



Scikit-Learn ile Metin Sınıflandırma

IOI

Ben

- Neslihan Şirin Saygılı
- fellow developer @ Prisync
-  /sirinnes
-  /sirin

Neden Scikit-Learn?

- NLTK → NumPy → sk-learn
- Veri analizi/madenciliği için basit ve yetenekli bir araç
- Birçok algoritmanın implemente edilmiş hali var

Neden metinle uğraşıyorum?

- Ham madde çok fazla ve sürekli artıyor.

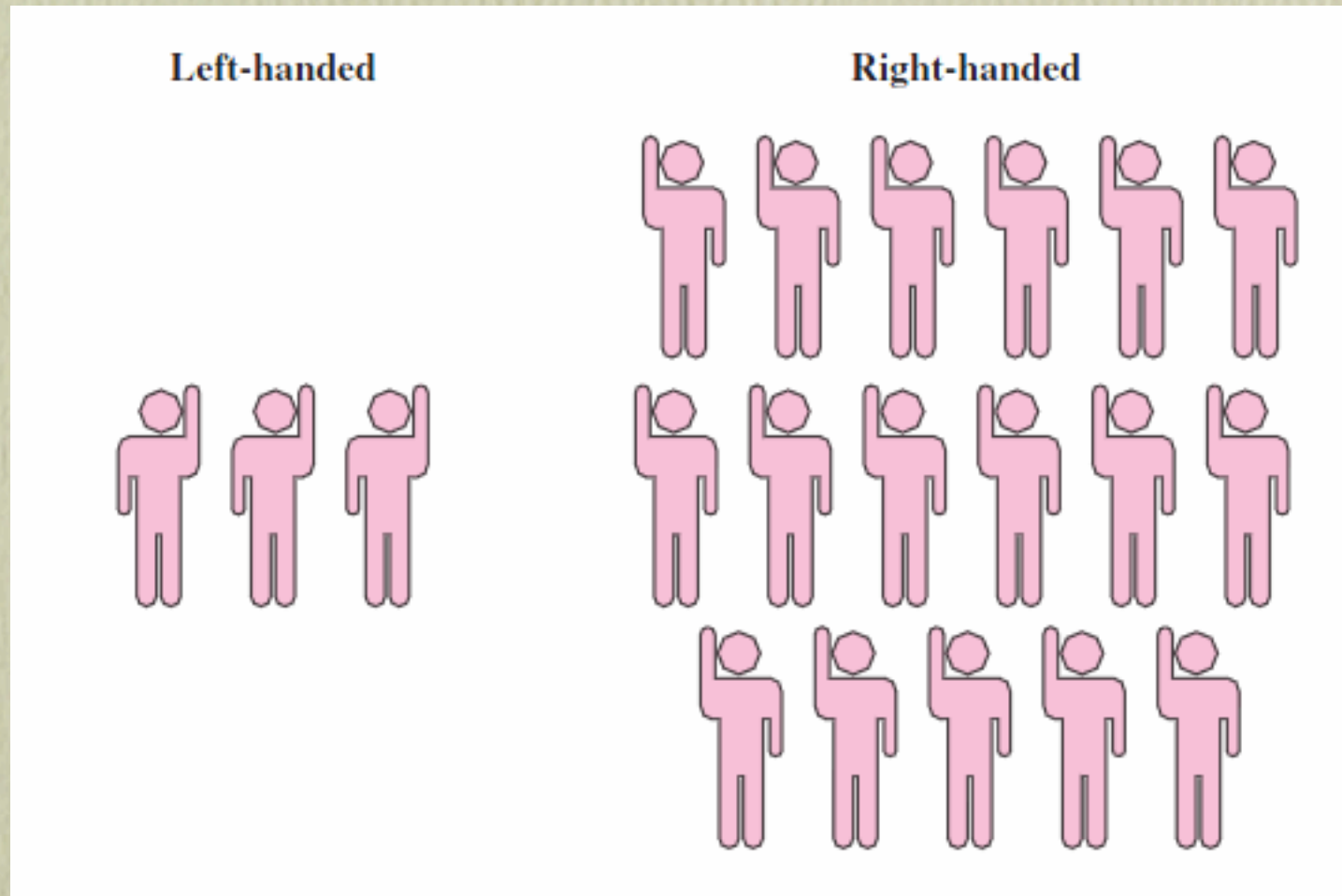


Problemimiz ne?

- Temel öğrenme problemleri şu yolu izler;
- Büyük bir veriyi ele alıp,
- Bu verinin bir parçasını kullanarak,
- Geri kalan verilerin özelliklerini bulmaya çalışır.

Peki Sınıflandırma nedir?

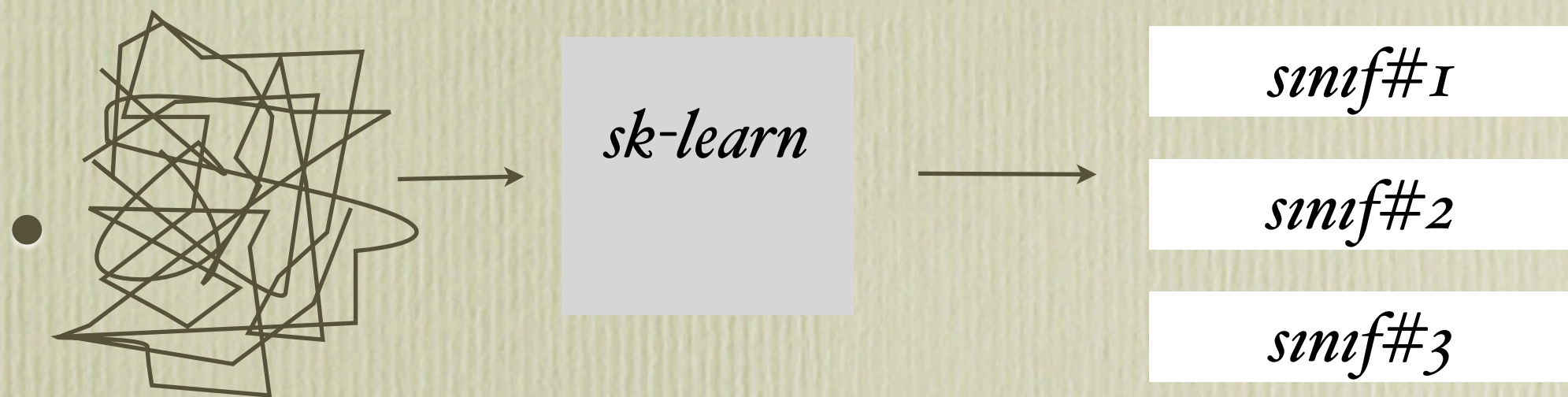
- İsmiyle müsemma, **sınıflandırma**,



Nasıl sınıflandırıyoruz?

- Veriyi al, **train** ve **test** olarak böl
- Numuneleri (train data) **sınıf#1**, **sınıf#2**, ..., **sınıf#n** olarak **etiketle**
- Algoritmayı **train data** ile **eğit**
- Algoritmaya **test** datasını ver

Sk-learn nasıl yapıyor?



sk-learn ile metin sınıflandırma adımları

- Veriyi train ve test olarak bölüp etiketleyen fonksiyonları kullan
- Veriyi sk-learn'ün kullanabileceği biçime getiren fonksiyonları kullan (**data to vector**)
- Algoritmayı train datasının vektör haliyle eğit
- Algoritmayı test datasının vektörüyle dene

Data to vector

- İşlemin adı: **Feature extraction** (veriyi machine learning algoritmasının anlayacağı dile çevirmek)
- Malzemeler: **data, vectorizer**
- Ürün: machine learning algoritmasının kullanmak için can attığı seksi bir **vektör**

Vectorizer

- TfidfVectorizer (metin işlemede çok kullanılır)
- tf-idf: bir kelimenin örnek dokümanlar içerisinde ne kadar önemli olduğunu anlamak için kullanılan bir ölçüdür
- sparse matrix: büyük çoğunluğu sıfırdan oluşan matrix

Support Vector Machines

- Çevirisi dümdüz : Destek Vektör Makineleri



•

Sonucu nasıl anlıyoruz?

- Sayarak. A ve B sınıfları olsun.
- Bütün adaylardan “bu A sınıfındadır” dediklerimin sayısı
- Bence A sınıfında olanlardan “gerçekten A sınıfında olması gerekenlerin” sayısı
- Bu iki sayıyı $\{*\text{topla}/\text{çıkar}/\text{çarp}/\text{böl}\} =$
- F1-Score (ideal olan sonuç 1)

	<i>FI Score</i>	<i>Doc. Count</i>
alt.atheism	0.85	319
comp.graphics	0.95	389
sci.space	0.95	394
talk.religion.misc	0.81	251
avg / total	0.90	1353

- Demo var.
- Teşekkür ederim.
- Soru ?