



MAKİNE ÖĞRENMESİ VE SCIKIT-LEARN

Dr. Öğr. Üyesi Caner Erden, cerden@subu.edu.tr



@canererden



/c/CanerErdenn

İçerik

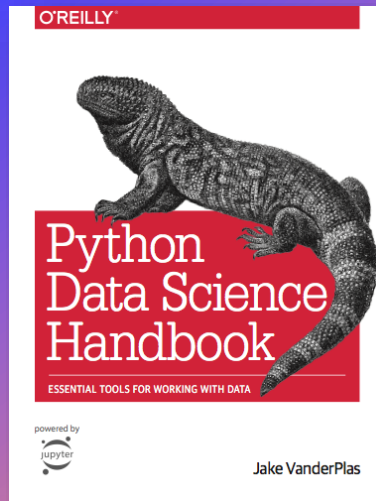
Makine
Öğrenmesi

Gözetimli ve Gözetimsiz Öğrenme
Eğitim ve Test Setleri
Performans Ölçütleri



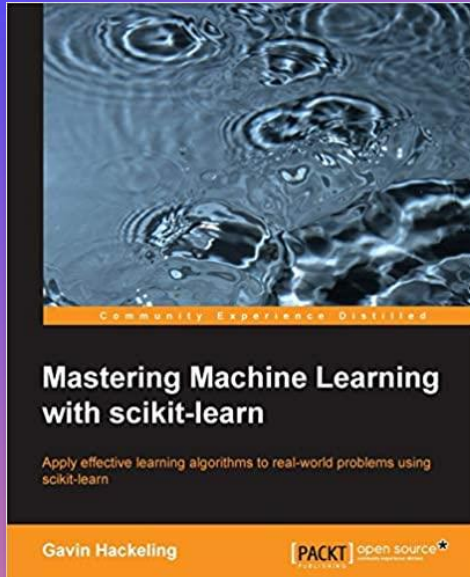
Scikit-Learn
Kütüphanesi

Genel bir bakış
Veri Setleri
Uygulamalar



VanderPlas, Jake. *Python Data Science Handbook: Tools and Techniques for Developers*. Sebastopol, CA, 2016.




Kaynaklar






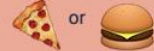
Hackling, Gavin. *Mastering Machine Learning with Scikit-Learn*. Birmingham: Packt Publishing, 2014.

MACHINE LEARNING IN EMOJI
















 SUPERVISED  UNSUPERVISED  REINFORCEMENT

 SUPERVISED human builds model based on input / output
 UNSUPERVISED human input, machine output
human utilizes if satisfactory
 REINFORCEMENT human input, machine output
human reward/punish, cycle continues

BASIC REGRESSION

 LINEAR `linear_model.LinearRegression()`
Lots of numerical data 
 LOGISTIC `linear_model.LogisticRegression()`
Target variable is categorical 

CLASSIFICATION

   NEURAL NET `neural_network.MLPClassifier()`
Complex relationships. Prone to overfitting
Basically magic. 
 K-NN `neighbors.KNeighborsClassifier()`
Group membership based on proximity 
 DECISION TREE `tree.DecisionTreeClassifier()`
If/then/else. Non-contiguous data
Can also be regression 
  RANDOM FOREST `ensemble.RandomForestClassifier()`
Find best split randomly
Can also be regression 
 SVM `svm.SVC()` `svm.LinearSVC()`
Maximum margin classifier. Fundamental
Data Science algorithm 
 NAIVE BAYES `GaussianNB()` `MultinomialNB()` `BernoulliNB()`
Updating knowledge step by step with new info 

CLUSTER ANALYSIS

 K-MEANS `cluster.KMeans()`
Similar datum into groups
based on centroids 
 ANOMALY DETECTION `covariance.EllipticalEnvelope()`
Finding outliers
through grouping 

FEATURE REDUCTION

T-DISTRIE STOCHASTIC NEIB EMBEDDING `manifold.TSNE()`
Visualize high dimensional data. Convert
similarity to joint probabilities 
PRINCIPLE COMPONENT ANALYSIS `decomposition.PCA()`
Distill feature space into components that
describe greatest variance 
CANONICAL CORRELATION ANALYSIS `decomposition.CCA()`
Making sense of cross-correlation
matrices 
LINEAR DISCRIMINANT ANALYSIS `lda.LDA()`
Linear combination of features that
separates classes 

OTHER IMPORTANT CONCEPTS

BIAS VARIANCE TRADEOFF 
UNDERFITTING / OVERFITTING 
INERTIA 
ACCURACY FUNCTION $(TP + TN) / (P + N)$ 
PRECISION FUNCTION $TP / (TP + FP)$ 
SPECIFICITY FUNCTION $TN / (FP + TN)$ 
SENSITIVITY FUNCTION $TP / (TP + FN)$

@emilyinamillion made this

Makine Öğrenmesi – Yapay Öğrenme

Gözetimli Öğrenme

- Etiketli
- Tahmin çalışmaları

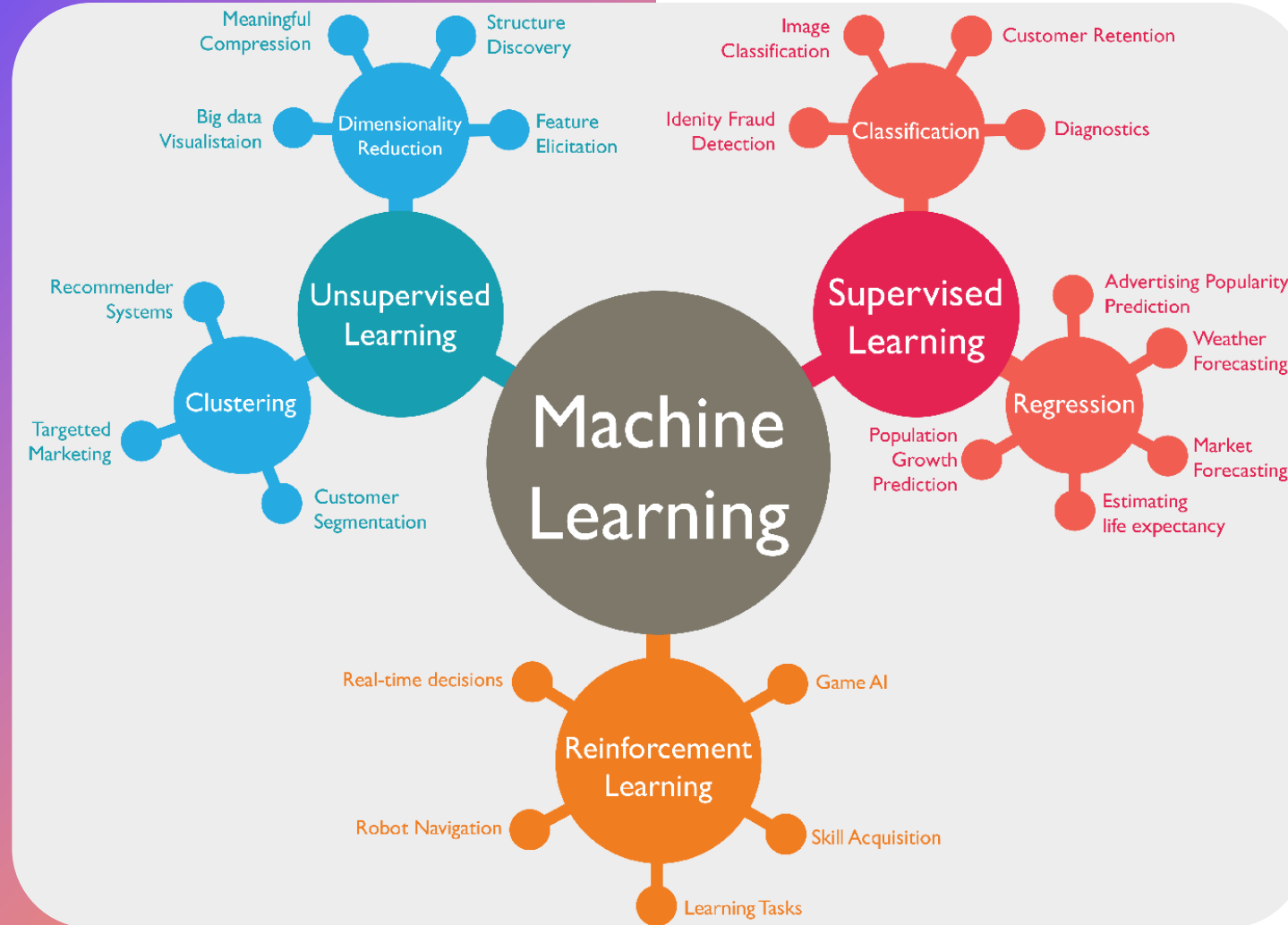
Gözetimsiz Öğrenme

- Etiket Yok
- Gizli örüntü keşfi

Takviyeli Öğrenme

- Ödül-ceza sistemi
- Deneme yanılma ile öğrenme

Uygulama Alanları



- <https://medium.com/@randylaosat/a-beginners-guide-to-machine-learning-dfadc19f6caf>

Sınıflandırma Örnekleri

Görev	Özellik Seti (x)	Sınıf etiketi (y)
Email mesajlarının kategorizasyonu	Email mesajından alınan metinler	SPAM ya da SPAM değil
Çiçek Türünün Belirlenmesi	Çiçeğin petal ve sepal uzunlukları ve genişlikleri	Şekillerine göre sınıflanan zambak çiçekleri
...



Canlı türleri sınıflandırması

Name	Body Temperature	Skin Cover	Gives Birth	Aquatic Creature	Aerial Creature	Has legs	Hibernates	Class Label
Human	warm-blooded	hair	yes	no	no	yes	no	mammal
python	cold-blooded	scales	no	no	no	no	yes	reptile
salmon	cold-blooded	scales	no	yes	no	no	no	fish
whale	warm-blooded	hair	yes	yes	no	no	no	mammal
frog	cold-blooded	none	no	semi	no	yes	yes	amphibian
komodo dragon	cold-blooded	scales	no	no	no	yes	no	reptile
bat	warm-blooded	hair	yes	no	yes	yes	yes	mammal
pigeon	warm-blooded	feathers	no	no	yes	yes	no	bird
cat	warm-blooded	fur	yes	no	no	yes	no	mammal
leopard shark	cold-blooded	scales	yes	yes	no	no	no	fish
turtle	cold-blooded	scales	no	semi	no	yes	no	reptile
penguin	warm-blooded	feathers	no	semi	no	yes	no	bird
porcupine	warm-blooded	quills	yes	no	no	yes	yes	mammal
eel	cold-blooded	scales	no	yes	no	no	no	fish
salamander	cold-blooded	none	no	semi	no	yes	yes	amphibian

Sınıflandırma Modeli

Eğitim seti:
Algoritmanın
öğrenme
gerçekleştireceği
veri seti

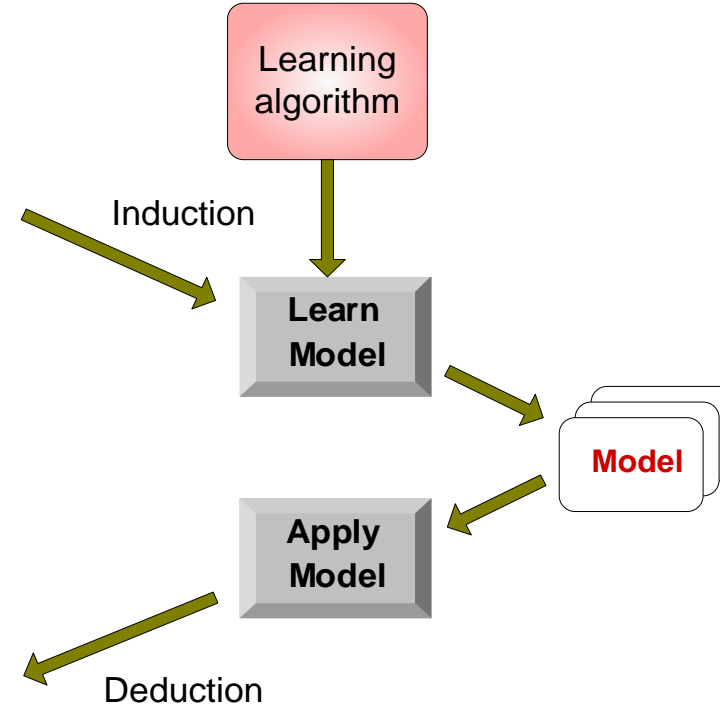
Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

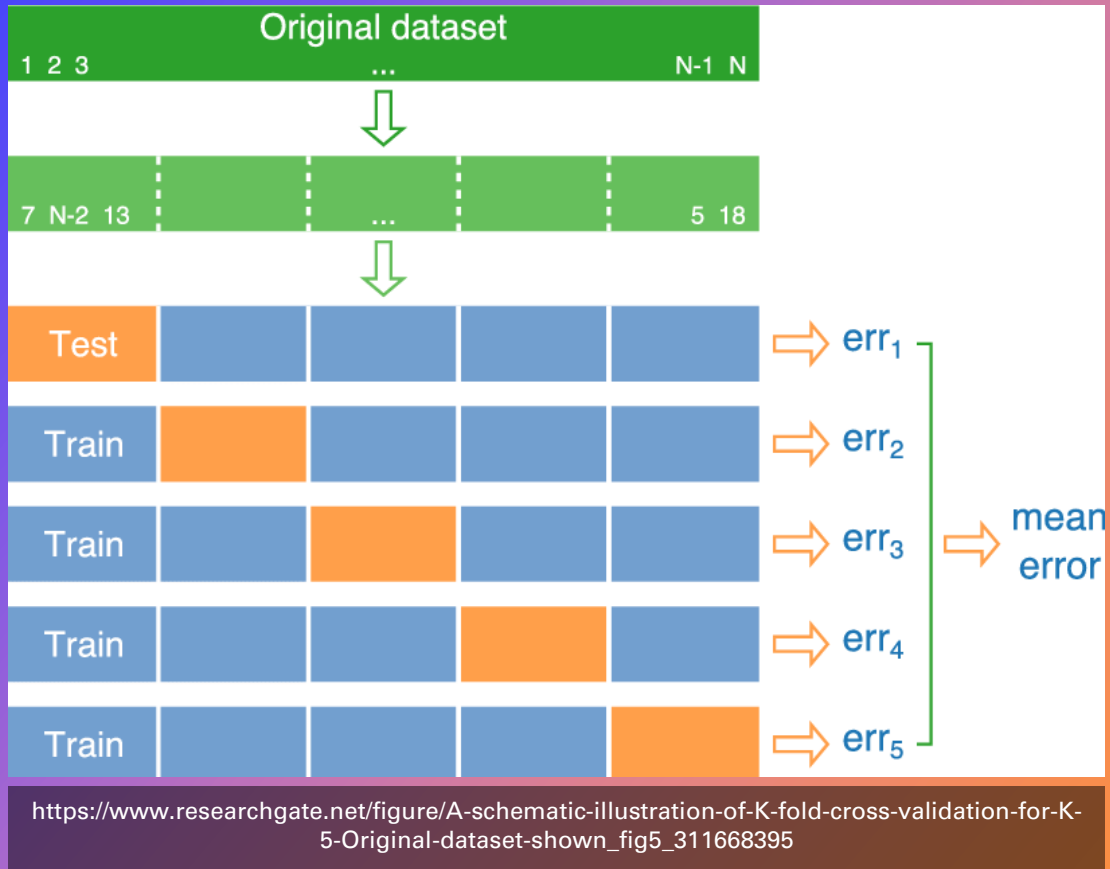
Training Set

Test seti:
Algoritmaların
performansının test
edildiği veri seti

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set





Çapraz Doğrulama (k-folds cross validation)

- Algoritmanın performansının seçilen eğitim ve test setinden bağımsız olarak ölçülebilmesi için geliştirilmiştir.
- Yeni verilerin tahmin becerisi hakkında izlenim verir.
- Veri seti belirli sayıda(k-folds) parçaya bölünür.
- Farklı veri setleri için en uygun k değeri için yöntemler vardır.

Performans Ölçüleri

		True Class	
		Positive	Negative
Predicted Class	Positive	TP	FP
	Negative	FN	TN

Doğruluk skoru (Accuracy): Toplam doğruların sayısı / toplam tahminlerin sayısı

Kesinlik Skoru(Precision): Pozitif doğru tahminlerin sayısı / toplam pozitiflerin sayısı

Duyarlılık (Recall): Pozitif doğru tahminlerin sayısı / Pozitif doğru + negatif yanlış

F1 Skoru: Kesinlik ve duyarlılık skorlarının harmonik ortalaması

$$F1 = 2 \times \frac{\text{kesinlik} \times \text{duyarlılık}}{\text{kesinlik} + \text{duyarlılık}}$$

Regresyonda Performans Ölçüleri

Ortalama Mutlak Hata(Mean Absolute Error):
Tahminlerin gerçek değerlerden mutlak farkının ortalamasıdır.

$$MAE = \frac{1}{n} \sum |Y_i - \hat{Y}_i|$$

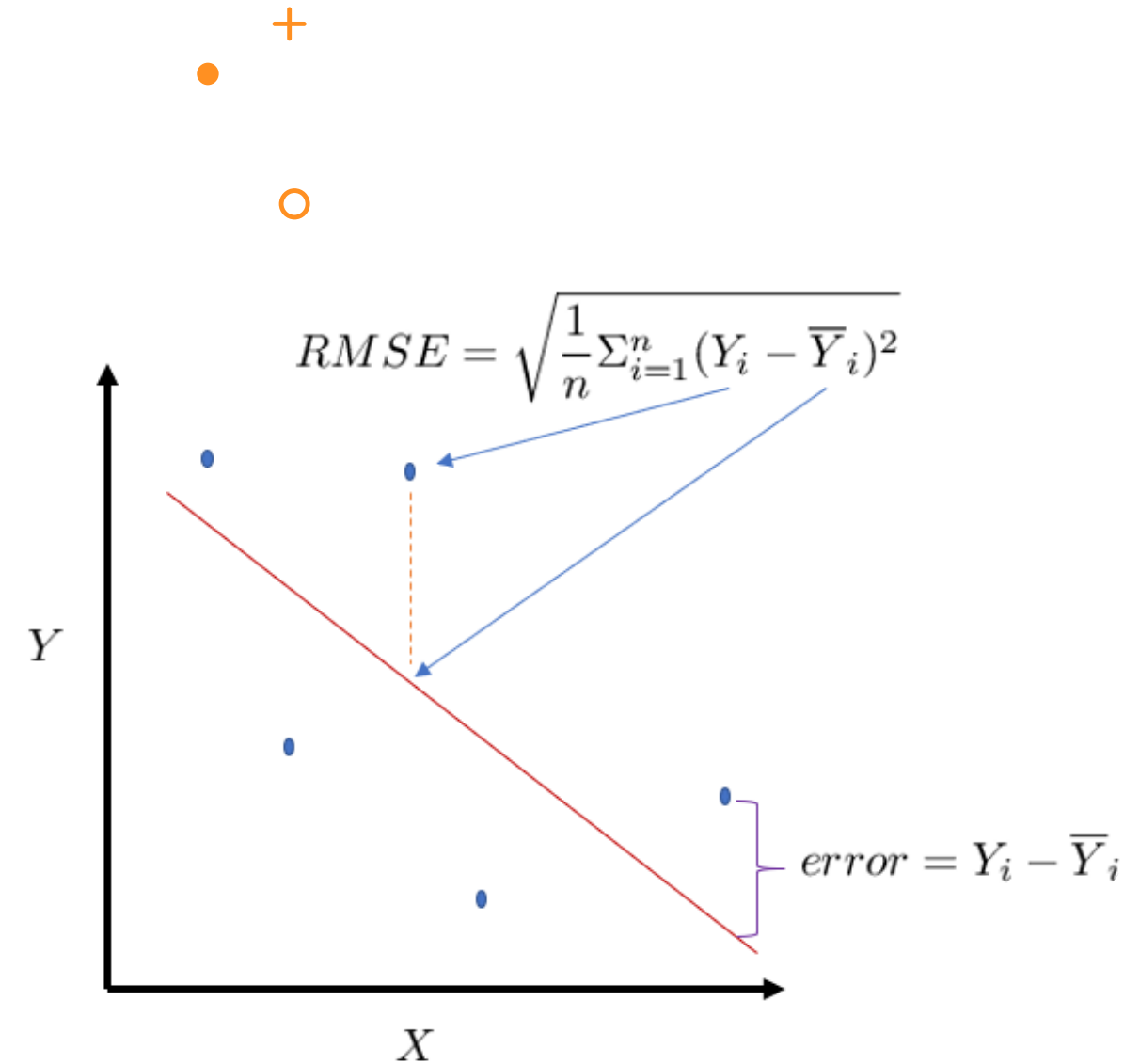
Ortalama Karesel Hata (Mean Squared Error)

$$MSE = \frac{1}{n} \sum (Y_i - \hat{Y}_i)^2$$

Ortalama Karesel Hataların Karekökü (Root Mean Squared Error)

$$RMSE = \sqrt{\frac{1}{n} \sum (Y_i - \hat{Y}_i)^2}$$

R² Skoru (Determinasyon Katsayısı): En iyi 1 olabilir.





Scikits

Veri bilimi kütüphaneleri

- Veri Analizi(Pandas + Numpy + Scikit-Learn)

Veri ön işleme (Pandas + Scikit-Learn)

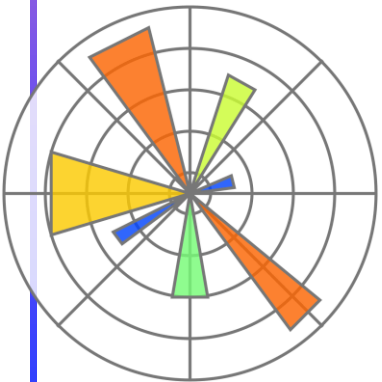
- Veri temizleme/ eksik veriler / Aykırı veriler
- Normalleştirme/ Standardizasyon

Eğitim ve Test Seti

Geleneksel makine

öğrenmesi algoritmaları

<https://scikit-learn.org/stable/>



Scikit-Learn Kütüphanesi



2007 yılında geliştirilmiştir.



Tahmine dayalı veri analizi için basit ve verimli araçlar



Herkes tarafından erişilebilir ve çeşitli bağlamlarda yeniden kullanılabilir.



NumPy, SciPy ve Matplotlib üzerine inşa edilmiştir.



Açık kaynak kodlu, ticari olarak kullanılabilir - BSD lisansına sahip.

Uygulama Alanı

Sınıflandırma

Regresyon

Kümeleme

Boyut Azaltma

Model Seçme

Veri Ön işleme