Binary Classification with Logistic Regression

Homework 1

1st PhD Student Önder Görmez
Computer Engineering
Faculty of Electrical and Electronics
Yıldız Technical University
İstanbul, Türkiye
ondergormez@gmail.com

2th Prof. Dr. Mine Elif KARSLIGİL
Computer Engineering
Faculty of Electrical and Electronics
Yıldız Technical University
İstanbul, Türkiye
elif@yildiz.edu.tr

***Abstract*—Bu makale ile bir firmaya iş başvurusunda bulunacak olan kişilerin mülakat sırasında tabi tutuldukları 2 sınav sonucunda işe kabul edilip edilmeyeceklerini tespit eden bir sınıflandırma çalışması yapılacaktır. Çalışma kapsamında lojistik regresyon kullanılacaktır.**

***Index Terms*—regression, logistic regression, classification, binary classification**

I. Introduction

2. bölümde Logistic Regression hakkında kısa bilgiler verilecek, 3. bölümde performans metrikleri tanıtılacak, 4. bölümde datanın ön işlemesi ile alakalı bölümler bulunacak, 5. bölümde regresyon yönteminin uygulanması ve performansların ölçümü yapılacak, 6. bölümde de sonuçlar ve gelecek çalışmalarda yapılacaklardan bahsedilecektir.

II. Logistic Regression Kisa Bilgi

A. Regularization Nedir?

Regularization modelin overfitting olmaması için coefficient tahminini 0 a yakınlaştırmaktır. Böylelikle model overfitting nedeniyle düzgün çalışmadığı zaman, modelin karşmaşıklığını kontrol altında tutabiliriz. Teknik olarak regularization, overfitting i modelin loss fonksiyonuna bir penaltı değeri ekleyerek engeller.

$$Regularization = LossFunction + Penalty$$

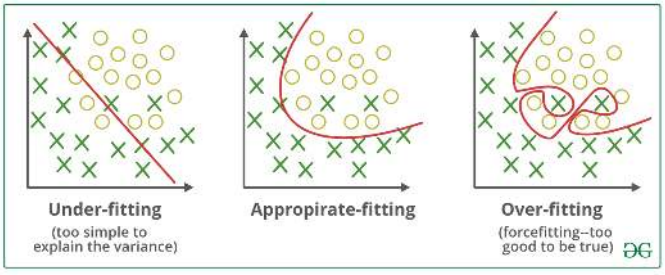

Fig. 1. Under-fitting vs. Appropriate-fitting vs. Over-fitting [1]

B. Sigmoid Aktivasyon Fonksiyonu

Sigmoid aktivasyon fonksiyonu, giriş değerini 0 ile 1 arasında bir değere dönüştüren bir aktivasyon fonksiyonudur. Fonksiyonun formülü (1)'deki gibidir:

$$S(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

Grafiğini çizdirirsek:

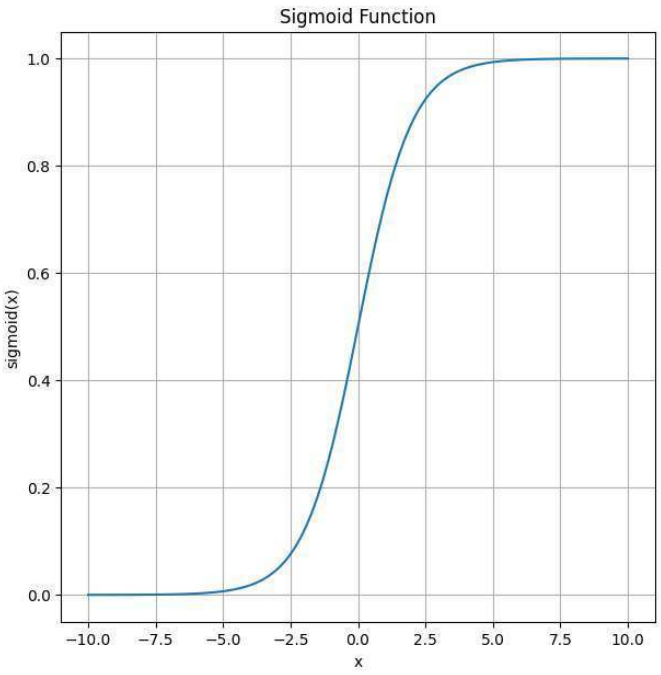

Fig. 2. Sigmoid Aktivasyon Fonksiyonu Grafiği

C. Cross Entropy Loss Fonksiyonu

Cross Entropy Loss fonksiyonu, sınıflandırma modellerinin tahminlerinin doğruluğunu ölçen bir loss fonksiyonudur.

- Log loss olarak da bilinir.
- Çıkışı 0 ile 1 arasında olan modeller için kullanılır.
- Tahmin edilen olasılık değeri, gerçek etiket değerine ne kadar yakınsa, loss değeri o kadar düşük olur.

- Tam tersi, tahmin edilen olasılık değeri, gerçek etiket değerine ne kadar uzaksa, loss değeri o kadar yüksek olur.

Formülü aşağıdaki gibidir:

$$CEL = -\left(y_{\text{true}} \log(y_{\text{predicted}}) + (1 - y_{\text{true}}) \log(1 - y_{\text{predicted}})\right) \tag{2}$$

D. Logistic Regression

Lojistik regresyon, binary classification için oldukça yaygın kullanılan bir eğiticili öğrenme algoritmasıdır.

- Lojistik regresyon, giriş verilerini bir doğrusal regresyon modeline besler ve ardından çıktıyı bir sigmoid fonksiyonuna geçirir. Bu, çıktının 0 ile 1 arasında olmasını sağlar ve bu çıktıyı bir eşik değeriyle karşılaştırarak sınıflandırma yapılabilir.
- Bir e-mailin spam olup olmadığını belirlemek, bir kullanıcının bir ürünü satın alıp almayacağını tahmin etmek gibi binary classification problemlerinde kullanılabilir.

Regularization modelin overfitting olmaması için coefficient tahminini 0 a yakınlaştırmaktır.

- Böylelikle model overfitting nedeniyle düzgün çalışmadığı zaman, modelin karşmaşıklığını kontrol altında tutabiliriz.
- Teknik olarak regularization, overfitting i modelin loss fonksiyonuna bir penaltı değeri ekleyerek engeller.

III. Performans Ölçüm Metrikleri

Çalışma yapılırken kullanılacak olan performans ölçüm metrikleri ve neleri ifade ettiklerine kısaca bir bakalım. Öncesinde kullanılacak olan tanımlardan bahsetmek gerekirse;

- True Positive (TP): Modelin pozitif olarak doğru tahmin ettiği örneklerdir.
- True Negative (TN): Modelin negatif olarak doğru tahmin ettiği örneklerdir.
- False Positive (FP): Modelin pozitif olarak yanlış tahmin ettiği örneklerdir.
- False Negative (FP): Modelin negatif olarak yanlış tahmin ettiği örneklerdir.

A. Accuracy

Accuracy, doğru sınıflandırılan örneklerin toplam örnek sayısına oranı ile hesaplanır. Modelin genel performansını ölçümlemek için kullanılan temel bir metriktir. Sınıflar arasında dengeli bir dağılım varsa accuracy iyi bir performans ölçütü olabilir. Fakat dengesiz veri setlerinde kullanılmak için uygun değildir. Örneğin fraud detection gibi unbalanced data set ler üzerinde accuracy yüksek çıksa da modelin genel performansı hakkında doğru bilgi vermez.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

B. Precision (Kesinlik)

Precision, modelin pozitif tahminlerinin ne kadarının doğru olduğunu gösterir. Precision, yanlış pozitiflerin (False Positive) maliyetinin yüksek olduğu durumlarda özellikle önemlidir. Yanlış pozitifler modelin pozitif olarak tahmin ettiği ancak gerçekte negatif olan örneklerdir. Örneğin tıbbi sonuçları olan testlerde, precision yüksek olmalıdır çünkü yanlış pozitifler (sağlıklı kişilerin hasta olarak değerlendirilmesi) gereksiz yere endişe, korkuya neden olduğu gibi sonrasında yanlış tedavi ile sağlığın bozulmasına kadar giden kötü sonuçlar doğurabilir.

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

C. Recall (Duyarlılık)

Recall, modelin pozitif sınıfı ne kadar iyi tespit edebildiğini gösteren bir metriktir. Yanlış negatiflerin (False Negative) maliyetinin yüksek olduğu durumlarda özellikle önemlidir. Örneğin tıbbi sonuçları olan testlerde, recall yüksek olmalıdır çünkü yanlış negatifler (hasta kişilerin sağlıklı olarak değerlendirilmesi) tedavinin gecikmesine ve hastanın sağlığının geri dönülmez bir şekilde kaybedilmesine ve hatta ölümlere bile yol açabilir. Geriye dönülmez sonuçlar doğurabilir.

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

D. F1 Score

F1 Score, precision (kesinlik) ve recall (duyarlılık) metriklerinin harmonik ortalamasıdır. Özellikle dengesiz veri setlerinde (pozitif ve negatif sınıflar arasında büyük bir dengesizlik olduğunda) kullanışlıdır. Bu tarz durumlarda accuracy metriğine değil F1 Score metriğine bakılmalıdır.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{6}$$

IV. Datanın Ön İşlenmesi

Data içerisinde 100 kişinin 2 sınav sonucuna göre işe kabul ve ret edilme durumları bilgisi verilmiştir.

Değerler 0 - 100 arasında değişmektedir.

A. Verinin Görselleştirilmesi

Veri içerisinde kabul ve ret olarak 2 tane sınıf vardır. İncelememizde 60 tane kabul edilmiş, 40 tane de reddedilmiş iş başvurusu olduğunu görüyoruz.

Fig. 3'de tüm veri setinin dağılımı ve kabul ve red alan örneklerin dağılımı görülmektedir. Fig. 3'i yorumlamak gerekirse;

- Herhangi **bir sınavdan çok yüksek almak** başarılı olmayı **garanti etmez**. 1 sınavdan çok iyi not alıyorsanız diğer sınavdan en az 40-45 almanız lazım ki başarılı olasınız.
- **2 sınavdan da 70 in üzerinde** not alırsanız **başarılı olma ihtimaliniz çok yüksektir**.

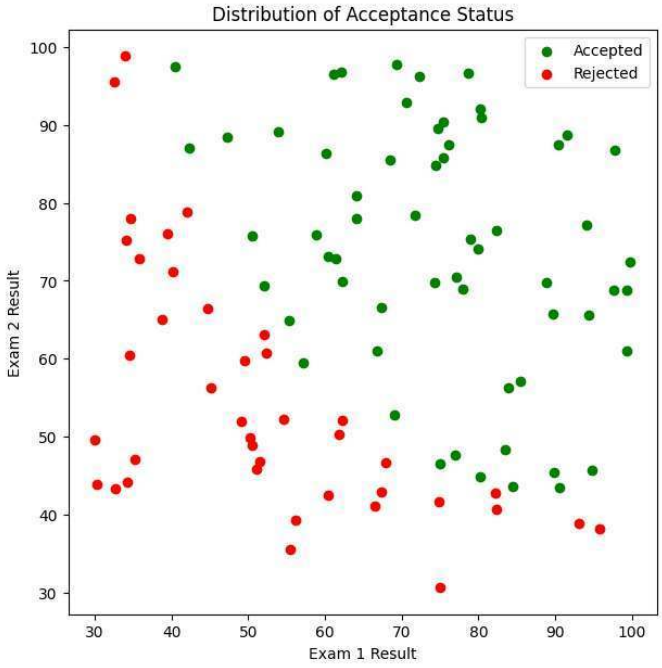

Fig. 3. Distribution of All Data

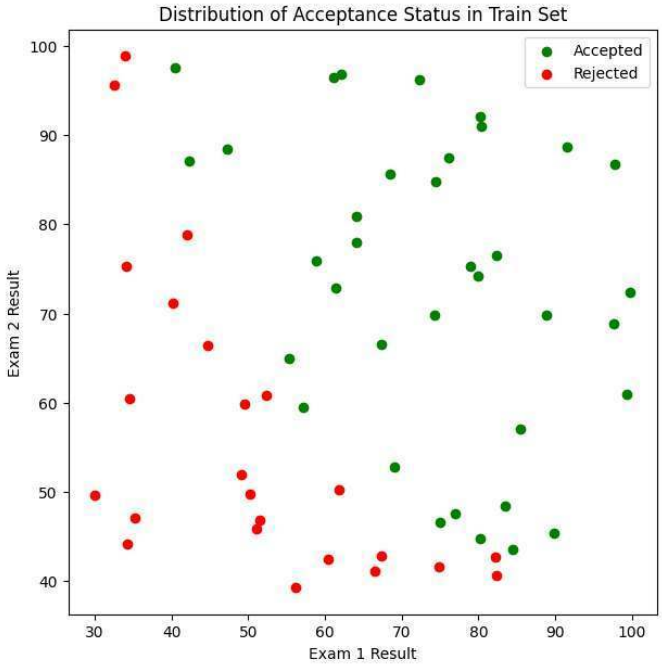

Fig. 4. Train Set

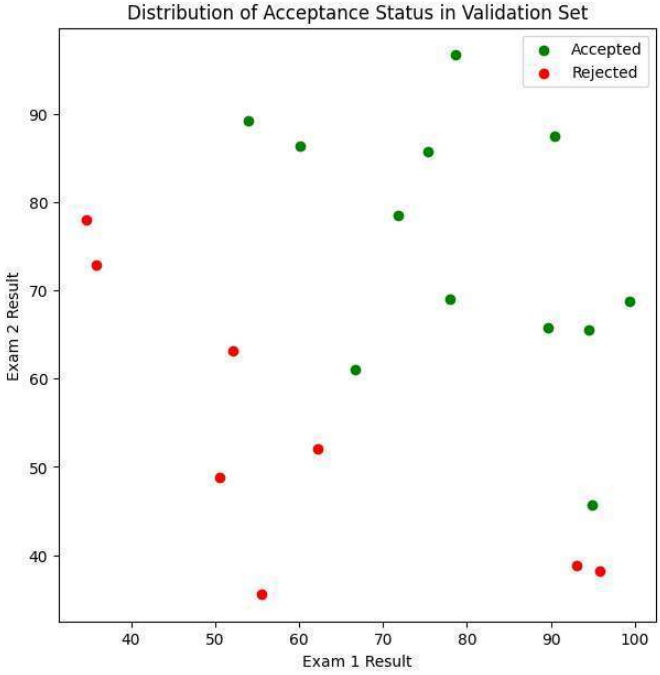

Fig. 5. Validation Set

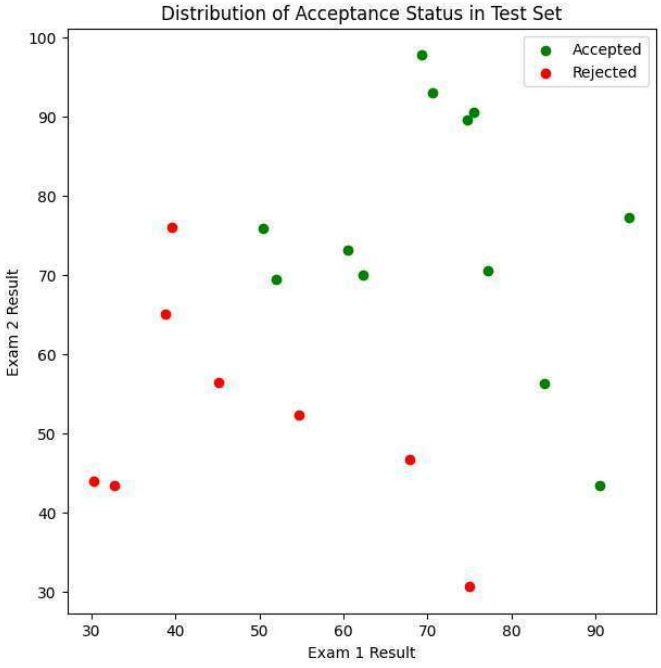

Fig. 6. Test Set

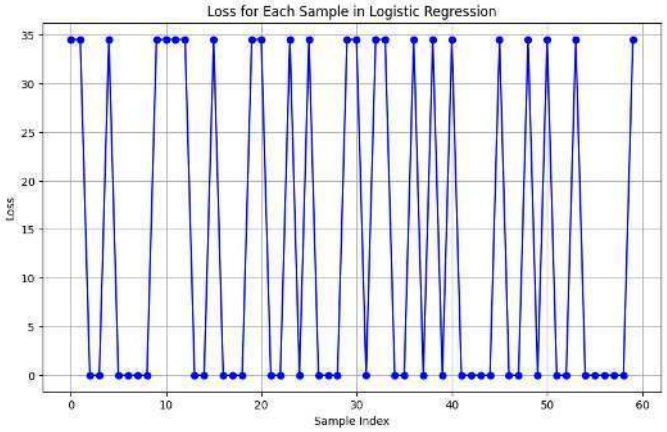

Fig. 7. Cross Entropy Loss

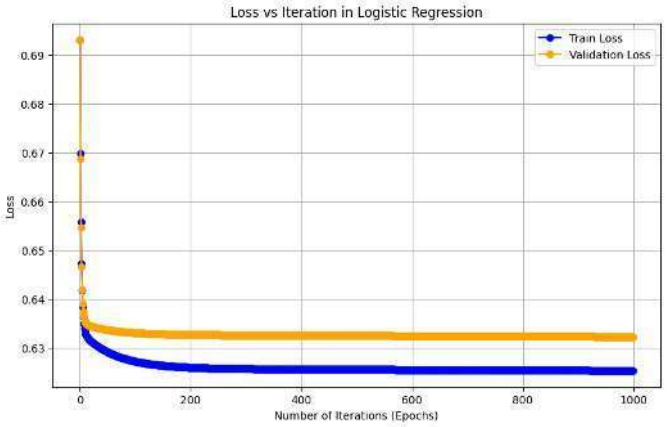

Fig. 8. Loss vs 1000 Iterations

B. Verinin Training, Validation ve Test Setlerine Ayrılması

Veriseti %60 training, %20 validation ve %20 test seti olarak ayrılmıştır.

- Fig. 4 training seti göstermektedir.
- Fig. 5 validation seti göstermektedir.
- Fig. 6 test seti göstermektedir.

V. Lojistik Regresyonun Uygulanması ve Performans Ölçümü

A. Loss vs Sample Grafiği

Fig. 7 üzerinde cross-entropy loss değerinin her sample için ayrı ayrı değerini görüyoruz.

B. Loss vs Epoch Grafikleri

Algoritmamızı learning rate 0.0001 ve iterasyon sayısı 1000 iken çalıştırdığımızda Fig. 8 daki gibi bir grafik elde edilmektedir. Bu grafik bize aslında modelin çok daha az iterasyonda fit edebildiğini ve loss un sıfıra yaklaştığını göstermektedir.

Yukarıda belirtilen nedenlerle iterasyon sayısı 100 değerine indirilerek tekrar modelimizi çalıştırırsak Fig. 9 daki gibi bir sonuç elde ederiz.

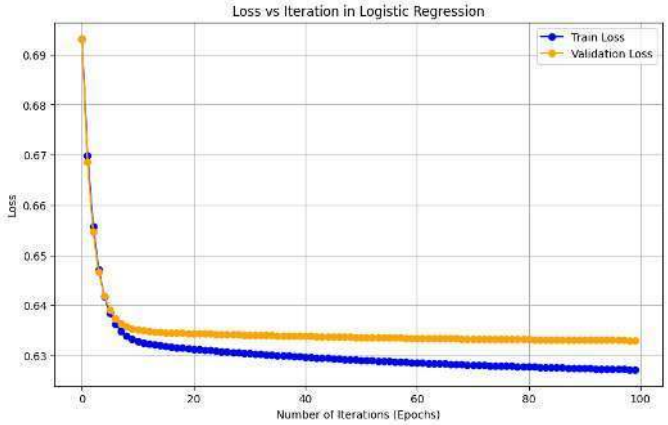

Fig. 9. Loss vs 100 Iterations

C. Performans Sonuçları

Kendi geliştirdiğimiz lojistik regresyon yönteminin uygulanması ile train, test ve validation setleri üzerinde elde edilen performans sonuçları Tablo I üzerinde gösterilmiştir.

Data	Accuracy	Precision	Recall	F1 Score
Train	0.900	0.917	0.917	0.917
Validation	0.900	0.857	1.000	0.923
Test	0.900	1.000	0.833	0.909

TABLE I
Performance metrics for Train, Validation, and Test datasets

D. Test Datası için Confusion Matrix Oluşturulması

Fig. 4'den anlaşılacağı üzere model hiç görmediği test datası üzerinde de iyi bir performans sergilemektedir. Bu çalışmada positif olan sınıf işe kabul edilenler olduğu için 2 tane false negatif olduğunu görüyoruz. Bu da recall ın neden %83.33 çıktığını açıklıyor.

VI. Sonuçlar ve Gelecek Çalışmalar

Liner Regresyonun ikili sınıflandırma (binary classification) problemleri için güzel çalıştığını ve doğrulunun yüksek çıktığını söyleyebiliriz. İncelediğimiz data balanced ve dengeli bir data olduğu için validation seti üzerinde optimizasyon ve düzenlemeye gitmeden test seti üzerinde yüksek başarım elde edilmiştir. Bu çalışmaya konu olmayan dengesiz veri seti (unbalanced dataset) üzerinde nasıl bir sonuç vereceği hakkında bir bilgimiz yoktur. İleride yapılacak olan çalışmalarda bu durum incelenebilir.

Acknowledgment

Bu makalenin oluşturulmasında kulllanılan veri setini bize sağlayan ve makalenin yayına hazır hale getirilmesinde geri bildirimlerini ve desteklerini esirgemeyen sayın hocamız Prof. Dr. Mine Elif KARSLIGİL'e teşekkürü bir borç bilirim.

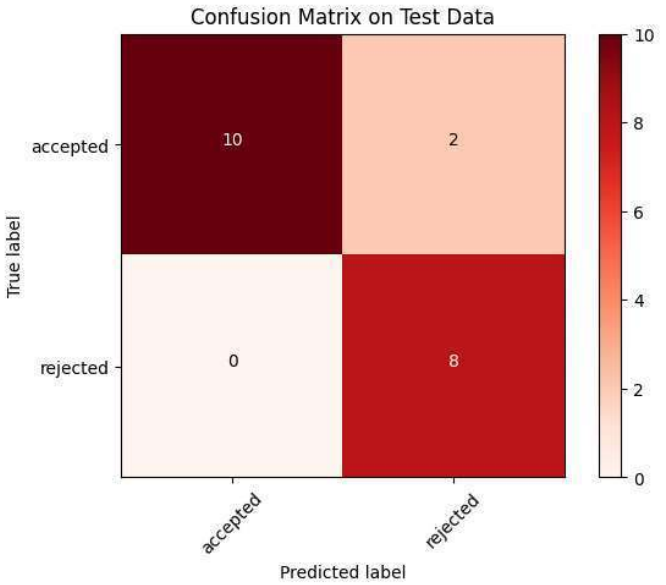

Fig. 10. Confusion Matrix on Test Data

REFERENCES

[1] "Overfitting," GeeksforGeeks. [Online]. Available: https://media.geeksforgeeks.org/wp-content/uploads/20190523171704/overfitting_21.png. [Accessed: 18-Nov-2024].

[2] K. Kushal, "Logistic Regression from Scratch," Medium, 12-Dec-2018. [Online]. Available: https://medium.com/@koushikkushal95/logistic-regression-from-scratch-dfb8527a4226. [Accessed: 18-Nov-2024].

[3] "Machine Learning Equations in LaTeX," blmoistawinde.github.io. [Online]. Available: https://blmoistawinde.github.io/ml_equations_latex/. [Accessed: 18-Nov-2024].

[4] F. K. Alarfaj, I. Malik, H. U. Khan, N. Almusallam, M. Ramzan, and M. Ahmed, "Credit Card Fraud Detection Using State-of-the-Art Machine Learning and Deep Learning Algorithms" *IEEE Access*, vol. 10, pp. 1-10, Apr. 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9755930. [Accessed: 18-Nov-2024].