T.C.

Ege Üniversitesi

Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

517 Yapay Zeka - Yüksek Lisans Dersi (3+0)

2019-2020 Güz Yarıyılı

MAKİNE ÖĞRENMESİ UYGULAMASI GELİŞTİRME

Hazırlayan

Ozan Türker 91190000051

Kasım 2019

İçindekiler

Proje Hakkında Genel Bilgiler

Problem tanımı

Proje kapsamında Portekiz'e ait "Vinho Verde" türündeki kırmızı şaraba ait beyaz şarapların kalitelerinin kimyasal veriler ışığında sınıflandırılması amaçlanmaktadır.

Veri Seti

Çalışma sırasında Portekiz'e ait "Vinho Verde" türündeki kırmızı şaraba ait veri seti kullanılacaktır. Veri seti 1600 adet şarap verisinden ve her bir şarabın kalite puanından oluşmaktadır. Veri seti UCI sitesinden edinilmiştir. Veri setinin linki kaynakça kısmında belirtilmiştir. Veri seti Paulo Cortez (Univ. Minho), Antonio Cerdeira, Fernando Almeida, Telmo Matos and Jose Reis tarafından 2009 yılında oluşturulmuştur. Veri setinde probleme ait 11 adet öznitelik ve 1 adet kaliteyi belirten çıktı sahası bulunur. Veri setinde bulunan öznitelikler sabit asit, uçucu asit, sitrik asit, artık şeker, klorür, serbest kükürt dioksit, toplam kükürt dioksit, yoğunluk, pH, sülfat, alkoldür. Çıktı değeri ise kalite değişkenidir. Veri setinin ilk 5 satırı aşağıda gösterilmiştir.

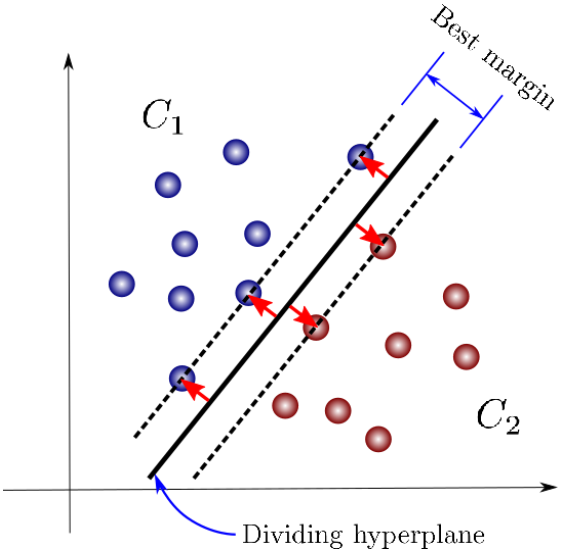
	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur diox...	total sulfur dio...	density	pH	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

Problemin Çözümü

Problem sınıflandırma kullanılarak çözülecektir. Veri setindeki kalite sınıfları kötü orta ve iyi olarak belirlenmiştir. Kalite puanı 1 ile 4 arasındakiler kötü, 5 ve 6 puana sahip olanlar orta ve son olarak 6 ve 9 kalite puana sahip şarap örnekleri iyi olarak sınıflandırılmıştır. Sınıflandırma, veri setinde belirlenmiş olan sınıflara verinin uygun olarak bölünmesi olarak tanımlanabilir. Bu işlemde sınıflar eğitim veri setindeki her bir değer için belirlidir. Sistem öğrenme işlemini eğitim veri setiyle tamamladıktan sonra sisteme daha önce verilmeyen bir veri için o veriyi uygun sınıfa yerleştirmeye çalışır. Bu problemin çözümünde 2 farklı öğrenme yöntemi üzerinde durulacaktır. Bunlardan birincisi Destek Vektör Makinesi'dir (Support Vector Machine - SVM). Diğeri ise Çok Katmanlı Algılayıcı'dır (Multi-Layer Perceptron - MLP).

Destek Vektör Makineleri (Support Vector Machines - SVM)

Destek vektör makinesi yöntemi hem regresyon hem de sınıflandırma için kullanılan bir yöntemdir. Destek vektör makineleri karar sınırlarını tanımlarken karar düzlemlerinden yararlanır. Karar düzlemleri sınıfları birbirinden ayıran düzlemlerdir. Bir veya birden fazla olabilir. Karar düzlemleri sınıflar arasında en fazla boşluğu oluşturacak şekilde yerleştirilmeye çalışılır.Şekilde karar düzlemlerinin örnek bir veride ortaya çıkardıkları grafik gösterilmiştir.

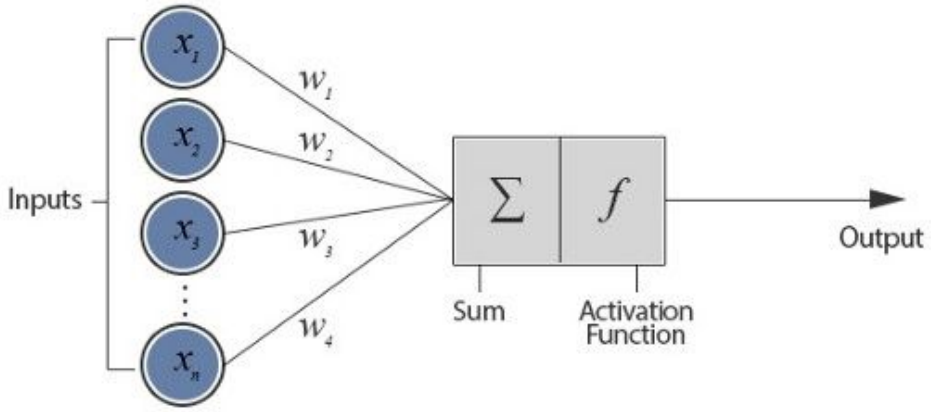


SVM örneği

Avantajları: Yüksek sayıdaki boyuttaki verilerle etkin çalışır. Boyutların sayısının örnek sayısından daha büyük olduğu durumlarda etkilidir. Eğitim noktalarını aynı zamanda karar verme fonksiyonunda da kullanıldığı için hafıza açısından daha az yer kaplar. Çok yönlüdür. Karar fonksiyonu için çeşitli fonksiyonlar kullanabilir.

Dezavantajları: Özelliklerin sayısı örnek sayısından çok daha fazlaysa, Çekirdek işlevlerini seçerken ezberleme durumundan kaçınmak gerekir. SVM'ler doğrudan olasılık tahminleri sağlamazlar, K-kat çapraz doğrulama kullanılarak hesaplanır. Buda maliyete sebep olur.

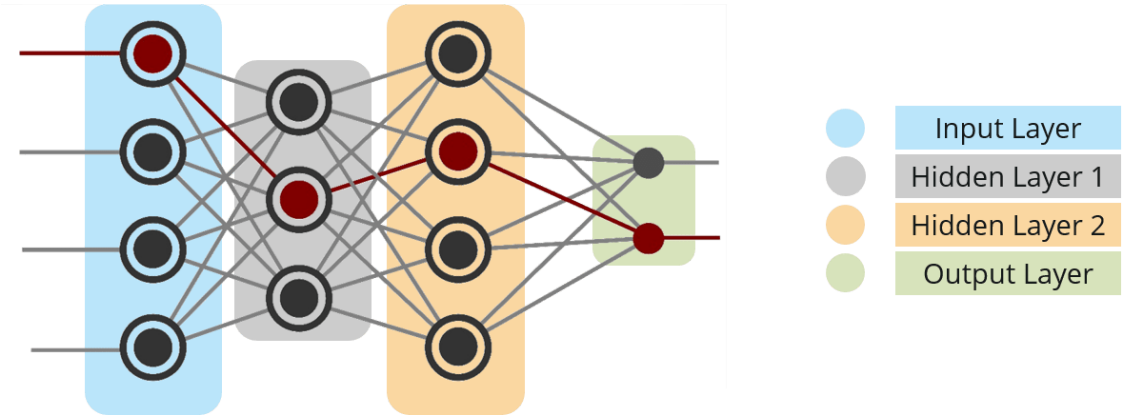
Çok Katmanlı Algılayıcı (Multi-Layer Perceptron - MLP)



Algılayıcı yapısı

Algılayıcı (Perceptron)

Algılayıcı çok katmanlı algılayıcının en küçük birimidir. Genel anlamda bir girdi ve çıktıdan oluşur. İkili (Binary) sınıflandırma işleminde kullanılır. Öğreticili öğrenme problemlerinde kullanılır. Algılayıcı girdileri alarak bir ağırlıkla çarpar ve bunları toplayıp bir aktivasyon fonksiyonu uygular. Aktivasyon fonksiyonu algılayıcının çıktı değerini oluşturur.



Çoklu algılayıcı yapısı

Birçok giriş için bir algılayıcı yeterli olmayabilir. Paralel işlem yapan birden fazla algılayıcıya ihtiyaç duyulduğunda katman kavramı devreye girer. Görüldüğü üzere Tek Perceptron Model'den farklı olarak arada gizli(hidden) katman bulunmaktadır. Giriş katmanı gelen verileri alarak ara katmana gönderir. Gelen bilgiler bir sonraki katmana aktarılır. Ara katman sayısı en az bir olmak üzere probleme göre değişir ve ihtiyaca göre ayarlanır. Her katmanın çıkışı bir sonraki katmanın girişi olmaktadır. Böylelikle çıkışa ulaşılmaktadır.

Deneyler ve sonuçları

Yöntem Adı : Destek Vektör Makinesi (SVM)

Parametre Adı	Kernel	C	Gamma	Veri Miktarı	Eğitim Verisi	Test Verisi	Eğitim Seti Başarı oranı	Test Seti Başarı Oranı
Parametre Değerleri	rbf	10	1	%100	%80	%20	0.834375	0.857701
	rbf	10	10	%100	%80	%20	0.846875	0.956997
	linear	10	1	%100	%80	%20	0.828125	0.824081

Yöntem Adı: Çok Katmanlı Algılayıcı (MLP)

Parametre Adı	Learning Rate (alpha)	Hidden layer size	Solver	Aktivasyon fonksiyonu	Veri Miktarı	Eğitim Verisi	Test Verisi	Eğitim Seti Başarı oranı	Test Seti Başarı Oranı
Parametre Değerleri	1e-5	100, 20	adam	relu	%100	%80	%20	0.824081	0.828125
	1e-5	5, 2	lbfgs	relu	%100	%80	%20	0.856919	0.853125
	1e-5	5, 2	adam	relu	%100	%80	%20	0.850664	0.846875

Normalizasyon

Verilerin veri bütünlüğünü bozacak şekilde, farklı ölçek ya da kod ile kaydedildiği durumlarda başvurulan bir yöntemdir. Buna örnek olarak maaş verisi, gelir, fiyat, tutar gibi finansal verilerin ayrık değerler olarak sistemde tutulmasını örnek gösterebiliriz. 3 farklı şekilde yapılabilir. Bunlar, ondalık ölçekleme, min-max normalleştirme ve Z-score standartlaştırmadır.

Proje çalışması sırasında min-max normalizasyon işlemi yapılarak bütün girdilerin 0-1 aralığına getirilmesi sağlanmıştır. Min-max normalizasyon işleminin formülü aşağıda belirtilmiştir. Normalizasyon işleminden sonra input değişkenlerinin değerleri aşağıdaki ekran görüntüsünde gösterilmiştir.

$$min - max\ normalizasyon = \frac{x - max}{max - min}$$

Min-max normalizasyon formülü

Normalize data

```
min_max_scaler = preprocessing.MinMaxScaler()
dataset_without_output = dataset[dataset.columns[:-1]]
scaled_dataset_without_output = min_max_scaler.fit_transform(dataset_without_output.values)
scaled_dataset_without_output_df = pd.DataFrame(scaled_dataset_without_output, columns= dataset.columns[:-1])
scaled_dataset_without_output_df.head()
```

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur diox...	total sulfur dio...	density	pH	sulphates	alcohol
0	0.247788	0.397260	0.00	0.068493	0.106845	0.140845	0.098940	0.567548	0.606299	0.137725	0.153846
1	0.283186	0.520548	0.00	0.116438	0.143573	0.338028	0.215548	0.494126	0.362205	0.209581	0.215385
2	0.283186	0.438356	0.04	0.095890	0.133556	0.197183	0.169611	0.508811	0.409449	0.191617	0.215385
3	0.584071	0.109589	0.56	0.068493	0.105175	0.225352	0.190813	0.582232	0.330709	0.149701	0.215385
4	0.247788	0.397260	0.00	0.068493	0.106845	0.140845	0.098940	0.567548	0.606299	0.137725	0.153846

Min-max normalizasyondan sonra veri seti

Öz Değerlendirme Tablosu

Madde Numarası	Puan	Var	Açıklama	Tahmini Puan
1) Veri Seti	20	☑	Kırmızı sarap veri seti kullanılmıştır.	20
2) 2 adet yöntem adı	10	☑	Proje kapsamında SVM ve MLP sınıflandırıcıları kullanılmıştır.	10
3) Deneysel Çalışma	30	☑	3'er farklı parametre setiyle 2 yöntem denenmiştir.	30
4) Normalizasyon	20	☑	Normalizasyon açıklanmıştır ve kod parçacığının projeden ekran görüntüsü eklenmiştir.	20
5) Öz Değerlendirme Tablosu	20	☑		20
Toplam	100			100