IST5106 Çok Değişkenli İstatistiksel Yöntemler Take home –FİNAL
Prof. Dr. Fatma NOYAN TEKELİ
Çok Değişkenli İstatistiksel Yöntemler Laboratuvarı- YTÜ- 26.05.2025
Teslim Tarihi: 16.06.2025, saat 12:55

Veri Setleri

1. **Weekly veri seti**
2. **Auto veri seti**

Soru 1: (50 Puan)

Aşağıdaki alt soruları, ***Weekly*** veri setini kullanarak yanıtlayınız. Bu veri seti, 1990 yılının başından 2010 yılının sonuna kadar olan dönemdeki haftalık hisse senedi getirilerine ilişkin bilgileri içermektedir. Veri seti 1.089 gözlemden oluşmaktadır.

a) ***Weekly*** veri setine ilişkin sayısal ve grafiksel tanımlayıcı istatistikleri elde ediniz. Verilerde gözlemlenebilecek herhangi bir örüntü olup olmadığını değerlendiriniz.

b) **Direction** (yön) değişkenini bağımlı değişken, *Lag1*, *Lag2*, *Lag3*, *Lag4*, *Lag5* ve *Volume* değişkenlerini ise bağımsız değişken olarak kullanarak bir lojistik regresyon modeli kurunuz. Model sonuçlarını yorumlayınız. Bağımsız değişkenlerden hangilerinin istatistiksel olarak anlamlı olduğunu belirtiniz.

c) Kurulan modelin tahmin başarımını değerlendirmek amacıyla karışıklık matrisi oluşturunuz ve doğru sınıflandırma oranını hesaplayınız. Elde ettiğiniz bulgular doğrultusunda modelin sınıflandırma hatalarını yorumlayınız.

d) Yalnızca *Lag2* değişkenini açıklayıcı olarak kullanarak, 1990–2008 yıllarına ait gözlemlerle modeli eğitip lojistik regresyon modeli kurunuz. Daha sonra 2009 ve 2010 yıllarına ait veriler(test) üzerinde tahminleme yaparak karışıklık matrisi ve doğru sınıflandırma oranını hesaplayınız.

e) (d) şıkkındaki veri bölünmesini ve değişkeni koruyarak, sınıflandırmayı bu kez *K-en yakın komşu (KNN)* yöntemiyle ve *K=1* olacak şekilde gerçekleştiriniz.

f) **(d)** Aynı veri bölünmesini ve değişkeni kullanarak, sınıflandırmayı *Naive Bayes* yöntemiyle tekrar ediniz.

g) (d), (e) ve (f) şıklarında uygulanan yöntemlerden hangisinin daha iyi performans gösterdiğini, elde edilen doğru sınıflandırma oranlarını karşılaştırarak değerlendiriniz.

h) Farklı değişken kombinasyonları, olası dönüşümler ve etkileşimler dâhil olmak üzere her bir yöntem için çeşitli model denemeleri yapınız. En iyi sonuçları veren yöntemi, kullanılan değişkenleri ve elde edilen karışıklık matrisini raporlayınız. *KNN* yöntemi için farklı *K* değerlerini deneyerek karşılaştırma yapmayı unutmayınız.

Soru 2: (50 Puan)

Bu soruda, ***Auto*** veri seti kullanılarak bir aracın yakıt verimliliğinin yüksek mi yoksa düşük mü olduğunu tahmin etmek amacıyla destek vektör makineleri (*Support Vector Machines – SVM*) uygulanacaktır. Aşağıdaki adımları izleyerek soruları yanıtlayınız.

(a) Araçların yakıt verimliliğini temsil eden Yakıt verimliliği (gas mileage)-*mpg* değişkenine göre, medyan değerin üzerinde yakıt verimliliğine sahip araçları 1, medyanın altında olanları 0 olarak kodlayan ikili (binary) bir hedef değişken oluşturunuz.

(b) Bağımlı değişken olarak (a) şıkkında oluşturduğunuz yeni değişkeni, bağımsız değişkenler olarak ise *mpg* dışındaki tüm sayısal değişkenleri kullanarak, çeşitli **C (cost)** parametre değerleri için **destek vektör sınıflayıcısı (SVM)** uygulayınız. 5 katlı çapraz doğrulama (**cross-validation**) kullanarak her C değeri için doğruluk skorlarını raporlayınız. Sonuçları yorumlayınız.
Not: Modeli oluştururken, yakıt verimliliği değişkenini (mpg) açıklayıcı değişkenler arasına dahil etmeyiniz.

(c) Aynı veri ve hedef değişken ile bu kez **radial** (RBF) ve **polinom** tabanlı çekirdek (kernel) fonksiyonlarını kullanarak SVM uygulayınız.
- RBF çekirdeği için farklı **gamma** ve **C** değerleri,
- Polinom çekirdeği için farklı **degree** (derece) ve **C** değerleri

kullanarak çapraz doğrulama doğruluklarını hesaplayınız ve karşılaştırmalı şekilde yorumlayınız.

(d) (b) ve (c) şıklarında elde ettiğiniz bulguları desteklemek amacıyla, parametre değişimlerine göre doğruluk skorlarını gösteren uygun grafikler oluşturunuz.
İpucu: *Eğer açıklayıcı değişken sayısı ikiden fazlaysa, her seferinde yalnızca iki değişkeni seçerek görselleştirme yapılabilir. Laboratuvar çalışmasında, eğitilmiş SVM'leri çizmek için plot_svm() fonksiyonunu kullanmıştık. Eğer* ***p > 2*** *(yani 2'den fazla açıklayıcı değişken varsa), her seferinde iki değişkenin grafiğini göstermek için features anahtar kelime argümanını kullanabilirsiniz.*

NOT: Bu soruları yanıtlamak için Python, R veya başka uygun bir istatistiksel yazılım aracı kullanabilirsiniz. Cevaplarınızda hem kod çıktıları hem de yorumlarınız birlikte yer almalıdır. Gerektiğinde grafik ve tablolarla analizlerinizi destekleyiniz.