CIE 5133 機器學習與深度學習導論
線上課程

開始之前 (10.06.2021)

- 請將你的麥克風靜音
- 請找個安全、舒適的空間
- 聽講時有任何問題請到 slido #073374 留言
- 我們會透過 Zoom, slido, 討論區，臉書社群來強化無法面對面所造成的互動不足，同學們有任何建議也請讓我們知道。
- HW1 was issued last week and TAs are ready to help!
- Stay Home! Happy Learning!!

CIE 5133 機器學習與深度學習導論

Course FaceBook

Today …

- Fundamentals and Landscape of Classical Machine Learning (II) and (III)

Review

Summary

Learning? What do we mean?

Is learning feasible?

- Machine learning: use data to compute hypothesis g that approximate unknown target f. → function
- In practice, learning algorithm $\mathcal{A}$ takes training examples $\mathcal{D}$ and hypothesis set $\mathcal{H}$ to get final hypothesis g. → new data
- Learning is only feasible in a *probabilistic* way and we can predict something useful outside the training set $\mathcal{D}$ using only $\mathcal{D}$.

Fundamentals and Landscape of Classical Machine Learning (II)

- Learn the framing of supervised learning
- Know the modern machine learning landscape
- Learn the basics of Scikit-Learn

https://www.sli.do/
#073374

Learn the framing of supervised learning

Supervised Machine Learning

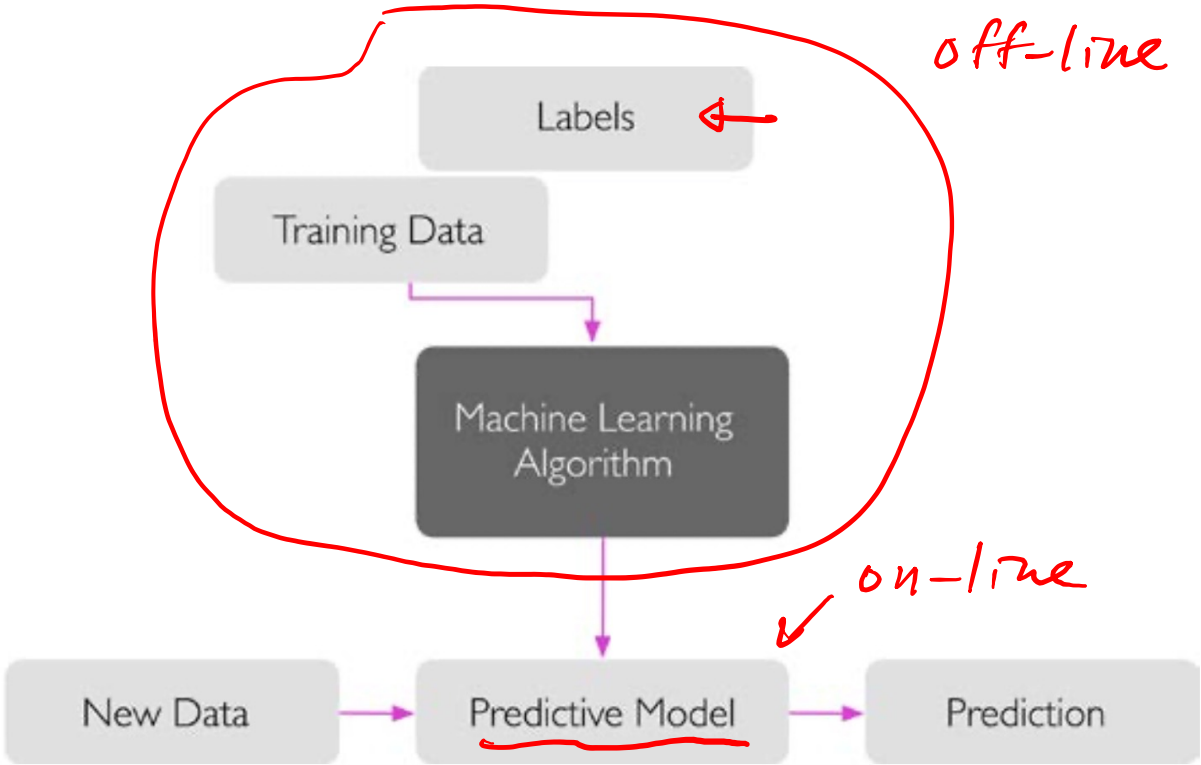
credit: Google, machine learning crash course,
https://developers.google.com/machine-learning/crash-course

ML models learn

how to combine input

to produce useful predictions

on never-before-seen data



Supervised Machine Learning

credit: Google, machine learning crash course,
https://developers.google.com/machine-learning/crash-course

ML models learn

how to combine input

to produce useful predictions

on never-before-seen data

https://www.tesla.com/autopilotAI

(The deep neural) networks learn from the most complicated and diverse scenarios in the world, iteratively sourced from our fleet of nearly 1M vehicles in real time. A full build of Autopilot neural networks involves 48 networks that take 70,000 GPU hours to train 🔥. Together, they **output 1,000 distinct tensors (predictions)** at each timestep.

Supervised Machine Learning Terminology

- **Label** is the variable we're predicting
  - Typically represented by the variable y
- **Features** are input variables describing our data
  - Typically represented by the variables $\{x_1, x_2, \ldots, x_D\}$
- **Example** is a particular instance of data, $\boldsymbol{x}$ (**bold** indicates a vector)
- **Labeled example** has {features, label}: $\{\boldsymbol{x}, y\}$
  - Used to train the model
- **Unlabeled example** has {features, ?}
  - Used to making prediction on new data
- **Model** maps unlabeled examples to predicted labels: y'
  - Defined by (training) parameters, which are learned.

from training examples

credit: Google, machine learning crash course,
https://developers.google.com/machine-learning/crash-course

Supervised Machine Learning (Credit Approval)

Labeled examples

age (feature)	gender (feature)	annual salary (feature)	year in residence (feature)	year in job (feature)	current debt (feature)	approval (label)
23	female	1,000,000	1	0.5	200,000	Yes
45	male	500,000	1	0.5	250,000	No
75	male	0	20	0	0	Yes

Unlabeled examples

age (feature)	gender (feature)	annual salary (feature)	year in residence (feature)	year in job (feature)	current debt (feature)
45	female	1,500,000	10	5	500,000

Fun Time: Supervised Machine Learning

https://www.sli.do/
#073374

Suppose you want to develop a supervised machine learning model to predict whether a given email is “spam” or “not spam.” Which of the following statements are true? (多選題)

(i) Emails not marked as "spam" or "not spam" are unlabeled examples.

(ii) The labels applied to some examples might be unreliable.

(iii) We'll use unlabeled examples to train the model.

(iv) Words in the subject header will make good labels.

credit: Google, machine learning crash course,
https://developers.google.com/machine-learning/crash-course

Suppose you want to develop a supervised machine learning model to predict whether a given email is "spam" or "not spam." Which of the following statements are true?

Emails not marked as "spam" or "not spam" are unlabeled examples. ✓

Because our label consists of the values "spam" and "not spam", any email not yet marked as spam or not spam is an unlabeled example.

正確答案共有 2 個，你目前選中了 2 個。

The labels applied to some examples might be unreliable. ✓

Definitely. It's important to check how reliable your data is. The labels for this dataset probably come from email users who mark particular email messages as spam. Since most users do not mark every suspicious email message as spam, we may have trouble knowing whether an email is spam. Furthermore, spammers could intentionally poison our model by providing faulty labels.

正確答案共有 2 個，你目前選中了 1 個。

We'll use unlabeled examples to train the model. ⊘

We'll use **labeled** examples to train the model. We can then run the trained model against unlabeled examples to infer whether the unlabeled email messages are spam or not spam.

請再試一次。

Words in the subject header will make good labels. ⊘

Words in the subject header might make excellent features, but they won't make good labels.

請再試一次。

credit: Google, machine learning crash course, https://developers.google.com/machine-learning/crash-course

Fun Time: Features and Labels

https://www.sli.do/
#073374

Suppose an online shoe store wants to create a supervised ML model that will provide personalized shoe recommendations to users. That is, the model will recommend certain pairs of shoes to Marty and different pairs of shoes to Janet. The system will use past user behavior data to generate training data. Which of the following statements are true? (多選題)

(i) "Shoes that a user adores" is a useful label.

(ii) "Shoe beauty" is a useful feature.

(iii) "The user clicked on the shoe's description" is a useful label.

(iv) "Shoe size" is a useful feature.

credit: Google, machine learning crash course,
https://developers.google.com/machine-learning/crash-course

Features and Labels

Explore the options below.

Suppose an online shoe store wants to create a supervised ML model that will provide personalized shoe recommendations to users. That is, the model will recommend certain pairs of shoes to Marty and different pairs of shoes to Janet. The system will use past user behavior data to generate training data. Which of the following statements are true?

"Shoes that a user adores" is a useful label. ⊘

Adoration is not an observable, quantifiable metric. The best we can do is search for observable proxy metrics for adoration.

請再試一次。

"Shoe beauty" is a useful feature. ⊘

Good features are concrete and quantifiable. Beauty is too vague a concept to serve as a useful feature. Beauty is probably a blend of certain concrete features, such as style and color. Style and color would each be better features than beauty.

請再試一次。

"The user clicked on the shoe's description" is a useful label. ✓

Users probably only want to read more about those shoes that they like. Clicks by users is, therefore, an observable, quantifiable metric that could serve as a good training label. Since our training data derives from past user behavior, our labels need to derive from objective behaviors like clicks that strongly correlate with user preferences.

正確答案共有 2 個，你目前選中了 1 個。

"Shoe size" is a useful feature. ✓

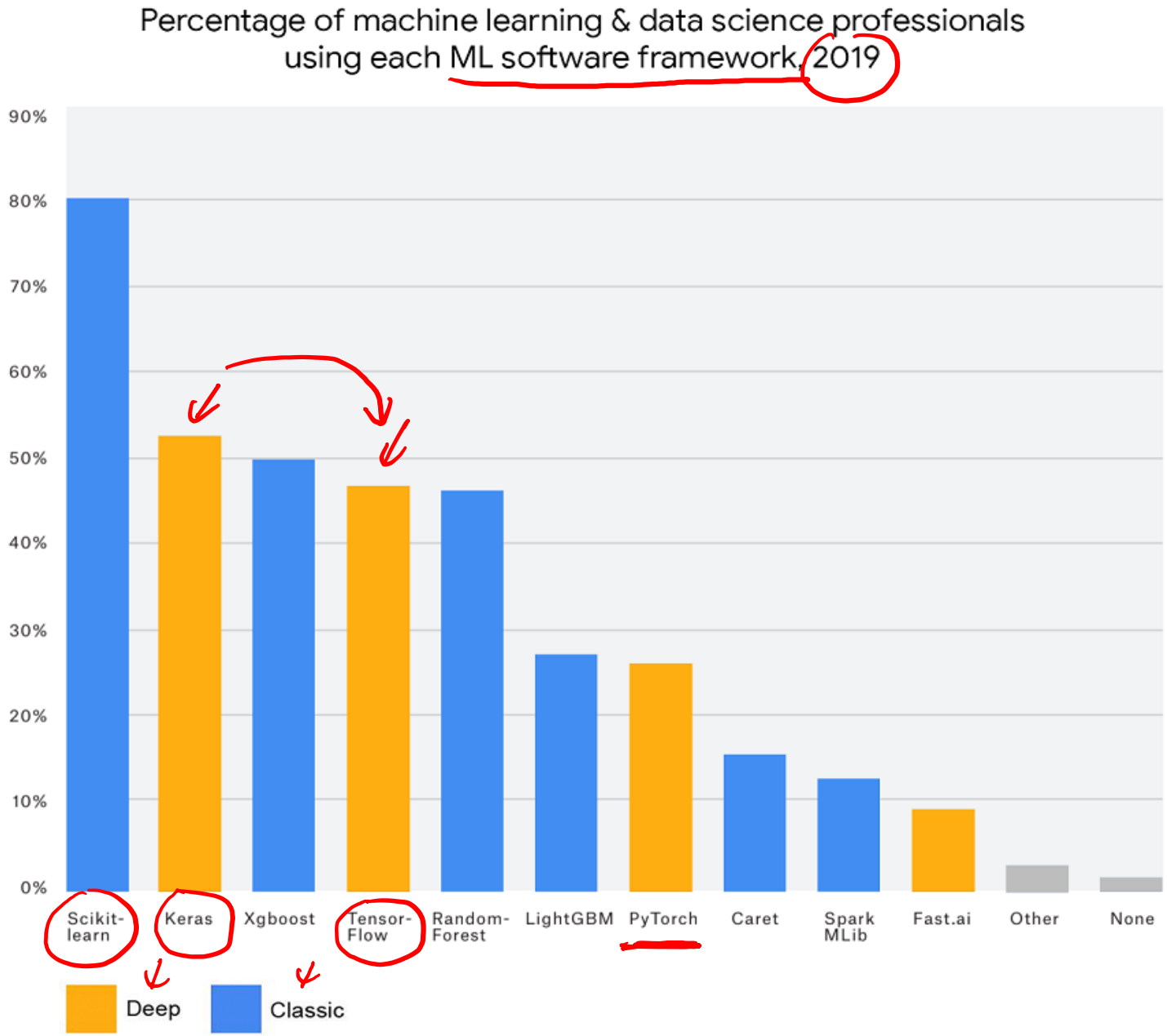
"Shoe size" is a quantifiable signal that likely has a strong impact on whether the user will like the recommended shoes. For example, if Marty wears size 9, the model shouldn't recommend size 7 shoes.

正確答案共有 2 個，你目前選中了 2 個。

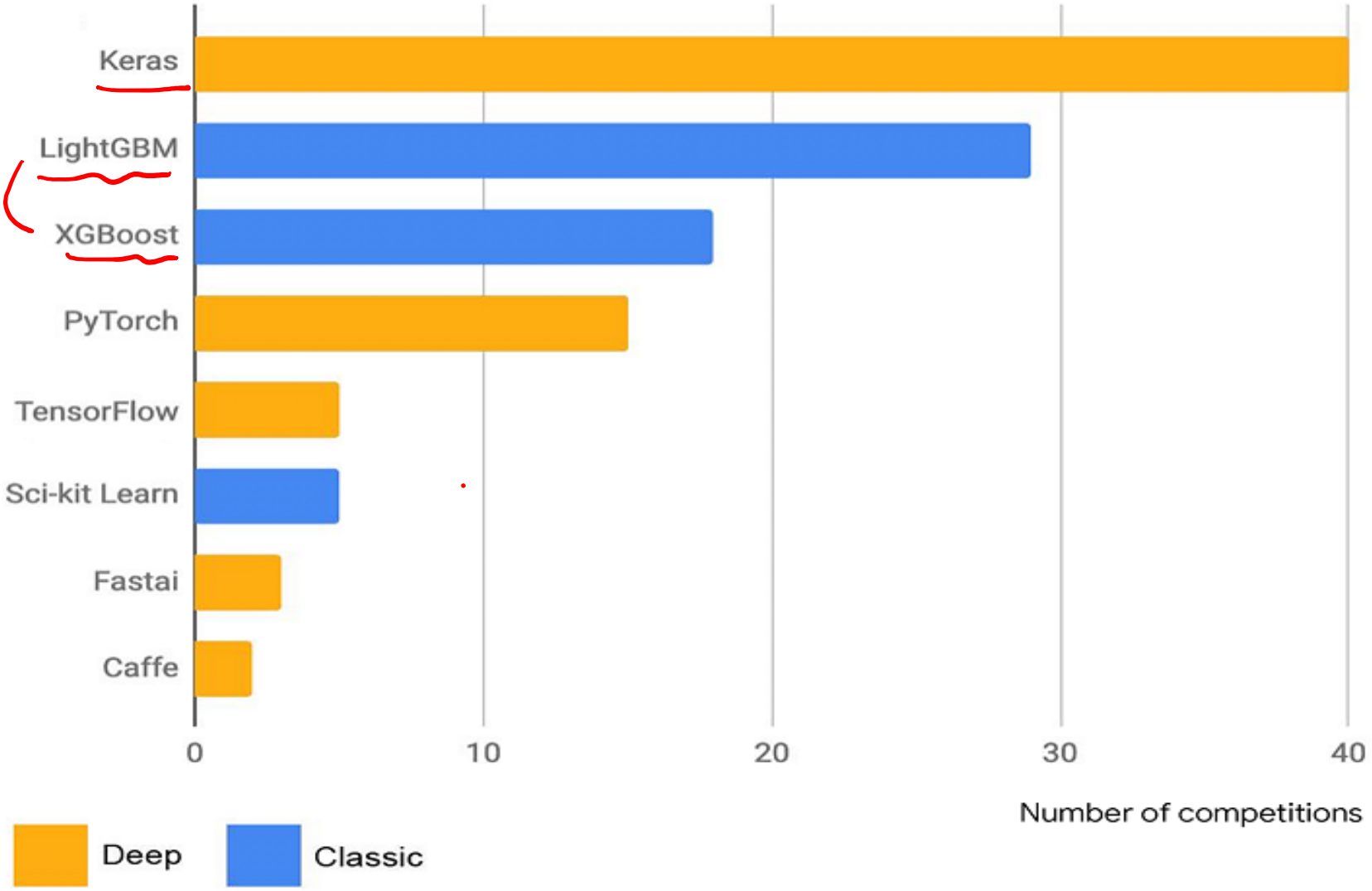
credit: Google, machine learning crash course, https://developers.google.com/machine-learning/crash-course

Know the modern machine learning landscape

C-S David Chen, Department of Civil Engineering, National Taiwan University


Percentage of machine learning & data science professionals using each ML software framework, 2019
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Scikit-learn
Keras
Xgboost
Tensor-Flow
Random-Forest
LightGBM
PyTorch
Caret
Spark MLib
Fast.ai
Other
None
Deep
Classic

Primary ML tool used by top-5 teams in Kaggle competitions, 2017-2018 (N=120)



Summary

Know the modern ML landscape

- Scikit-Learn and Keras (now part of TensorFlow) are mostly widely used ML software frameworks by ML professionals.

- From 2016 to 2020, the entire machine learning and data science industry has been dominated by these two approaches: deep learning and gradient boosted trees. Specifically, gradient boosted trees is used for problems where structured data is available, whereas deep learning is used for perceptual problems such as image classification.

- Users of gradient boosted trees tend to use Scikit-Learn, XGBoost or LightGBM. Meanwhile, most practitioners of deep learning use Keras, often in combination with its parent framework TensorFlow.

- The common point of these tools is they're all Python libraries: Python has is by far the most widely-used language for machine learning and data science.

Learn the basics of Scikit-Learn

C-S David Chen, Department of Civil Engineering, National Taiwan University

Machine Learning with Scikit-Learn

Extensions to **SciPy** (Scientific Python) are called **SciKits**. **SciKit-Learn** provides machine learning algorithms.

- **Algorithms for supervised & unsupervised learning**
- **Built on SciPy and Numpy**
- **Standard Python API interface**
- **Probably the best general ML framework out there.**

credit: Introduction to Machine Learning with Scikit-Learn, District Data Lab

Machine Learning with Scikit-Learn

Primary Features

- Generalized Linear Models
- SVMs, kNN, Bayes, Decision Trees, Ensembles
- Clustering and Density algorithms
- Cross Validation
- Grid Search
- Pipelining
- Model Evaluations
- Dataset Transformation
- Dataset Loading

credit: Introduction to Machine Learning with Scikit-Learn, District Data Lab

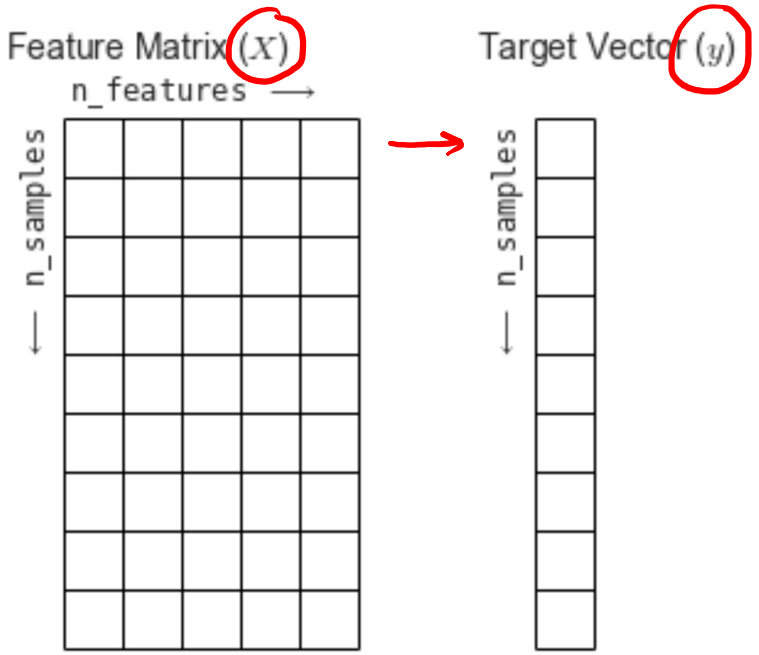
Machine Learning with Scikit-Learn

Object-oriented interface centered around the concept of an Estimator:

"**An estimator is any object that learns from data**; it may be a classification, regression or clustering algorithm or a transformer that extracts/filters useful features from raw data."

credit: Introduction to Machine Learning with Scikit-Learn, District Data Lab

Machine Learning with Scikit-Learn



Estimators

- fit(X,y) sets the state of the estimator.
  - X is usually a 2D numpy array of shape (num_samples, num_features)
  - y is a 1D array with shape (n_samples,)
- predict(X) returns the class or value

See Introducing_Scikit-Learn.pdf

credit: Introduction to Machine Learning with Scikit-Learn, District Data Lab