BM5702 MAKİNE ÖĞRENMESİNE GİRİŞ

Hafta 1

Doç. Dr. Murtaza CİCİOĞLU

Derse Giriş

- Gün ve Saati: Pazartesi – 09:40 – 12:00
- Görüşme Zamanları: Çarşamba: 13:00 – 14:00 (Randevu alınarak)
- İletişim: murtazacicioglu@uludag.edu.tr
- Notlandırma:
  - Vize Sınavı: %40 + Final Sınavı: %60
  - Vize → Sunum + En az bir uygulama + ilgili 5 Adet Makale İncelemesi (IEEE, Elseiver, Wiley, Springer) SCI, SCI-Exp – En geç Cumartesi 23:59'a kadar classroom a yüklenmesi
  - Final + Proje → Özgün bir veri setinde en az 5 farklı ML algoritmaları kullanılarak makale hazırlamak (Benzerlik oranı %20'nin altında)

Ana başlıkların planlanması

- Ana metnin hazırlanması
  - Başlık seçimi
  - Öz ve anahtar kelimeler
  - Giriş
  - Materyal ve yöntem
  - Bulgular
  - Tartışma ve Sonuç
  - Kaynakça

Final Ödevi

- Konunun özgünlüğü (problem cümlesi, literatürde bu probleme önerilen çözümler ve sizin önerdiğiniz çözüm)
- Akademik Yazım Tarzı (Özet, Giriş, İlgili Çalışmalar, Önerilen Sistem, Performans Sonuçları, Sonuç, Kaynakça)
- Giriş bölümünde özellikle problem ve alana katkılar net olarak açıklanmalı
- Literatür Taraması: (konu ile ilgili en az 15 çalışma) yazarlar neler yapmış, hangi teknikleri kullanmış ve neler önermiş, eksikler nelerdir, bölümün sonunda bir tablo halinde çalışmalar karşılaştırılabilir

Final Ödevi

- Önerilen çalışmanın kuramsal açıklaması, matematiksel modeller,
- Performans sonuçları en az 5 farklı makine öğrenmesi algoritması kullanılarak sonuçların karşılaştırılması ve görselleştirilmesi
- Referanslar IEEE formatında verilecek

Derse Giriş

- Kaynak Kitaplar:
  1. Introduction to Machine Learning, Ethem ALPAYDIN , MIT
  2. Introduction to Machine Learning with Python: A Guide for Data Scientists - Sarah Guido, Andreas C. Mueller O'Reilly Media (2016)
  3. Mark E. Fenner - Machine Learning With Python for Everyone-Addison-Wesley Professional_Pearson education (2020)
  4. Intro to Python for Computer Science and Data Science: Learning to Program with AI, Big Data and the Cloud, Deitel, Pearson Education
- Harici Kaynaklar;
  1. http://www.saedsayad.com
  2. http://www.veridefteri.com
  3. https://erdincuzun.com/makine_ogrenmesi/

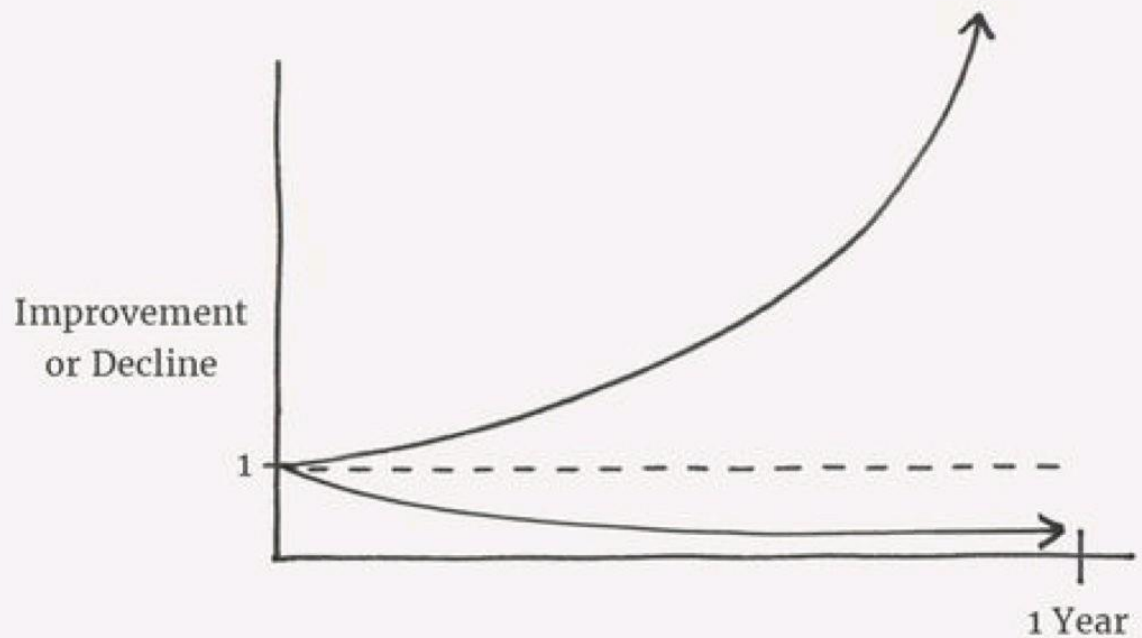
Konular

Hafta	Teori
1	Makine Öğrenmesine Giriş
2	Karar Ağaçları
3	Örnek Tabanlı Öğrenme
4	Bayesçi Öğrenme
5	Lojistik Regresyon
6	Sinir Ağları
7	Destek Vektör Makineleri
8	Kümeleme, k-ortalama
9	Maksimum Beklenti, Gauss Karışım
10	Topluluk Öğrenmesi
11	Rastgele Orman
12	Çekişmeli Öğrenme
13	Takviyeli Öğrenme
14	LDA ve PCA

The Power of Tiny Gains

1% better every day $1.01^{365} = 37.78$

1% worse every day $0.99^{365} = 0.03$



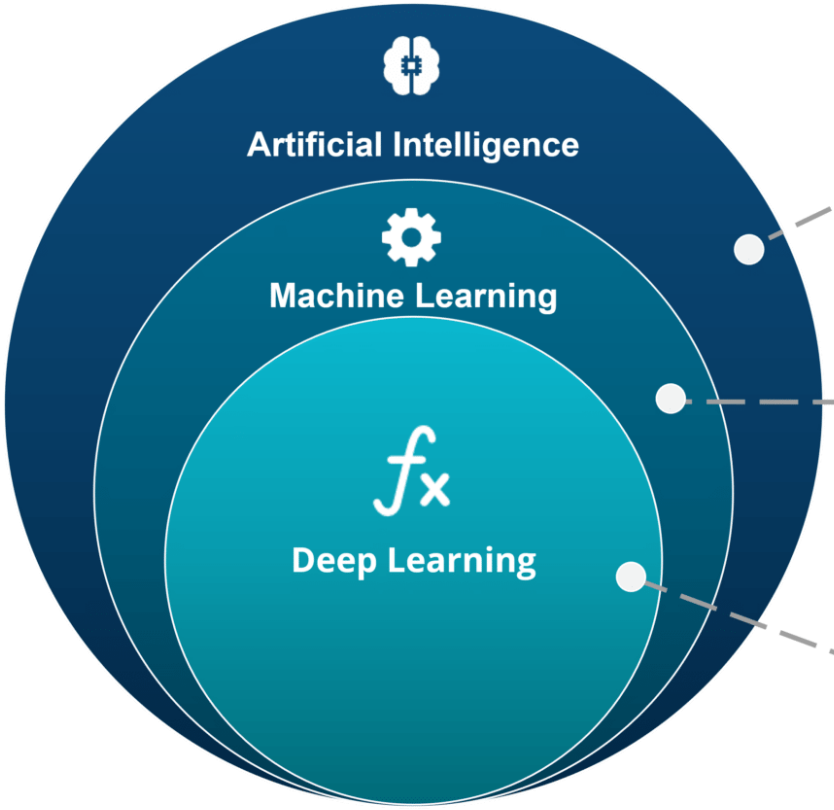
Machine Learning: Classification, Regression and Clustering


Makine Öğrenmesi
Denetimli
Regresyon
Doğrusal Regresyon
Çok Değişkenli Uyarlanabilir Regresyon Eğrileri(MARS)
Locally Weighted Scatterplot Smoothing (LOWESS)
Destek Vektör Regresyon (SVR)
Karar Ağacı RegresyonCART)
Rastgele Ağaçlar Regresyon
K-En Yakın Komşu Regresyon (KNN)
Sinir Ağları
İleri Beslemeli Sinir Ağlar (FFNN)
Uzun Ömürlü Kısa Dönem Bellek Ağları (LSTM)
Geçitli Yinelemeli Birim (GRU)
Evrişimsel Sinir Ağlar (CNN)
Üretken Çekişmeli Ağlar (GAN)
Otokodlayıcılar
Denetimli Boyut Azaltma
Doğrusal Ayırma Analizi
Sınıflandırma
K-En Yakın Komşu Sınıflandırma (KNN)
Aşırı Gradyan Hızlandırma (XGBoost)
Gradyan Hızlandırma
Uyarlamalı Hızlandırma (AdaBoost)
Rastgele Orman Sınıflandırma
Karar Ağacı Sınıflandırma (CART)
Destek Vektör Makineleri (SVM)
Naive Bayes
Lojistik Regresyon
Denetimsiz
Kümeleme
K-Ortalamalar
EM-Algoritması
Gauss Karışım Modelleri (GMM)
Hiyerarşik Kümeleme
Gürültülü Veritabanlarında Yoğunluk Tabanlı Uzamsal Kümeleme (DBSCAN)
Destek Vektör Kümeleme (SVC)
Boyut Azaltma
Temel bileşenler Analizi (PCA)
Çok Boyutlu Ölçekleme (MDS)
İzometrik Haritalama Gömülmesi
Yerel Doğrusal Gömülme
Modifiye Edilmiş Yerel Doğrusal Gömülme
Apriori Algoritması
Birliktelik Kuralları
Pekiştirmeli
Monte Carlo Yöntemler
Zamansal-Fark Öğrenme
Politika Gradyanları
Proksimal Politika Optimizasyonu (PPO)
SARSA (Durum-Eylem-Ödül-Durum-Eylem)
Q-Öğrenmesi
Deep Q Neural Network (DQN)
Yarı-Denetimli
İletken Destek Vektör Makinesi (TSVM)
Etiket Yayma
Etiket Yayılım Algoritması
Kendi Kendine Eğitim Sınıflandırıcısı
Diğerleri
Olasılıksal Grafik Modelleri
Bayesçi İnanç Ağları

Introduction to Machine Learning

- **machine learning**—one of the most exciting and promising subfields of **artificial intelligence**
- You'll see how to quickly solve challenging and intriguing problems that novices and most experienced programmers probably would not have attempted just a few years ago.
- Big, complex topic.

Introduction to Machine Learning



ARTIFICIAL INTELLIGENCE
A technique which enables machines to mimic human behaviour

MACHINE LEARNING
Subset of AI technique which use statistical methods to enable machines to improve with experience

DEEP LEARNING
Subset of ML which make the computation of multi-layer neural network feasible

What Is Machine Learning?

- Can we really make our machines (that is, our computers) learn?
- “Secret sauce” is **data, and lots of it**
- Rather than programming expertise into our applications, we program them to learn from data
- Build working machine-learning models then use them to make remarkably accurate predictions

Challenges

Challenges in Machines Learning

While Machine Learning is rapidly evolving, making significant strides with cybersecurity and autonomous cars, this segment of AI as whole still has a long way to go. The reason behind is that ML has not been able to overcome number of challenges. The challenges that ML is facing currently are:

Quality of data: Having good-quality data for ML algorithms is one of the biggest challenges. Use of low-quality data leads to the problems related to data preprocessing and feature extraction.

Time-Consuming task: Another challenge faced by ML models is the consumption of time especially for data acquisition, feature extraction and retrieval.

Lack of specialist persons: As ML technology is still in its infancy stage, availability of expert resources is a tough job.

No clear objective for formulating business problems: Having no clear objective and well-defined goal for business problems is another key challenge for ML because this technology is not that mature yet.

Issue of overfitting & underfitting: If the model is overfitting or underfitting, it cannot be represented well for the problem.

Curse of dimensionality: Another challenge ML model faces is too many features of data points. This can be a real hindrance.

Difficulty in deployment: Complexity of the ML model makes it quite difficult to be deployed in real life.

Prediction

- Improve **weather forecasting** to save lives, minimize injuries and property damage
- Improve **cancer diagnoses** and **treatment regimens** to save lives
- Improve **business forecasts** to maximize profits and secure people's jobs
- **Detect fraudulent credit-card purchases** and **insurance claims**
- Predict **customer "churn"**, what prices houses are likely to sell for, ticket sales of new movies, and anticipated revenue of new products and services
- Predict the **best strategies for coaches and players** to use to win more games and championships
- All of these kinds of predictions are happening today with *machine learning*.

Popular Machine Learning Applications

- Anomaly detection
- Chatbots
- Classifying emails as spam or not spam
- Classifying news articles as sports, financial, politics, etc.
- Computer vision and image classification
- Credit-card fraud detection
- Customer churn prediction
- Data compression
- Data exploration
- Data mining social media (like Facebook, Twitter, LinkedIn)
- Detecting objects in scenes
- Detecting patterns in data
- Diagnostic medicine
- Facial recognition
- Insurance fraud detection
- Intrusion detection in computer networks
- Handwriting recognition
- Marketing: Divide customers into clusters
- Natural language translation (English to Spanish, French to Japanese, etc.)
- Predict mortgage loan defaults
- Recommender systems ("people who bought this product also bought…")
- Self-Driving cars (more generally, autonomous vehicles)
- Sentiment analysis (like classifying movie reviews as positive, negative or neutral)
- Spam filtering
- Time series predictions like stock-price forecasting and weather forecasting
- Voice recognition

Scikit-Learn

- Scikit-learn machine learning library
- Scikit-learn, also called sklearn, conveniently packages the most effective machine-learning algorithms as estimators.
- Each is encapsulated, so you don't see the intricate details and heavy mathematics of how these algorithms work.
- With scikit-learn and a small amount of Python code, you'll create powerful models quickly for analyzing data, extracting insights from the data and most importantly making predictions.

Scikit-Learn

- You'll use scikit-learn to train each model on a subset of your data, then test each model on the rest to see how well your model works.
- Once your models are trained, you'll put them to work making predictions based on data they have not seen.
- Scikit-learn has tools that automate training and testing your models.
- Although you can specify parameters to customize the models and possibly improve their performance.

Which Scikit-Learn Estimator Should You Choose for Your Project

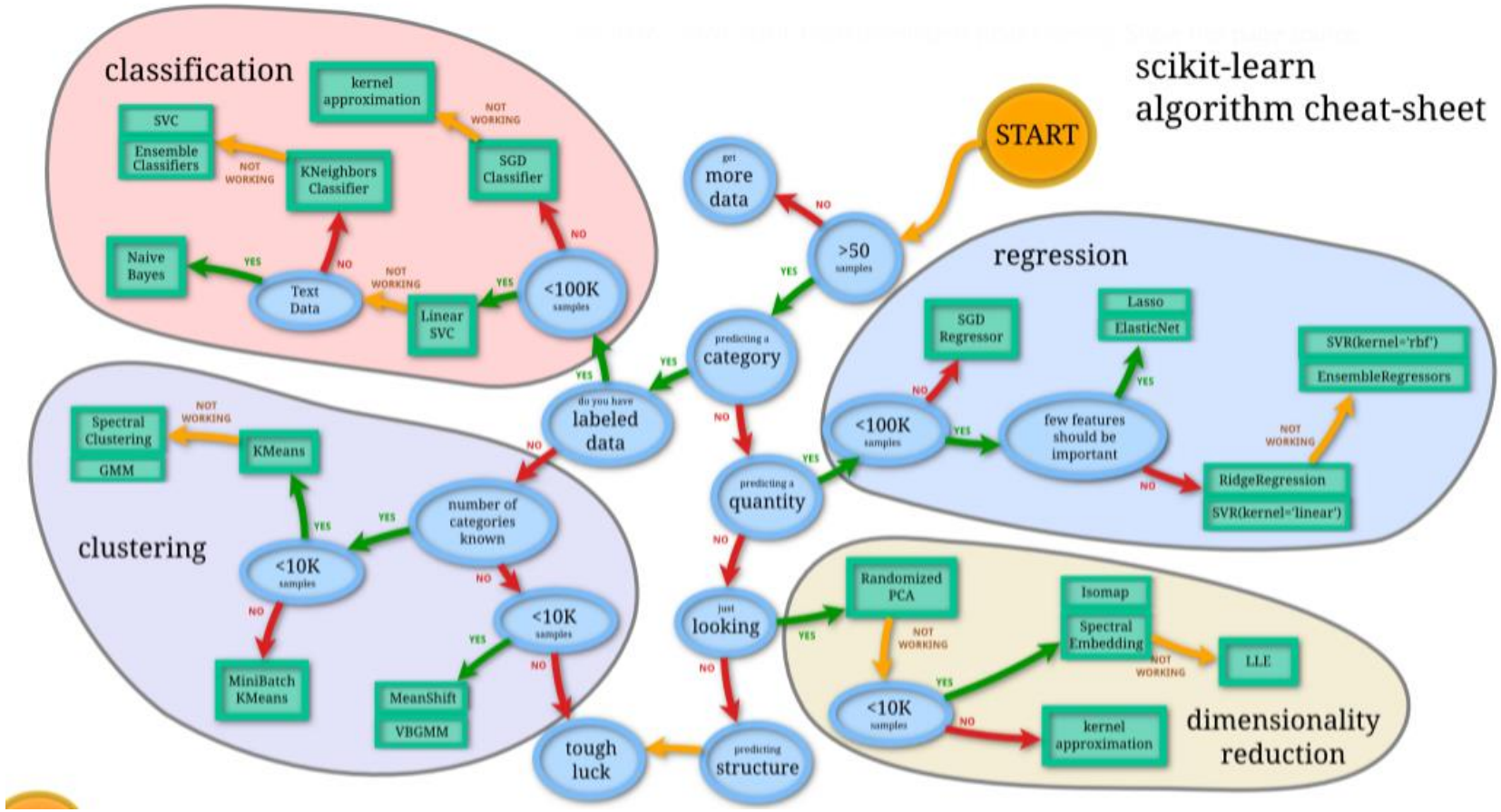
- It's difficult to know in advance which model(s) will perform best on your data, so you typically try many models and pick the one that performs best.
- A popular approach is to run many models and pick the best one(s).
- How do we evaluate which model performed best?
- You'll want to experiment with lots of different models on different kinds of datasets.

Which Scikit-Learn Estimator Should You Choose for Your Project

- You'll rarely get to know the details of the complex mathematical algorithms in the sklearn estimators, but with experience, you'll become familiar with which algorithms may be best for particular types of datasets and problems.
- Even with that experience, it's unlikely that you'll be able to intuit the best model for each new dataset.
- So scikit-learn makes it easy for you to "try 'em all."
- The models report their performance so you can compare the results and pick the model(s) with the best performance.

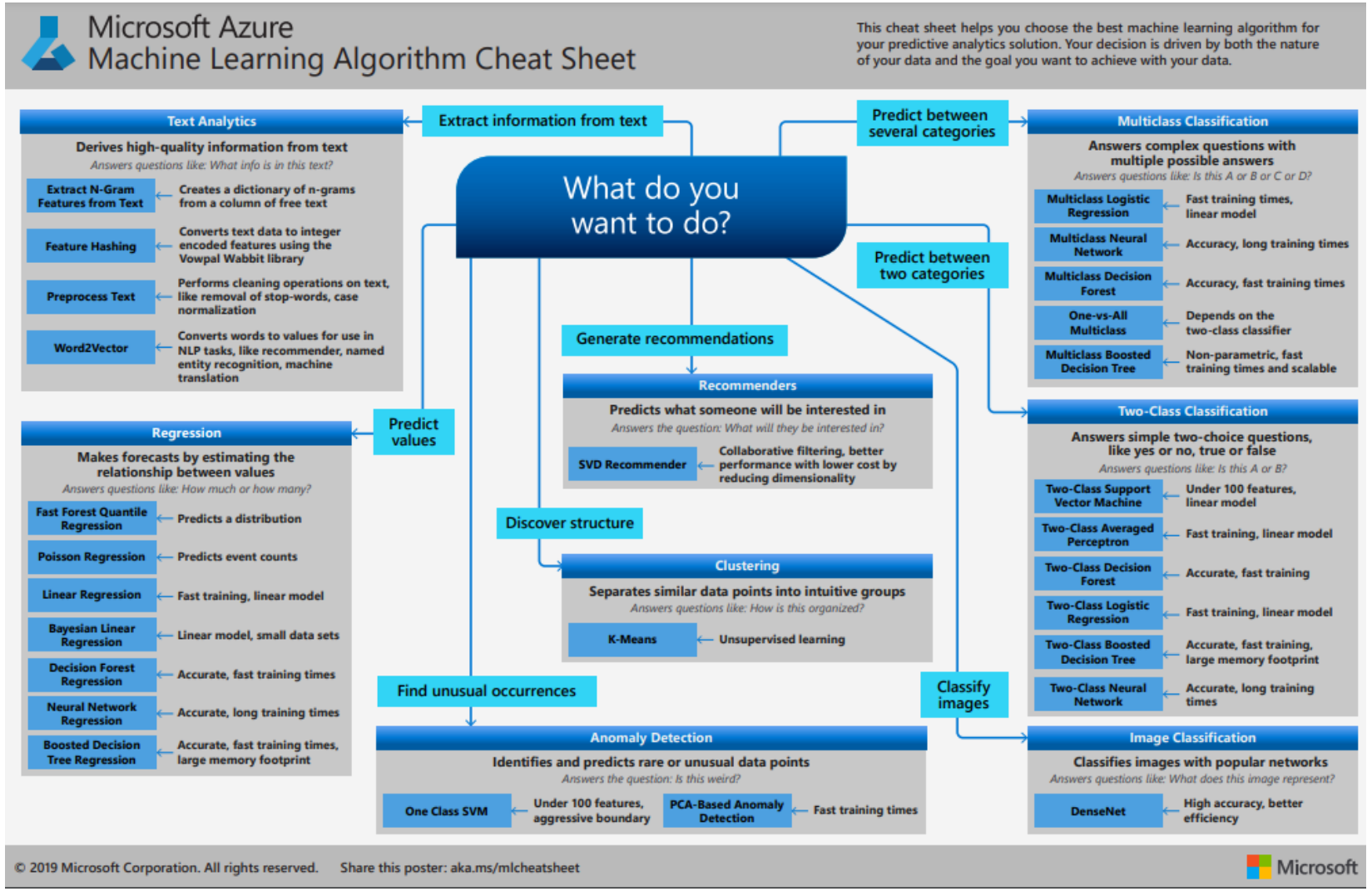
Scikit-Learn Estimator Diagram

- https://scikit-learn.org/stable/tutorial/machine_learning_map/



Azure ML Algorithm

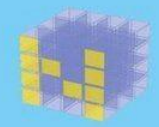
- [Link]



PYTHON FRAMEWORKS FOR DATA SCIENCE

NUMPY

It is a Python library that handles most of the numerical computing done using Python. It provides support for multi-dimensional arrays and matrices and comes with an impressive collection of routines to operate the arrays.



SCIPY

SciPy is a Python library that is commonly used in applications that call for scientific computing by scientists, engineers, and other technical fields. It has modules for signal processing, integration, solving ODEs, linear algebra, and more.

TENSORFLOW

TensorFlow is a platform that was created by the Google Brain Team with the sole purpose of making it easy for you to build Machine Learning (ML) models.

KERAS

Keras is a Python-based API that can run on TensorFlow, Theano, or CNTK. It is essentially a neural-network library that is designed to facilitate quick experiments on neural networks.

MATPLOTLIB

Matplotlib is mainly used for data visualization through plotting. Matplotlib is analogous to MATLAB in terms of application with the advantage of allowing you to program using Python which also means that it is open-source and free.

PANDAS

Pandas is a library that is used for data computation and analysis. It is extensively used for data wrangling which explains its popularity when any form of data analysis is involved.

Types of Machine Learning

- Supervised, which works with labeled data
- Unsupervised, which works with unlabeled data.
- computer vision application to recognize dogs and cats
- How can looking at unlabeled data be useful?
- Online booksellers → recommendation systems

Supervised Machine Learning

- Supervised machine learning falls into two categories;
  - Classification
  - Regression
- You train machine-learning models on datasets that consist of rows and columns.
- Each row represents a data sample.
- Each column represents a feature of that sample.
- In supervised machine learning, each sample has an associated label called a target (like “dog” or “cat”).
- This is the value you’re trying to predict for new data that you present to your models.

Datasets

- You'll work with some **"toy" datasets**, each with a small number of samples with a limited number of features.
- You'll also work with several richly featured **real-world datasets**, one containing a few thousand samples and one containing tens of thousands of samples.
- In the world of big data, datasets commonly have, millions and billions of samples, or even more.
- There's an enormous number of free and open datasets available for data science studies.

Datasets

- Libraries like scikit-learn package up popular datasets for you to experiment with and provide mechanisms for loading datasets from various repositories (such as openml.org).
  - http://archive.ics.uci.edu/ml/datasets.php
  - https://www.openml.org
  - https://www.kaggle.com/datasets
  - https://registry.opendata.aws
  - https://toolbox.google. com/datasetsearch
  - https://msropendata.com

Datasets

- Governments, businesses and other organizations worldwide offer datasets on a vast range of subjects.
  - https://data.tuik.gov.tr
  - https://data.ibb.gov.tr
  - https://data.gov

Classification

- one of the **simplest classification algorithms**, **k-nearest neighbors**, to analyze the **Digits dataset** bundled with **scikit-learn**.
- **Classification algorithms predict the discrete classes (categories) to which samples belong.**
- **Binary classification** uses two classes, such as “spam” or “not spam” in an email classification application.
- **Multi-classification** uses more than two classes, such as the 10 classes, 0 through 9, in the Digits dataset.
- A classification scheme looking at movie descriptions might try to classify them as “action,” “adventure,” “fantasy,” “romance,” “history” and the like.

Regression

- Regression models predict a continuous output, such as the predicted temperature output in the weather time series analysis.
- perform simple linear regression using scikit-learn's LinearRegression estimator.
- Next, use a LinearRegression estimator to perform multiple linear regression with the California Housing dataset that's bundled with scikit-learn.
- predict the median house value of a U. S. census block of homes, considering eight features per block, such as the average number of rooms, median house age, average number of bedrooms and median income.
- The LinearRegression estimator, by default, uses all the numerical features in a dataset to make more sophisticated predictions than you can with a single-feature simple linear regression

Unsupervised Machine Learning

- unsupervised machine learning with clustering algorithms
- dimensionality reduction (with scikit-learn's TSNE estimator) to compress the Digits dataset's 64 features down to two **for visualization purposes**.
- This will enable us to **see how** nicely the **Digits data "cluster up."**
- Digit dataset contains handwritten digits like those the post office's computers must recognize to route each letter to its designated zip code.
- This is a challenging computer-vision problem, given that each person's handwriting is unique.

Unsupervised Machine Learning

- Yet, we'll build this **clustering model** with just a **few lines of code** and achieve **impressive results**.
- And we'll do this **without having to understand the inner workings of the clustering algorithm**.
- This is the beauty of **object-based programming**.
- We'll see this kind of convenient object-based programming, when we'll build powerful **deep learning** models using the **open source Keras library**.

K-Means Clustering and the Iris Dataset

- simplest unsupervised machine-learning algorithm, k-means clustering
- dimensionality reduction (with scikit-learn's PCA estimator) to compress the Iris dataset's four features to two for visualization purposes.
- the **clustering of the three *Iris* species** in the dataset and graph each cluster's **centroid**, which is the **cluster's center point**.
- Finally, we'll **run multiple clustering estimators** to compare their ability to divide the Iris dataset's samples effectively into three clusters.

K-Means Clustering and the Iris Dataset

- You normally specify the desired number of clusters, k.
- K-means works through the data trying to divide it into that many clusters.
- As with many machine learning algorithms, **k-means** is **iterative** and gradually zeros in on the clusters to match the number you specify.
- K-means clustering can find similarities in unlabeled data.
- This can ultimately **help with assigning labels** to that data **so that supervised learning estimators can then process it**.
- Given that it's tedious and error-prone for humans to have to assign labels to unlabeled data, and given that the vast majority of the world's data is unlabeled, unsupervised machine learning is an important tool.

Big Data and Big Computer Processing Power

- The amount of data is already enormous and continues to grow exponentially.
- **The data produced in the world in the last few years equals the amount produced up to that point since the dawn of civilization.**
- We commonly talk about **big data**, but "big" may not be a strong enough term to describe truly how huge data is getting.
- People used to say "*I'm drowning in data and I don't know what to do with it.*"
- With machine learning, we now say, "*Flood me with big data so I can use machine-learning technology to extract insights and make predictions from it.*"

Big Data and Big Computer Processing Power

- This is occurring at a time when computing power is exploding and computer memory and secondary storage are exploding in capacity while costs dramatically decline.
- All of this enables us to think differently about the solution approaches.
- We now can program computers to **learn** from data, and lots of it.
- **It's now all about predicting from data.**

Datasets Bundled with Scikit-Learn

"Toy" datasets	*Real-world datasets*
Boston house prices	Olivetti faces
Iris plants	20 newsgroups text
Diabetes	Labeled Faces in the Wild face recognition
Optical recognition of handwritten digits	Forest cover types
Linnerrud	RCV1
Wine recognition	Kddcup 99
Breast cancer Wisconsin (diagnostic)	California Housing

Steps in a Typical Data Science Study

- loading the dataset
- exploring the data with pandas and visualizations
- transforming your data (converting non-numeric data to numeric data because scikit-learn requires numeric data; we'll discuss the issue again in the "Deep Learning" chapter)
- splitting the data for training and testing
- creating the model
- training and testing the model
- tuning the model and evaluating its accuracy
- making predictions on live data that the model hasn't seen before.
- These are important steps in cleaning your data before using it for machine learning.

Tavsiyeler!!!

- AlphaGo - The Movie | Full Documentary
- In the Age of AI (full film) | FRONTLINE → https://www.youtube.com/watch?v=5dZ_lvDgevk&t=4351s
- Machine Learning: Living in the Age of AI | A WIRED Film
- https://www.youtube.com/watch?v=ZJixNvx9BAc&t=299s
- How is Artificial Intelligence changing China → https://www.youtube.com/watch?v=cu731-8Bj60
- China - Surveillance state or way of the future? | DW Documentary
- https://www.youtube.com/watch?v=7gSU_Xes3GQ&t=1867s