Akbank Makine Öğrenmesi Bootcamp: Yeni Nesil Proje Kampı

Bu projede katılımcılar makine öğreniminde Gözetimli öğrenme alanında bir çalışma gerçekleştireceklerdir.

Öğrenciler tahmin amaçlı olarak verileri kategoriler halinde sınıflandırmayı veya girdi özelliklerine dayalı olarak sürekli değerleri tahmin etme üzerine bir proje gerçekleştirecektir.

Bu proje, öğrencilere yapay zeka ve makine öğrenmesi alanında Gözetimli Öğrenme konusunda veri analizi, model geliştirme ve değerlendirme teknikleri konusunda pratik deneyim kazandırmayı amaçlamaktadır.

Bonus: Dokümanın sonunda değinileceği üzere, katılımcılar Gözetimli Öğrenme'ye ek olarak aynı veri setiyle Gözetimsiz Öğrenme çalışabilir, uçtan uca projesini GPU kütüphaneleriyle çalıştırabilir, projesini deploy ederek web arayüzünden sunabilir.

Proje Konusu

Seçeceğiniz veri setinde Gözetimli Öğrenme teknikleriyle bir proje geliştireceksiniz. **Bootcamp sonunda, projenizi Kaggle ortamında sergiliyor olacaksınız. Geliştirme aşamasında VSCode, PyCharm vs. Her türlü IDE'yi kullanabilirsiniz. Ancak projenizin finalinde, notebookunuz(veya notebooklarınız) mutlaka Kaggle'da yer almak zorunda. Kaggle linklerini de, GitHub reponuzdaki README.md dosyanıza koymak zorundasınız.**

Bonus kısmını da projenize dahil etmek isterseniz, Gözetimsiz Öğrenme için de aynı veri setini kullanacaksınız.

1 - Veri Setinizi Seçin

Veri Kriterlerimiz:

- Filtrelerken veri seti boyutunun 10 MB'den büyük olması gerekmekte veya;
- Veri seti en az 10 bin veri noktası içermeli.

Dokümanın sonunda birtakım veri seti kaynakları bulacaksınız, ancak kriterlerimizi karşıladığı sürece bu kaynakların dışından da herhangi bir veri seti kullanabilirsiniz.

2 - Bir Notebook Dosyası Oluşturun

* Projeniz, .ipynb uzantılı dosyalarda gerçekleştirilmeli.

* Bu dosyalarda, kod satırlarının yanı sıra projenizin teknik detaylarını açıkladığınız yorum hücreleriniz olmalı.

3 - Bir GitHub Reposu & README.md Dosyası oluşturun

* Bir Github reposu oluşturmalısınız.

* Bu repoda upload edeceğiniz .ipynb uzantılı proje dosyanız ve README.md dosyanız olmalı.

* Detaylı teknik anlatımlarınız .ipynb dosyasında yer almalıdır.

*** README.md dosyanızda, projenizin final çıktılarına yer vermelisiniz. Hangi algoritmayı seçtiniz, geliştirdiğiniz proje veri seti üzerinde hangi problemi çözdü, gerçek hayatta nasıl işimize yarayabilir, bu proje daha da geliştirmek istense daha başka nasıl geliştirilebilir gibi sözel anlatımlara da README.md dosyanızda yer vermelisiniz.**

*** Kaggle'daki notebook linkiniz de README.md dosyasında yer almalı.**

4 - Keşifsel Veri Analizi (EDA - Exploratory Data Analysis)

Pandas, matplotlib, seaborn vb. ilgili kütüphaneleri kullanarak seçtiğiniz veri setini açıklayınız. Görselleştirme kütüphaneleri ile görseller, grafikler oluşturup, veri setine ait genel bilgileri aktarın. Analizlerinizi açıklamak için .ipynb içerisinde yorum hücrelerini kullanmalısınız.

Kaynak: https://gokerguner.medium.com/machine-learning-1-7d4581caa291

5 - Veri Ön İşleme

Veri kümenize bağlı olarak gerekli ön işleme adımlarını yapmalısınız. Veri setini temizlemek, normalleştirme, label-encoding veya one-hot encoding (kategorik değişkenleriniz varsa), veri kümenizi eğitim ve test kümelerine bölmek vb.

6 **- Algoritma Seçimi & Hiperparametre Optimizasyonu**

Örnek Algoritmalar

Gözetimli Öğrenme Algoritmaları:

1. **Doğrusal Regresyon (Linear Regression)**
2. **Lojistik Regresyon (Logistic Regression)**
3. **Karar Ağaçları (Decision Trees)**
4. **k-En Yakın Komşu (k-Nearest Neighbors - KNN) Sınıflandırıcısı**
5. **Destek Vektör Makineleri (Support Vector Machines - SVM)**

(Bonus)Gözetimsiz Öğrenme Algoritmaları:

1. **k-Ortalama (k-Means) Kümeleme**
2. **Apriori Algoritması**
3. **Hiyerarşik Kümeleme (Hierarchical Clustering)**
4. **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**
5. **Gaussian Karışım Modelleri (Gaussian Mixture Models - GMM)**

* Her iki öğrenme türünde de, projenize uygun birkaç model seçin ve bunları ön işlemeden geçirdiğiniz verilerinizle eğitin.
* Çapraz doğrulamayı(cross validation) kullanarak seçilen modellerin performanslarını inceleyin.
* En iyi performansı gösteren model ile ilerleyin.
* Önceki adımda seçilen modelin hiper parametrelerini uygun bir yöntemle optimize edin (Grid Search, Randomized Search veya uygun gördüğünüz herhangi biri).

Not: Bu adımda, birden fazla model denemek yerine, direkt olarak hiperparametre optimizasyonuyla seçtiğiniz bir modelin parametrelerini değiştirerek performans artırma yoluna da gidebilirsiniz. README.md dosyasında neden o modeli seçtiğinize dair açıklamalarınızı da bekliyor olacağız.

Proje dosyasında, yalnızca ilerlemeye karar verdiğiniz modele ve hiperparametrelere yer verebilirsiniz.

7 - Model Değerlendirme

Uygun model değerlendirme yöntemlerini kullanarak optimize edilmiş modeli değerlendirin.

- Örneğin Regresyon için kayıp hesaplaması: Ortalama Karesel Hata(Mean Squared Error), Ortalama Mutlak Hata(Mean Absolute Error) vb. ile,
- Sınıflandırma için karışıklık matrisi(Confusion matrix) oluşturarak:
- Doğruluk(accuracy), kesinlik(precision), duyarlılık(recall), F1 puanı(F1 score) özelliklerine yer vererek.

https://gokerguner.medium.com/machine-learning-2-korelasyon-matrisi-özellik-seçimi-sınıfların-dengesizliği-karar-ağaçları-af993bd8ea66

8 - Sonuç

Çalışmalarınızın sonucunda, reponuz şu şekilde görünmelidir: https://github.com/gokerguner/example-repo

---

BONUS

1) Aynı veri seti üzerinde olmak kaydıyla, bir Gözetimsiz Öğrenme çalışması da gerçekleştirebilirsiniz.

2) Çalışmanızı, GPU üzerinde çalışan kütüphanelerle tekrarlayabilir, aradaki hız farkı vb. Kavramlara dair çıkarımlarınızı paylaşabilirsiniz. Kaynak

3) Çalışmanızı bir arayüzden erişilebilecek şekilde deploy edebilirsiniz.

Veri Setleri

1. *Kaggle* (https://www.kaggle.com/datasets)
2. *UCI Machine Learning Repository* (https://archive.ics.uci.edu/ml/index.php)
3. *Google Dataset Search* (https://datasetsearch.research.google.com/)
4. *AWS Public Datasets* (https://registry.opendata.aws/)
5. *Microsoft Research Open Data* (https://msropendata.com/)
6. *Data.gov* (https://www.data.gov/)
7. *Data.world* (https://data.world/)
8. *FiveThirtyEight* (https://data.fivethirtyeight.com/)
9. *The World Bank Data* (https://data.worldbank.org/)

10. *GitHub* - Awesome Public Datasets Collection (https://github.com/awesomedata/awesome-public-datasets)
11. *NASA Planetary Data System* (https://pds.nasa.gov/datasearch/data-search/)
12. *Reddit r/datasets* (https://www.reddit.com/r/datasets/top/?sort=top&t=all&rdt=64096)
13. *OpenML* (https://www.openml.org/search?type=data&sort=runs&status=active)
14. *PaperWithCodes* (https://paperswithcode.com/datasets)
15. *Hugging Face Datasets* (https://huggingface.co/datasets)
16. *TensorFlow Datasets*(https://www.tensorflow.org/datasets)
17. *Yelp Veri Seti* (https://www.yelp.com/dataset)
18. *Amazon İnceleme Veri Seti* (https://nijianmo.github.io/amazon/index.html)