürültüsü (Attribute-Value Noise)

- Öznitelik değerlerine ve sınıf etiketlerine, güvenilir olmayan bilgi kaynakları, zayıf ölçüm cihazları, yazım hataları, kullanıcının kafa karışıklığı ve diğer birçok nedenden dolayı genellikle güvenilemez.
- Veriler çeşitli gürültü türlerinden muzdarip olabilir.
- Stokastik (olasılıksal) gürültü rastgeledir. Örneğin gün içinde vücut ağırlığımız değiştiği için sabah aldığımız değer akşamki değerden farklıdır. Bir insan hatası da rol oynayabilir: Hastanın kan basıncını ölçecek zamanı olmayan ihmalkar bir hemşire, önceki okumanın bir modifikasyonunu kullanır.
- Buna karşılık, sistematik gürültü tüm değerleri aynı yönde sürükler. Örneğin, zayıf kalibre edilmiş bir termometre her zaman olması gerekenden daha düşük bir okuma verir.

Verilerle İlgili Sorunlar

Sınıf-Etiket Gürültüsü (Class-Label Noise)

- Bir uzman tarafından önerilen etiketler düzgün bir şekilde kaydedilmemiş olabilir; alternatif olarak, bazı örnekler kendilerini iki sınıf arasında "gri bir alanda" bulurlar ve bu durumda doğru etiketler kesin değildir.
- Her iki durum da stokastik gürültüyü temsil eder ve bunlardan ikincisi tipik olarak iki sınıf arasındaki sınır bölgesinden gelen örnekleri etkiler.
- Bununla birlikte, sınıf etiketi gürültüsü de sistematik olabilir: bir doktor, kanıtlar ezici olmadıkça nadir görülen bir hastalığı teşhis etme konusunda isteksiz olabilir – bu durumda, sınıf etiketlerinin pozitiften ziyade negatif olması daha olasıdır. Son olarak, sınıfların yanlış giden otomatik bir süreç tarafından sağlandığı alanlarda, sınıf etiketlerinde rastgele yapay nesnelerle karşılaşılır.
- **Sınıf etiketi gürültüsü genellikle nitelik değeri gürültüsünden daha tehlikelidir.** Bir niteliğin yanlış bir değeri, örneğin genel özelliklerini yalnızca biraz değiştirebilir ve bu nedenle, sınıflandırıcıyı yalnızca marjinal olarak etkileyebilir. Buna karşılık, olumsuz olarak yanlış etiketlenmiş olumlu bir örnek oldukça yanıltıcı olabilir.

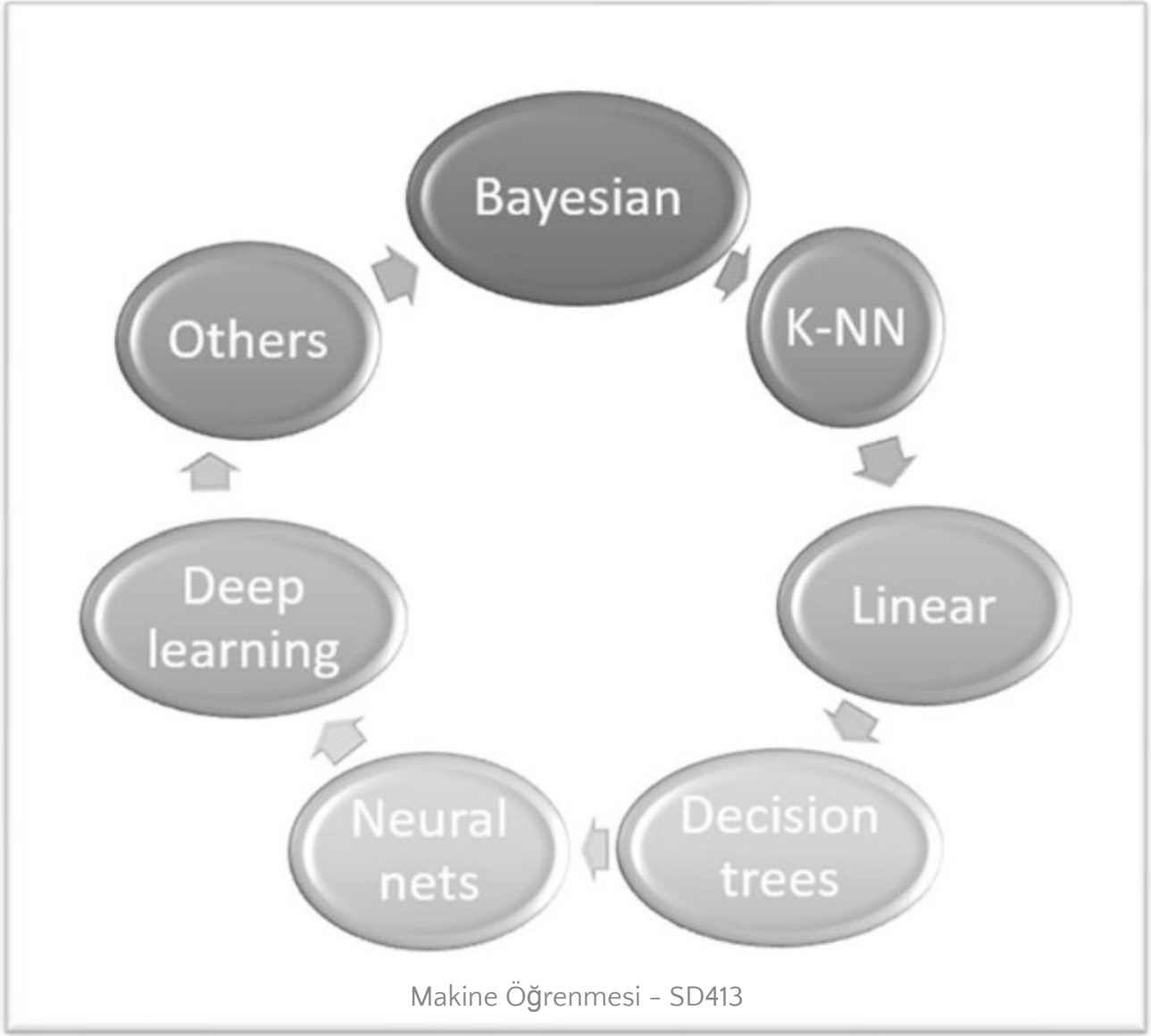
Aşağıdaki terimleri açıklayın:

- Alakasız ve gereksiz nitelikler, eksik nitelikler ve eksik nitelik değerleri. Her birini pasta örneğini kullanarak açıklayın.
- "Tutarsız eğitim seti" nedir? Nedeni ne olabilir? Öğrenme sürecini nasıl etkileyebilir?
- Ne tür gürültüler biliyoruz? Olası kaynakları nelerdir? Gürültü, öğrenme girişiminin başarısını ve/veya başarısızlığını ne şekilde etkileyebilir?

Kavram Öğrenmeye Giden Yollar

- Şimdi, önceden sınıflandırılmış eğitim örneklerinden öğrenmenin kolay olmadığını anlamış durumdayız.
- Eğitim seti mükemmel ve gürültüsüz olsa bile, tüm eğitim örneklerini doğru bir şekilde sınıflandırabilen ancak öğrenme sırasında görülmeyen örnekleri ele almalarında farklılık gösterecek birçok sınıflandırıcı bulunabilir.
- En iyisi nasıl seçilir?

Kavram Öğrenmeye Giden Yollar



Gerçek Dünyayla Yüzleşmek

- Eğitim örnekleri nadiren mükemmeldir. Çoğu zaman, sınıf etiketleri ve nitelikleri gürültülüdür, mevcut bilgilerin çoğu alakasız, gereksiz veya eksiktir, eğitim seti tüm kritik yönleri yakalamak için çok küçük olabilir.
- Bütün bir makine öğrenmesi disiplini yukarıda bahsedilen tüm meselelerle uğraşmaya ve altta yatan görevlerin tüm karmaşık komplikasyonlarını aydınlatmaya çalışmaktadır.
- Önceki şekilde belirtildiği gibi, mühendislerin emrinde, her biri farklı özelliklerle işaretlenmiş, her biri somut bir göreve uygulandığında farklı avantajlar ve eksiklikler sergileyen birkaç büyük ve bazı küçük paradigmalar vardır.

Olasılıklar

- Eğitim verilerini mükemmel bir şekilde sınıflandıran, ancak gelecekteki veriler üzerinde farklı davranışları olan birçok sınıflandırıcının olabileceğini gördük.
- Mühendis hangi sınıflandırıcıları kullanmalı?
- Bu soruya cevap vermenin bir yolu, zaman içinde kendini kanıtlamış olasılık teorisine güvenmektir.
- Eğitim setindeki dairesel veya kare pastaların **bağıl frekansları** kesinlikle gelecekteki bir pastanın pozitif veya negatif sınıftan olduğuna dair ipuçları veriyor mu? Bu, Bayes sınıflandırıcıları tarafından takip edilen yoldur.

Benzerlikler

- Başka bir fikir, benzerliklere güvenmektir.
- **Aynı sınıfa ait nesneler bir şekilde benzerdir.**
- Bu akıl yürütme, en yakın komşu sınıflandırıcılarının temelini oluşturur.
- Bir örnek çiftinin karşılıklı benzerliğinin, onları tanımlayan öznitelik vektörleri arasındaki geometrik uzaklıkla yakalandığı varsayılır.

Karar Yüzeyleri

- Bir başka büyük felsefe, çok boyutlu "uzay metaforu" etrafında inşa edilmiştir.
- Basitlik için, tüm niteliklerin sayısal olduğunu, böylece her örneğin, N'nin niteliklerin sayısı olduğu N boyutlu bir uzayda tek bir nokta ile tanımlanabileceğini varsayalım.
- **Aynı sınıfa ait örneklerin kendilerini geometrik olarak birbirine yakın bulma eğiliminde oldukları doğruysa, o zaman olumlu örneklerin işgal ettiği bir bölgeyi ve olumsuz örneklerin işgal ettiği başka bir bölgeyi betimlemek mümkün olmalıdır.**
- Bu bölgeler daha sonra bir "karar yüzeyi" ile ayrılabilir: bir tarafta olumlu örnekler ve diğer tarafta olumsuz örnekler.

İleri Konular

- Farklı türdeki sınıflandırıcılar bir araya getirilerek daha başarılı modeller üretilebilir.
- Örneğin, oy kullanan sınıflandırıcı grupları birleştirilerek sınıflandırma performansı artırılabilir.
- Temel sınıflandırıcılarda, genellikle kavram kayması, dengesiz sınıflar ve önyargı gibi problemler vardır.

Derin Öğrenme

- Makine öğreniminde şu anda en ünlü atılımlardan bazıları, düşük seviyeli öznitelikleri (örneğin, her biri bir bilgisayar ekranında bir pikselin yoğunluğunu veren), daha sonra tanıma amacıyla kullanılan anlamlı yüksek seviyeli özelliklere dönüştüren yeni mekanizmalar tarafından elde edildi.
- Bu dönüştürme mekanizmaları genellikle birkaç (veya çok) katmana sahip yapay sinir ağlarını kullandığından, araçlar toplu olarak derin öğrenme adı altında bilinir hale geldi.
- Bu nispeten yeni teknoloji, bilgisayarlı görmede iyi duyurulan atılımlar sayesinde ünlü oldu. Örneğin, lisans dersleri bile artık bir resimde göz veya burun gibi belirli nesneleri tanımayı öğrenen bir bilgisayar programının nasıl yazılacağını öğretiyor.

Makine Öğrenmesinin Diğer Konuları

- Sınıflandırıcıların oluşturulması, en popüler makine öğrenimi görevidir, ancak tek görev değildir!

Denetimsiz Öğrenme (Unsupervised Learning)

- Sınıflarla etiketlenmemiş örneklerden bile birçok bilgi toplanabilir.
- Örneklerin benzer öznitelik vektörlerinden oluşan kümeler oluşturduğu ortaya çıkabilir.
- Bu tür kümelerin her biri, incelenmeyi hak edebilecek farklı özellikler sergileyebilir.
- Ortaya çıkan iki boyutlu matris, verilerin klasik küme analizinden farklı şekillerde görselleştirilmesine yardımcı olur.
- Örnek uzayının hangi kısımlarının yoğun, hangi kısımlarının seyrek olduğu görülebilir, hatta kaç tane istisna olduğunu öğrenebiliriz.

Pekiştirmeli Öğrenme (Reinforcement Learning)

- Makine öğreniminin en büyük zaferleri arasında, belki de en büyüleyici olanı, bilgisayarların satranç, Tavla ve Go gibi oyunlarda en iyi insanları yenmesidir.
- Nesiller boyunca, bu tür başarılar imkansız kabul edildi! Ve en nihayetinde, bu noktaya geldik.
- Bilgisayar programları, kendi kendilerine sayısız oyun oynayarak ve bu deneyimden bir şeyler öğrenerek ustalaşmayı öğrenebilirler. Bu başarıların ardındaki sır, genellikle yapay sinir ağları ve derin öğrenme ile birlikte **pekiştirmeli öğrenme** olarak bilinen tekniklerdir.
- Uygulama alanı sadece oyun oynamaktan çok daha geniştir. Gerçek dünya ortamlarında hareket etme, bu ortamdaki değişikliklere tepki verme, kutup dengelemeden araç navigasyonuna ve ayrıntılı teknik açıklaması olmayan alanlarda gelişmiş karar vermeye kadar değişen görevlerde makinenin davranışını optimize etme yeteneği geliştirilir.

Gizli Markov Modelleri (Hidden Markov Models)

- Elimizdeki tek bilgi ağaç halkalarıysa, 14. yüzyılda kaç yılın sıcak, kaçının soğuk olduğunu tahmin edebilir miyiz?
- Cevap evet – sıcak ya da soğuk yıllarda ağaç halkalarının küçük, orta ya da büyük olma olasılıklarını ve örneğin, soğuk bir yılı sıcak bir yılın izlemesinin ne kadar muhtemel olduğunu biliyorsak. Orta çağlarda doğrudan sıcaklık ölçümleri olmadığı halde, ağaç halkalarının sağladığı dolaylı bilgilerden hala makul görüşler geliştirebiliriz.
- Amaç dolaylı değişkenlere dayalı **zaman serisi tahminleri** yapmaktır.
- Bu tür problemler, finans, doğal dil işleme, biyoinformatik ve finans dahil olmak üzere etkileyici bir dizi uygulamaya uygulanmıştır. Makine öğrenmesinin görevi, mevcut verilerden çeşitli olasılıklardan oluşan güvenilir bir model ortaya çıkarmaktır: X durumunu Y durumunun izlemesi ne kadar olasıdır, temel durum X ise A gözleminin yapılma olasılığı ne kadardır vb.

Sorular

1. Pasta örneğinde, örnek uzayının herhangi bir alt kümesinin farklı bir sınıflandırıcıyı temsil edebilmesi koşuluyla, tüm sınıflandırıcıların uzayının boyutu 2^{108}'dir. Yalnızca öznitelik-değer çiftlerinin bağlaçları biçiminde sınıflandırıcılara izin verirsek, arama uzayı ne kadar küçülür?
2. Sizce pasta verisinde ne tür bir gürültü olabilir? Bu gürültünün kaynağı ne olabilir? Başka hangi sorunlar bu tür eğitim setlerini mükemmel olmaktan çıkarabilir?

Sorular

3. Bazı sınıflandırıcılar, açıklamalar açısından fazla bir şey sunmayan kara kutular gibi davranırlar. Kara kutu sınıflandırıcılarının pratik olmadığı alan örnekleri önerin ve bu sınırlamanın önemli olmadığı alanlar önerin.
4. Verilerle ilgili problemler bağlamında, bunlardan hangileri gerçekten ciddi ve hangileri tolere edilebilir?
5. Gereksiz nitelikler ile alakasız nitelikler arasındaki fark nedir?

Sorular

6. Tanımlaması zor olduğunu düşündüğünüz bir konu düşünün – örneğin, karmaşık bir biyolojik nesnenin (meşe ağacı, devekuşu vb.) veya bir müzik türünün (rock, folk, caz vb.) tanınması. Potansiyel eğitim örneklerini tanımlamak için nitelikler listesi önerin. Bu niteliklerin değerlerini elde etmek kolay olacak mı? Bu bölümde tartışılan problemlerden hangisinin öğrenme sürecini zorlaştırmasını bekliyorsunuz?
7. Makine öğrenmesi araştırmasının bir kolu, sınıflarla etiketlenmemiş örneklerden öğrenmeye odaklanır. Sizce bu tür programların pratik faydaları neler olabilir?
8. Oyun oynarken pekiştirmeli öğrenmenin başarıları etkileyicidir, ancak belki de gerçek dünyada, örneğin endüstride veya ekonomide çok yararlı değildir. Bu tekniklerden faydalanabilecek daha pratik bir uygulama alanı düşünebiliyor musunuz?

Bilgisayar Ödevi

- "Johnny'nin sevdiği pastalar" tanımını aramayı gerçekleştirecek bir program yazın. Kendi genelleme ve uzmanlaşma operatörlerinizi tanımlayın. Değerlendirme işlevi, eğitim örneklerinde gözlemlenen hata oranına bağlı olacaktır.

- Ders Sonu.