MNİST VERİ SETİ

Sunum 1

Özet

Veri setinin özellikleri ve kullanılan yöntemler

1-)MNIST veri seti nedir ve ne için kullanılır

MNIST, "Modified National Institute of Standards and Technology" kısaltmasıdır. MNIST veri seti, el yazısı rakamların (0'dan 9'a kadar) 28x28 piksel boyutundaki siyah-beyaz görüntülerinin bulunduğu bir veri kümesidir. Her bir görüntü, bir rakamı temsil etmektedir ve bu görüntüler elle yazılmış rakamların dijitalleştirilmiş halleridir

bir makine öğrenimi modeli MNIST veri seti üzerinde eğitildiğinde, bu model bir verilen görüntüdeki rakamı tanıyabilir ve sınıflandırabilir. MNIST, makine öğrenimi algoritmalarının temelini anlamak, yeni algoritmaların geliştirilmesi ve mevcut algoritmaların karşılaştırılması için de sıkça kullanılan bir benchmark olarak kabul edilir.

Not: Bu veri seti, genellikle makine öğrenimi ve derin öğrenme alanlarında kullanılsa da, bu projede bu amaçla kullanılmayacak. Biz, bu veri setinin hangi yöntemlerle kullanıldığını, bu yöntemlerin özelliklerini, başarı oranlarını ve en son olarak bunları bir grafiğe aktaracağız.

2-)mnist veri setinin içeriği

Etiketler: Her görüntü, temsil ettiği rakama karşılık gelen bir etiketle ilişkilendirilmiştir.

Sınıf Sayısı: Toplamda 10 sınıf bulunmaktadır, yani rakamlar 0'dan 9'a kadar.

Görüntü Boyutu: Her bir görüntü 28x28 piksel boyutundadır.

Toplam Eğitim Örneği Sayısı: 60.000

Toplam Test Örneği Sayısı: 10.000

Her Sınıftaki Örnek Sayısı (Eğitim ve Test veri setlerinde genellikle benzerdir):

0: ~6.000	, 1: ~6.000	,2: ~6.000	,3: ~6.000	,4: ~6.000
5: ~6.000	,6: ~6.000	,7: ~6.000	,8: ~6.000	,9:~6.000

3-) mnist veri setinde kullanılan yöntemler ve başarı oranları

YÖNTEM	ÖZELLİKLER	BAŞARI ORANI	AVANTAJLAR	DEZAVANTAJLAR
k-en yakın komşu(KNN)	Basit ve etkili	%95 civarı	Hesaplama açısından ucuz	Karmaşık modellerde zayıf performans
Destek Vektör Makineleri (SVM)	Doğrusal olmayan sınırlara uygun	%97-98	Yüksek doğruluk	Hesaplama açısından daha pahalı
Karar Ağaçları	Kurallara dayalı model	%97-98	Yorumlanabilirlik	Karmaşık modellerde zayıf performans
Sinir Ağları	Karmaşık örüntüleri öğrenme yeteneği	%99'un üzerinde	En yüksek doğruluk	Daha fazla veri ve işlem gücü gerektirir

Tabloda hakkında:

- KNN algoritması basit ve hızlıdır, ancak karmaşık modellerde zayıf performans gösterebilir.
- SVM algoritması yüksek doğruluk sağlar, ancak daha fazla hesaplama gücü gerektirir.
- Karar ağaçları yorumlanabilirlik sunar, ancak karmaşık modellerde zayıf performans gösterebilir.
- Sinir ağları en yüksek doğruluğu sağlayabilir, ancak daha fazla veri ve işlem gücü gerektirir.

4-) KNN YÖNTEMİ

KNN, Türkçe'de K-En Yakın Komşu anlamına gelen bir makine öğrenmesi algoritmasıdır. Hem sınıflandırma hem de regresyon problemlerinde kullanılabilir.

KNN Özellikleri:

KNN yöntemi, parametrik olmayan bir algoritmadır. Bu, algoritmanın herhangi bir veri dağılımı varsayımı yapmadığı anlamına gelir.

KNN yöntemi, tembel bir öğrenme algoritmasıdır. Bu, algoritmanın eğitim verilerini ezberlediği ve yeni veri noktaları için tahmin yapmak için bu bilgileri kullandığı anlamına gelir.

KNN yöntemi, yorumlanabilir bir algoritmadır. Bu, algoritmanın nasıl çalıştığını ve hangi faktörlerin tahminleri etkilediğini anlayabildiğimiz anlamına gelir.

Nasıl Çalışır?

Veri seti: Algoritma, önceden etiketlenmiş bir veri seti üzerinde eğitilmelidir. Bu veri seti, her bir veri noktasının özelliklerini ve sınıf etiketini içerir.

Uzaklık hesaplama: Yeni bir veri noktası için, algoritma, eğitim setindeki her veri noktasına olan uzaklığını hesaplar.

En yakın komşular: Algoritma, en yakın K komşuyu belirler. K değeri, kullanıcı tarafından belirlenir.

Sınıflandırma: KNN algoritması, yeni veri noktasının sınıfını, en yakın komşularının sınıfına göre tahmin eder. Sınıflandırma problemlerinde, en yakın K komşunun çoğunluk sınıfı, yeni veri noktasının sınıfı olarak kabul edilir.

Regresyon: Regresyon problemlerinde, yeni veri noktasının değeri, en yakın K komşularının değerlerinin ortalaması olarak tahmin edilir.

KNN Yönteminin Avantajları:

- -Çok yönlü bir algoritmadır ve hem sınıflandırma hem de regresyon problemlerinde kullanılabilir.
- -Herhangi bir veri dağılımı varsayımı yapmaz.
- -Hızlı bir algoritmadır.

KNN Yönteminin Dezavantajları:

- -Gürültüye karşı hassastır.
- -K değerinin seçimi önemlidir ve performansı etkileyebilir.
- -Yüksek boyutlu verilerde hesaplama açısından pahalı olabilir.

KNN Yönteminin Dezavantajlarının Çözümü:

- Gürültüye karşı hassasiyet: Veri setini ön işlemek ve gürültülü verileri ayıklamak, gürültüye karşı hassasiyeti azaltabilir.
- K değerinin seçimi: Çapraz doğrulama tekniği, K değerinin en uygun değerini seçmek için kullanılabilir.
- Yüksek boyutlu veriler: Boyut indirgeme teknikleri, yüksek boyutlu verilerin boyutunu azaltmak için kullanılabilir.

KNN Yönteminin Kullanım Alanları:

- -Görüntü sınıflandırma
- -El yazısı tanıma
- -Tıbbi teşhis
- -Müşteri segmentasyonu
- -Reklam hedefleme

KNN Yönteminin Başarı Oranını Etkileyen Faktörler

Veri Seti:

Boyutu: Daha büyük veri setleri, genellikle daha yüksek başarı oranlarına yol açar.

Kalitesi: Gürültü ve eksik veri içeren veri setleri, başarı oranını düşürebilir.

Denge: Sınıflar dengesizse, algoritma azınlık sınıfını doğru şekilde tahmin edemeyebilir.

K Değeri:

K değeri çok küçükse, algoritma gürültüye karşı hassas hale gelir.

K değeri çok büyükse, algoritma overfitting sorununa maruz kalabilir.

Uzaklık Ölçümü:

Öklid uzaklığı, Manhattan uzaklığı ve Minkowski uzaklığı gibi farklı uzaklık ölçümleri kullanılabilir.

KNN Yönteminin Başarı Oranını Arttırmak İçin Yapılabilecekler:

- Daha büyük ve daha kaliteli bir veri seti kullanmak
- Gürültü ve eksik verileri temizlemek
- Sınıf dengesizliğini gidermek
- En uygun K değerini seçmek
- Veri setine ve probleme en uygun uzaklık ölçümünü seçmek
- Boyut indirgeme teknikleri kullanmak
- Algoritmanın parametrelerini optimize etmek
- Doğru ön işleme tekniklerini kullanmak
- Algoritmayı doğru şekilde uygulamak