Module 3

Alireza Samar

UTM UNIVERSITI TEKNOLOGI MALAYSIA

MLDS MACHINE LEARNING FOR DATA SCIENCE

Who Am I?

Alireza Samar @alirezasmr

Founder, CTO at Appoint AI

Graduate Research Assistant at UTM MLDS

Curator at Machine Learning Weekly - mlweekly.com

AppointAI

UTM MLDS MACHINE LEARNING FOR DATA SCIENCE

ML

MLDS

What We've Learned in Previous Modules

- Anaconda Package Manager
- Install and run code on Jupyter
- Version Controlling Concept
- Git
- GitHub and GitHub Desktop

MLDS

What We've Learned in Previous Modules

- Statistics
- Linear Algebra
- Optimization
- Bayes Rule
- Maximum Likelihood, Gradient Descent, ...

MLDS

What We've Learned in Previous Modules

- Python Basics (Syntax, Arrays, Loops, Functions and etc)
- Numpy
- Matplotlib
- Exploratory Data Analysis
- SciPy - Scientific Computing Toolkit

MLDS

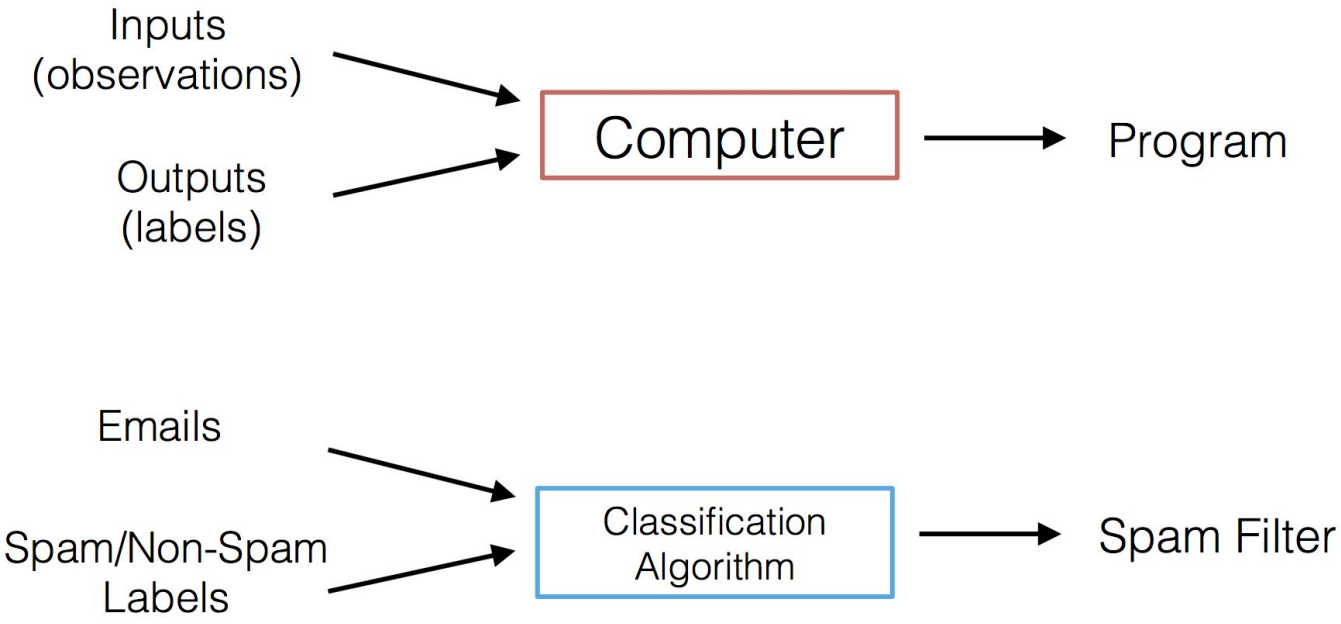
Workshop Materials

GitHub: **UTMMLDS**

https://github.com/utmmlds

MLDS

What is Machine Learning?



MLDS

3 Types of Learning

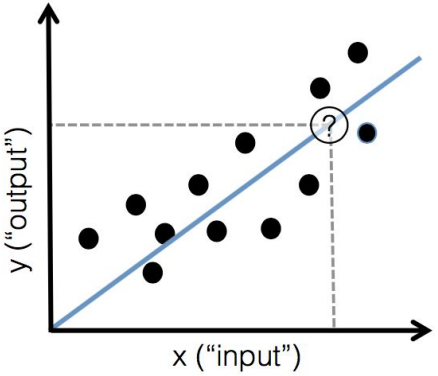
- Supervised
- Unsupervised
- Reinforcement

MLDS

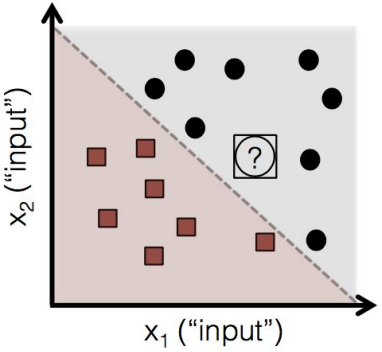
Working with Labeled Data

Supervised Learning

Regression



Classification

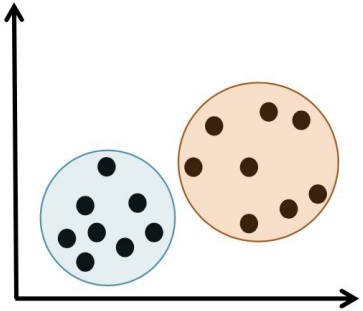


MLDS

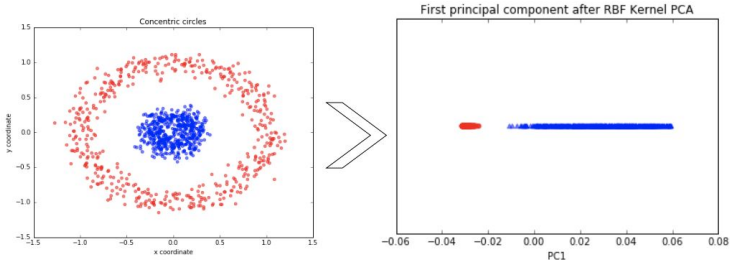
Working with **Un**labeled Data

Unsupervised Learning

Clustering



Compression



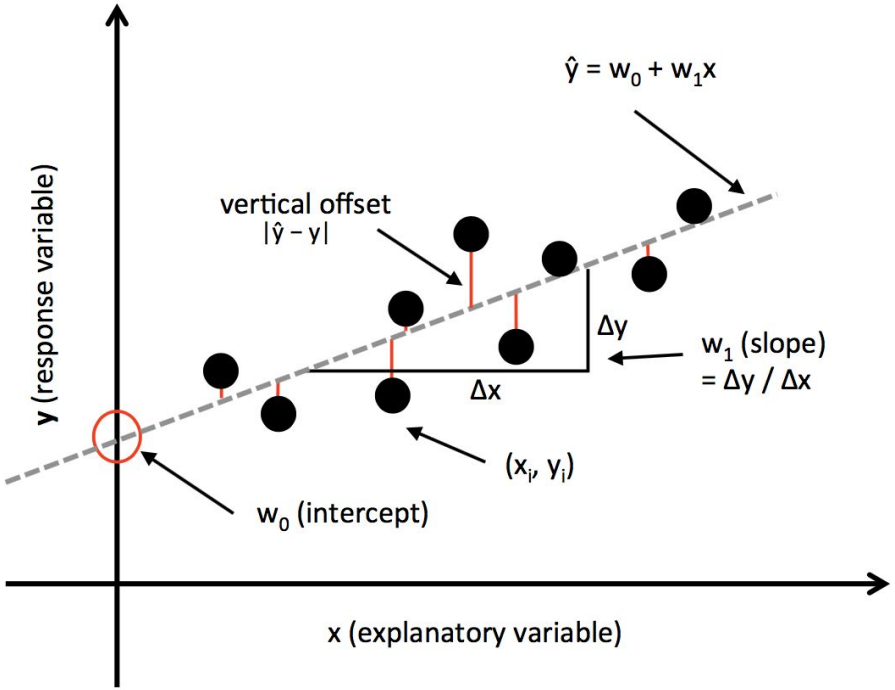
MLDS

Topics

1. Introduction to Machine Learning
2. **Linear Regression**
3. Introduction to Classification
4. Feature Preprocessing & scikit-learn Pipelines
5. Dimensionality Reduction: Feature Selection & Extraction
6. Model Evaluation & Hyperparameter Tuning

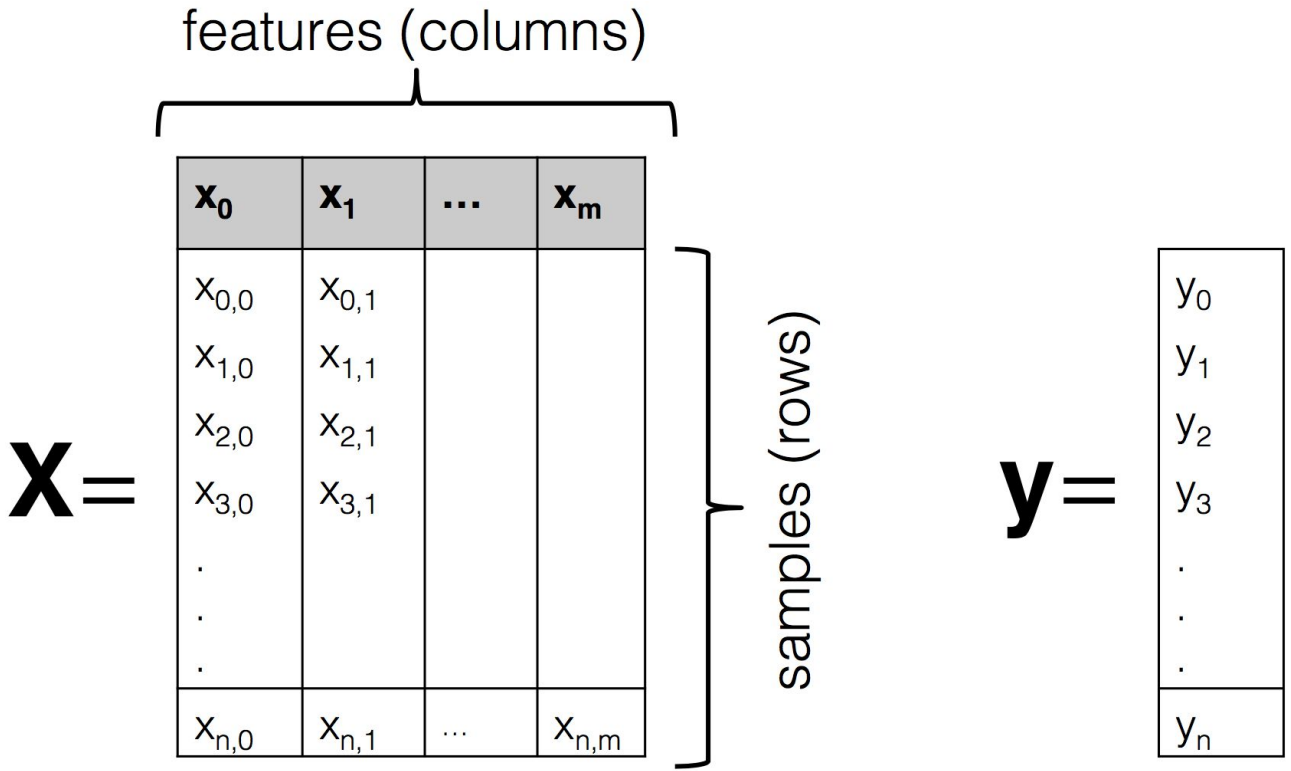
MLDS

Simple Linear Regression



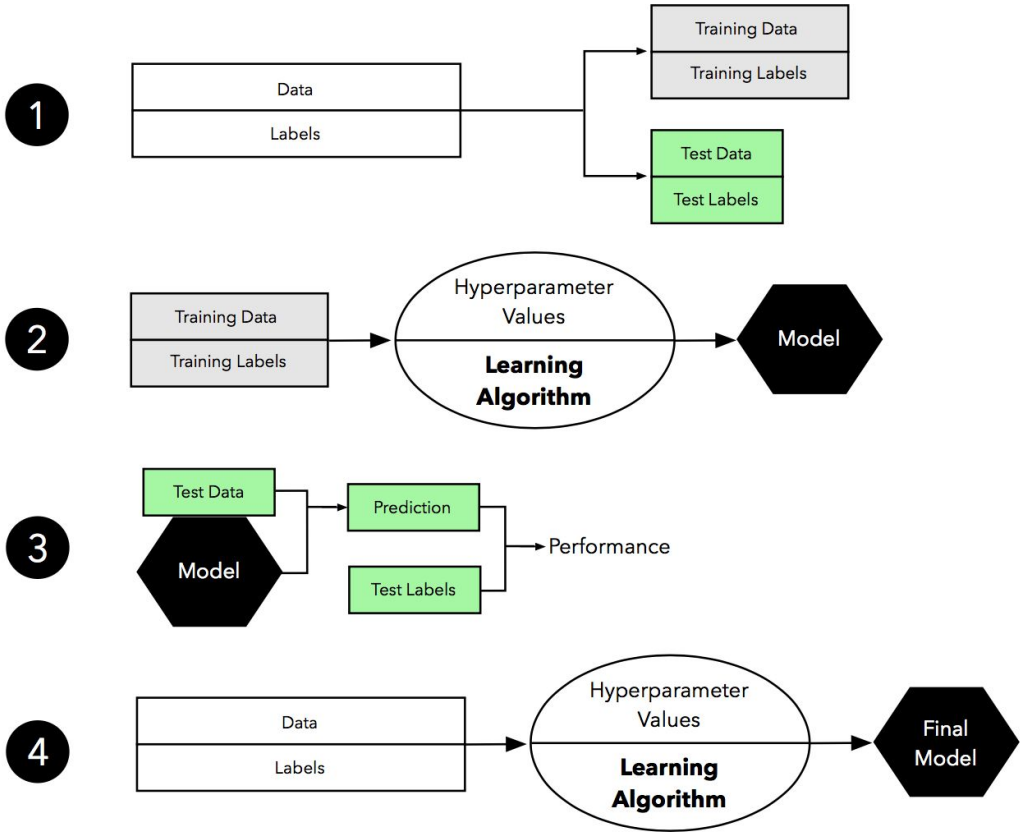
MLDS

Data Representation



MLDS

“Basic” Supervised Learning Workflow



MLDS

Coding Time!

MLDS

Topics

1. Introduction to Machine Learning
2. Linear Regression
3. **Introduction to Classification**
4. Feature Preprocessing & scikit-learn Pipelines
5. Dimensionality Reduction: Feature Selection & Extraction
6. Model Evaluation & Hyperparameter Tuning

MLDS

Scikit-learn API

```
class SupervisedEstimator(...):
    def __init__(self, hyperparam, ...):
        ...
    def fit(self, X, y):
        ...
        return self
    def predict(self, X):
        ...
        return y_pred
    def score(self, X, y):
        ...
        return score
    ...
```

MLDS

Iris Dataset

Iris-Setosa

Iris-Setosa

Iris-Versicolor

MLDS

Iris Dataset

features (columns)

$\mathbf{X} =$

	sepal length [cm]	sepal width [cm]	petal length [cm]	petal width [cm]
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
50	6.4	3.5	4.5	1.2
	.			
	.			
	.			
150	5.9	3.0	5.0	1.8

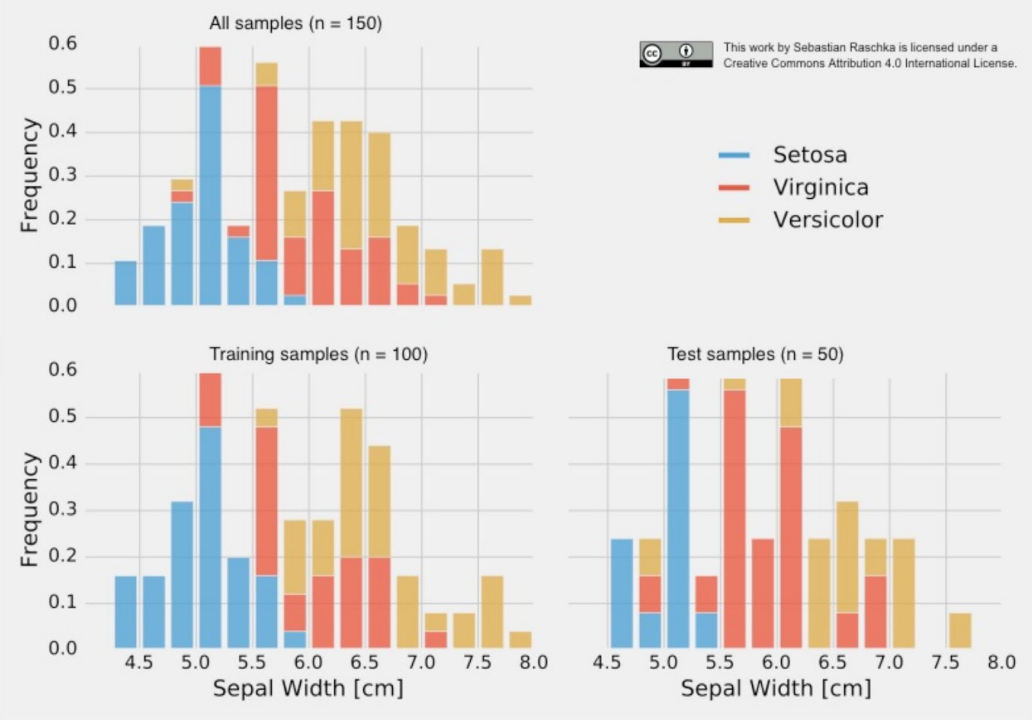
samples (rows)

$\mathbf{y} =$

setosa
setosa
versicolor
.
.
.
virginica

MLDS

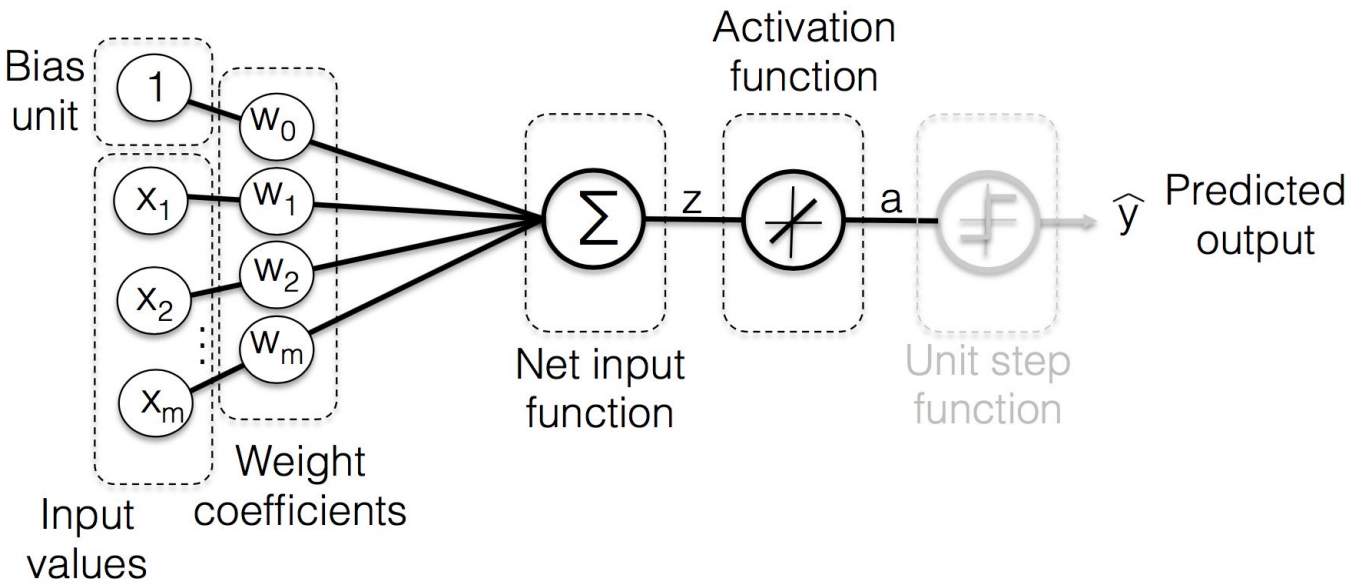
Note about Non-Stratified Splits



- training set → 38 x Setosa, 28 x Versicolor, 34 x Virginica
- test set → 12 x Setosa, 22 x Versicolor, 16 x Virginica

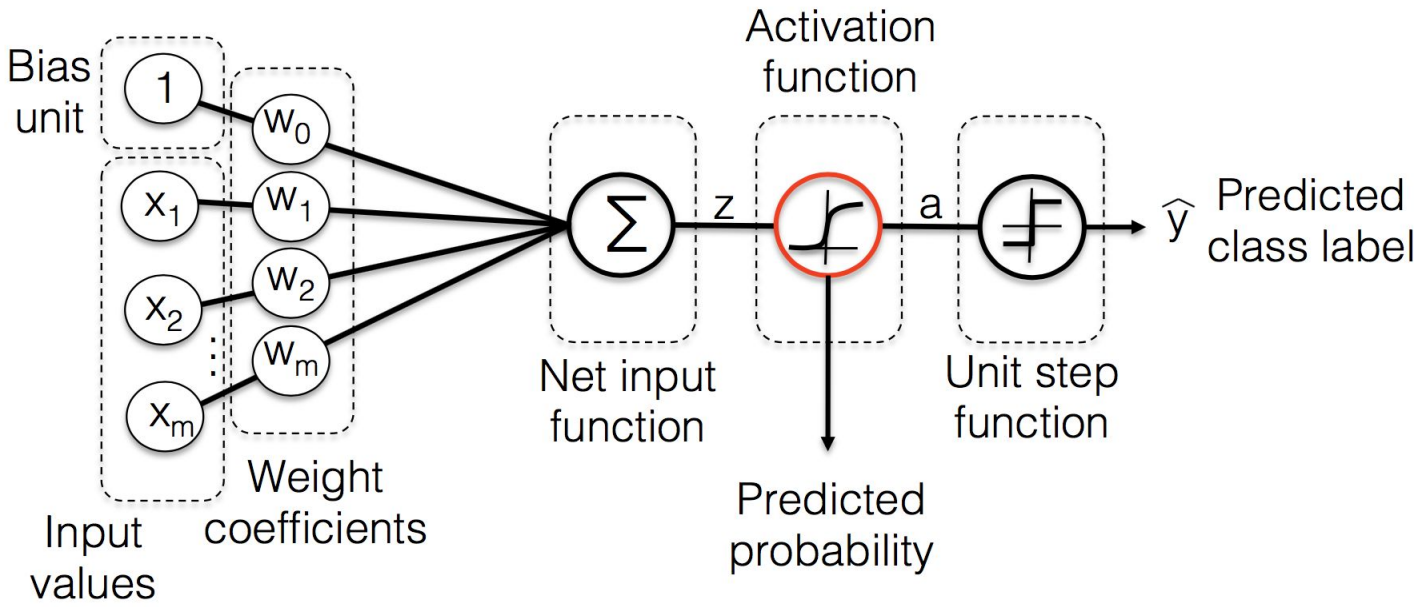
MLDS

Linear Regression Recap



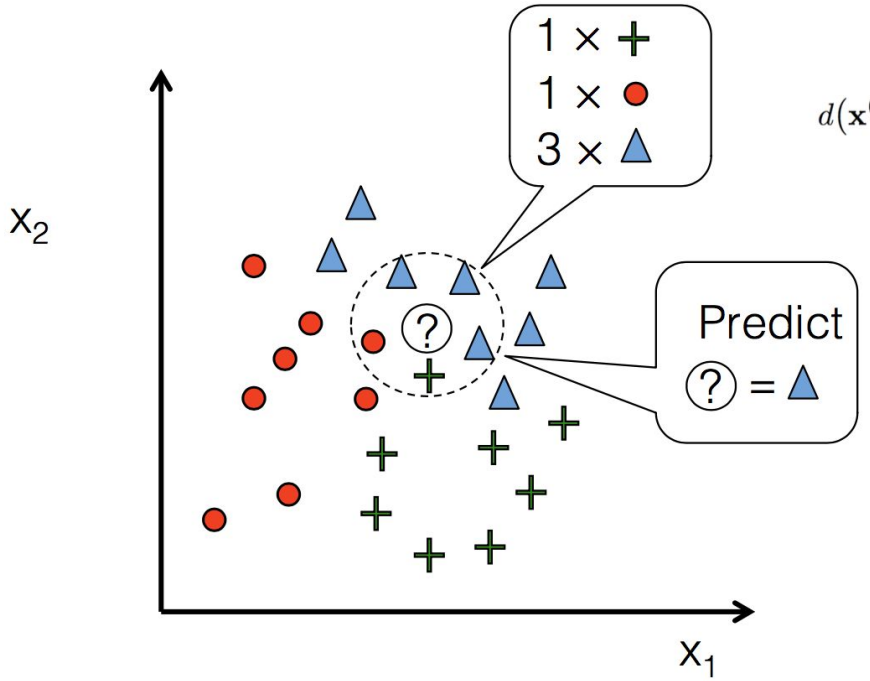
MLDS

Logistic Regression, a Generalized Linear Model



MLDS

A “Lazy Learner:” K-Nearest Neighbors Classifier



$$d(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \sqrt[p]{\sum_k |x_k^{(i)} - x_k^{(j)}|^p}$$

MLDS

Coding Time!

MLDS

Topics

1. Introduction to Machine Learning
2. Linear Regression
3. Introduction to Classification
4. **Feature Preprocessing & scikit-learn Pipelines**
5. Dimensionality Reduction: Feature Selection & Extraction
6. Model Evaluation & Hyperparameter Tuning

MLDS

Categorical Variables

color	size	price	class label
red	M	$10.49	0
blue	XL	$15.00	1
green	L	$12.99	1

MLDS

Encoding Categorical Variables

color	size	price	class label
red	M	$10.49	0
blue	XL	$15.00	1
green	L	$12.99	1

red	blue	green
1	0	0
0	1	0
0	0	1

size
0
2
1

MLDS

Feature Normalization

Min-max scaling

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Z-score standardization

$$z = \frac{x - \mu}{\sigma}$$

feature	minmax	z-score
1.0	0.0	-1.46385
2.0	0.2	-0.87831
3.0	0.4	-0.29277
4.0	0.6	0.29277
5.0	0.8	0.87831
6.0	1.0	1.46385

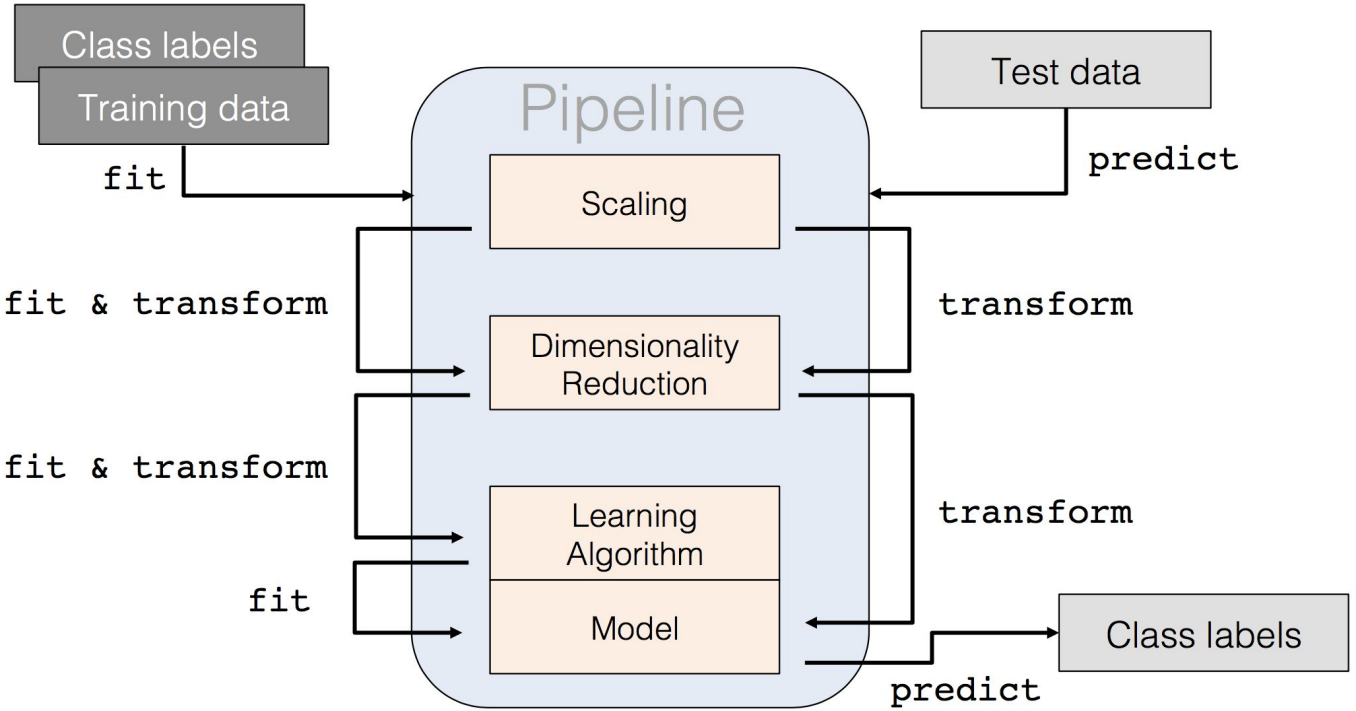
MLDS

Scikit-learn API

```
class UnsupervisedEstimator(...):
    def __init__(self, ...):
        ...
    def fit(self, X):
        ...
        return self
    def transform(self, X):
        ...
        return X_transf
def predict(self, X):
        ...
        return pred
```

MLDS

Scikit-learn Pipelines



MLDS

Coding Time!

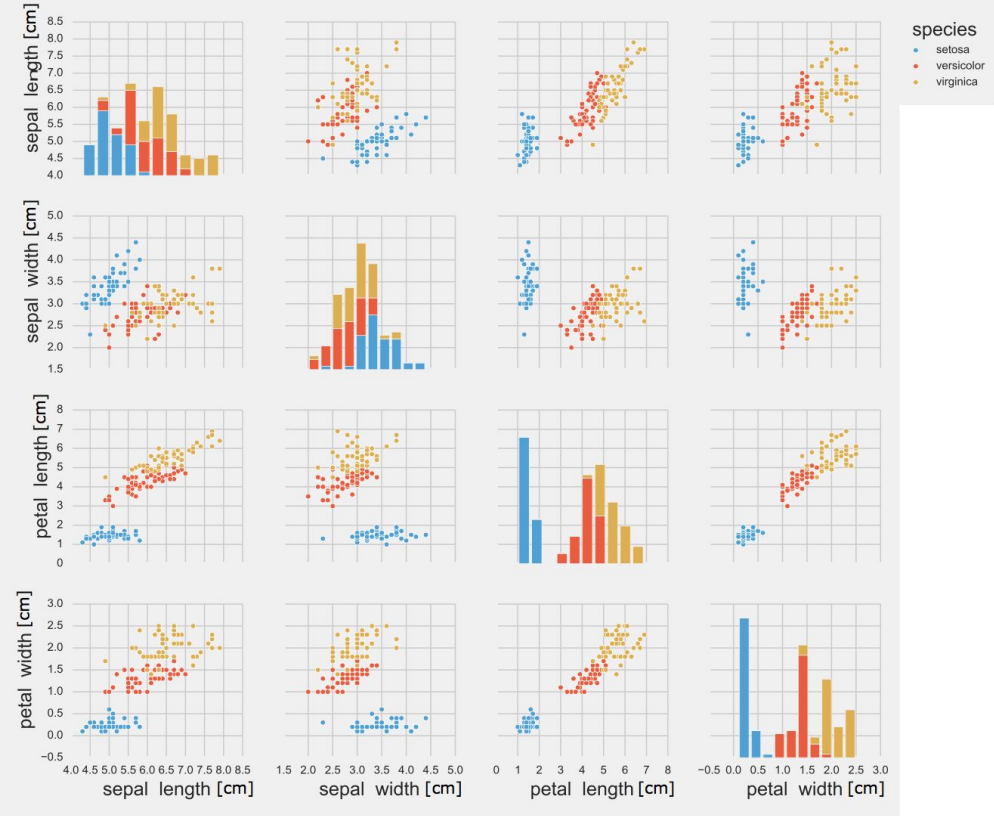
MLDS

Topics

1. Introduction to Machine Learning
2. Linear Regression
3. Introduction to Classification
4. Feature Preprocessing & scikit-learn Pipelines
5. **Dimensionality Reduction: Feature Selection & Extraction**
6. Model Evaluation & Hyperparameter Tuning

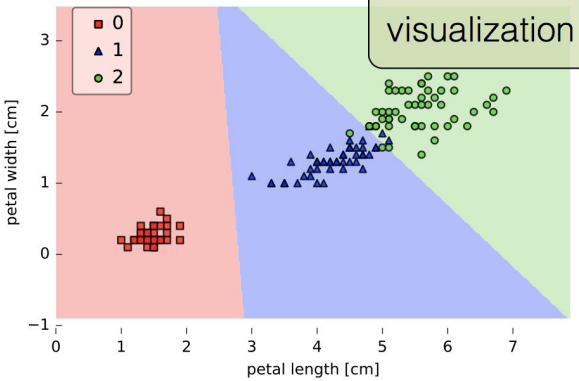
MLDS

Dimensionality Reduction – why?



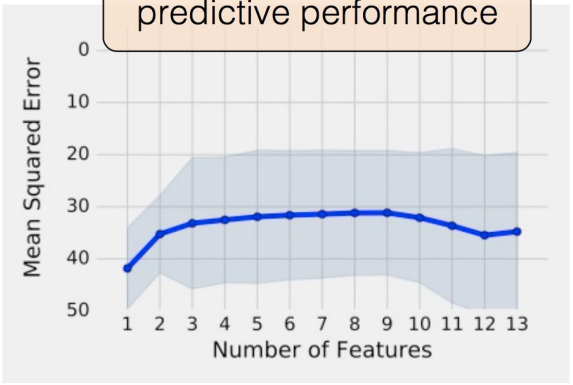
MLDS

Dimensionality Reduction – why?



visualization & interpretability

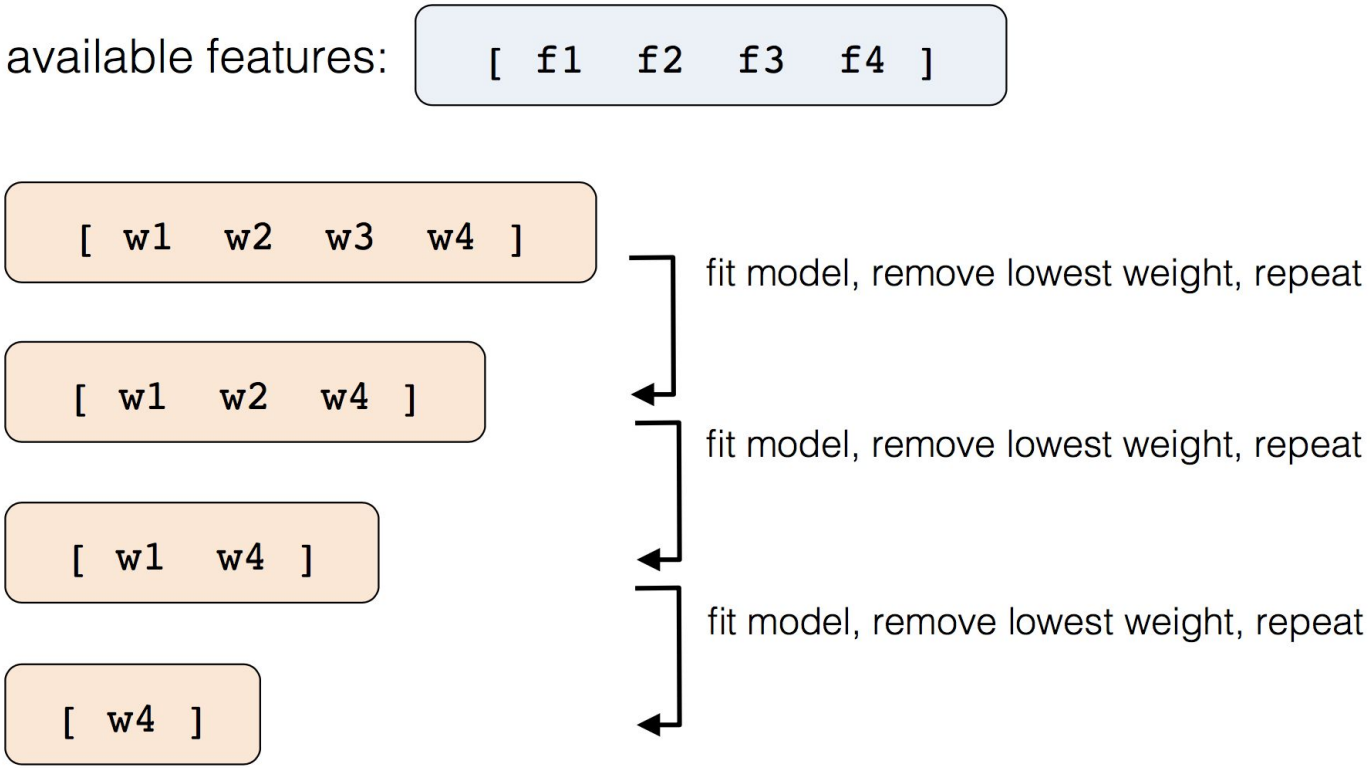
storage & speed



predictive performance

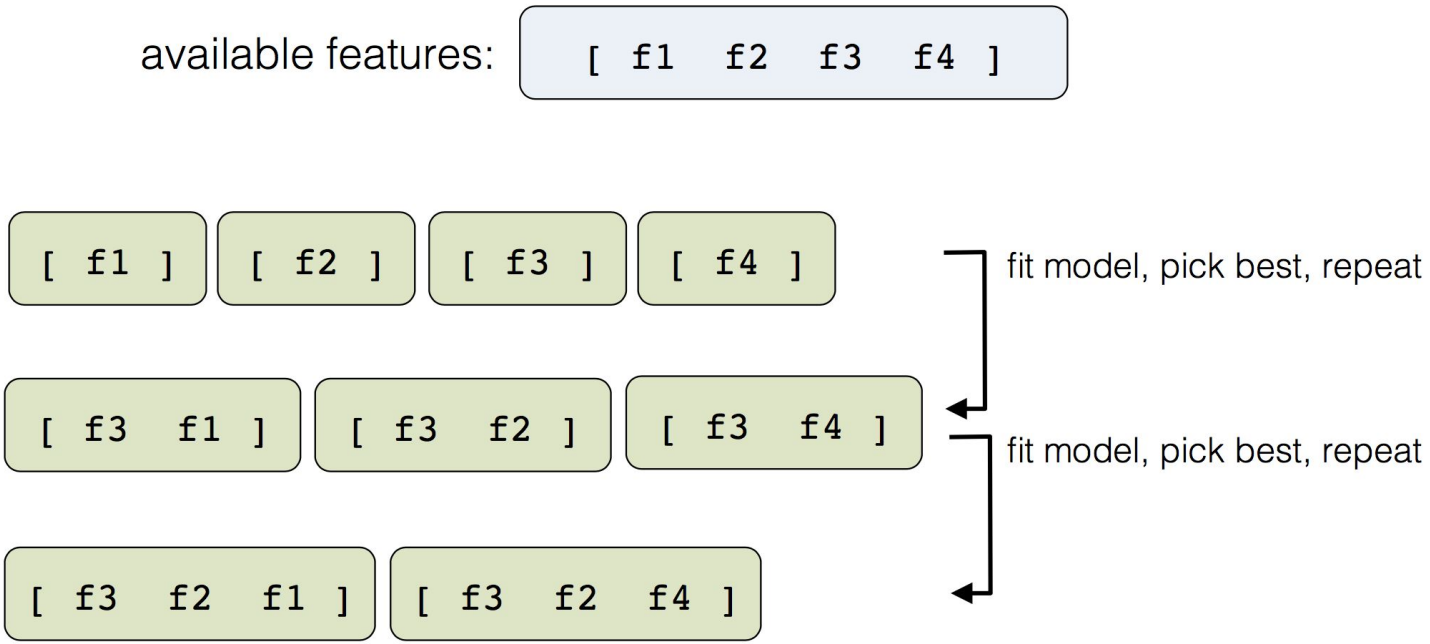
MLDS

Recursive Feature Elimination



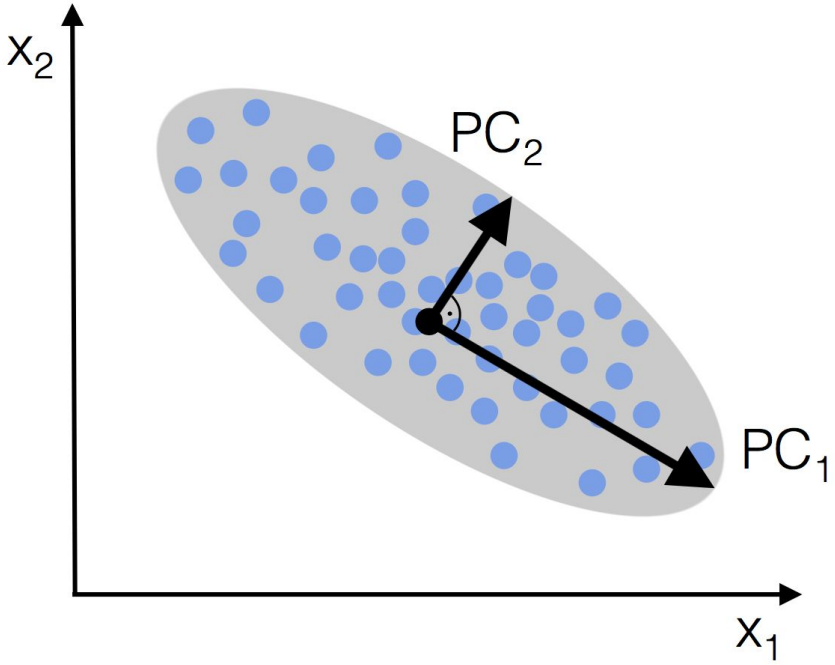
MLDS

Sequential Feature Selection



MLDS

Principal Component Analysis



MLDS

Coding Time!

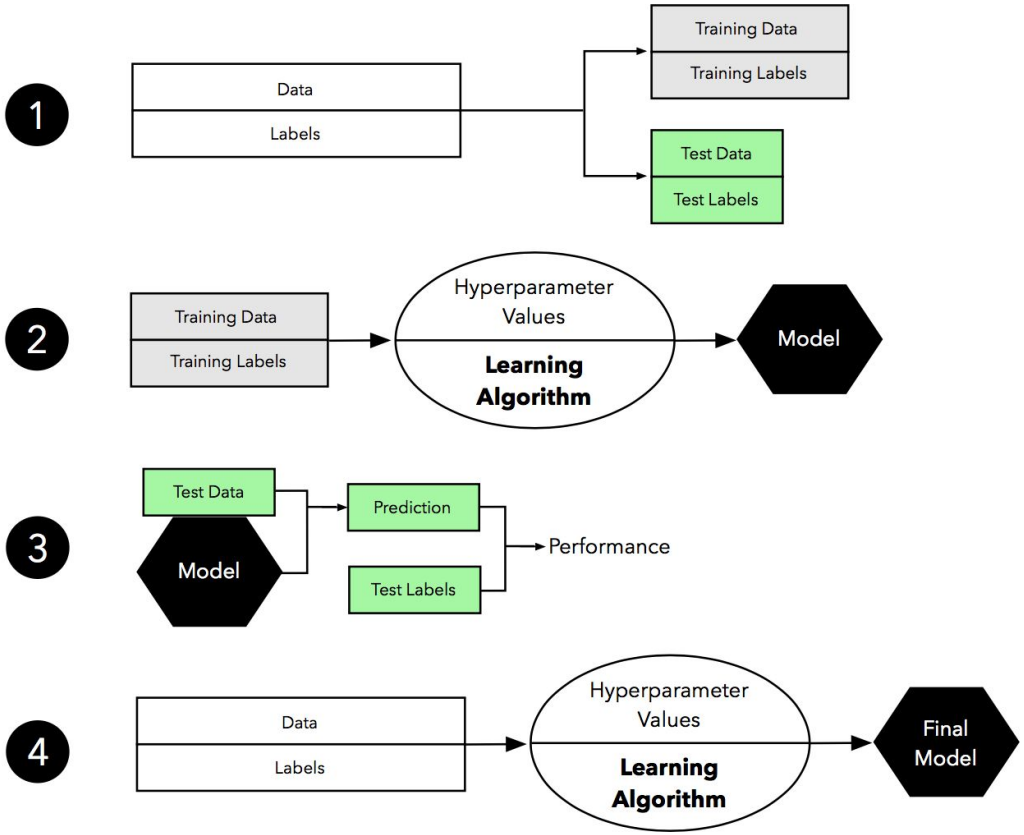
MLDS

Topics

1. Introduction to Machine Learning
2. Linear Regression
3. Introduction to Classification
4. Feature Preprocessing & scikit-learn Pipelines
5. Dimensionality Reduction: Feature Selection & Extraction
6. **Model Evaluation & Hyperparameter Tuning**

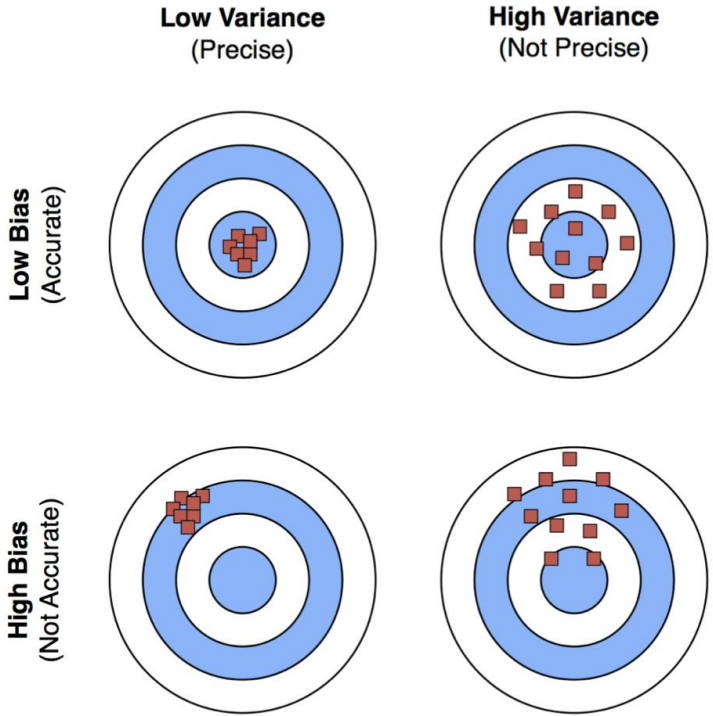
MLDS

“Basic” Supervised Learning Workflow



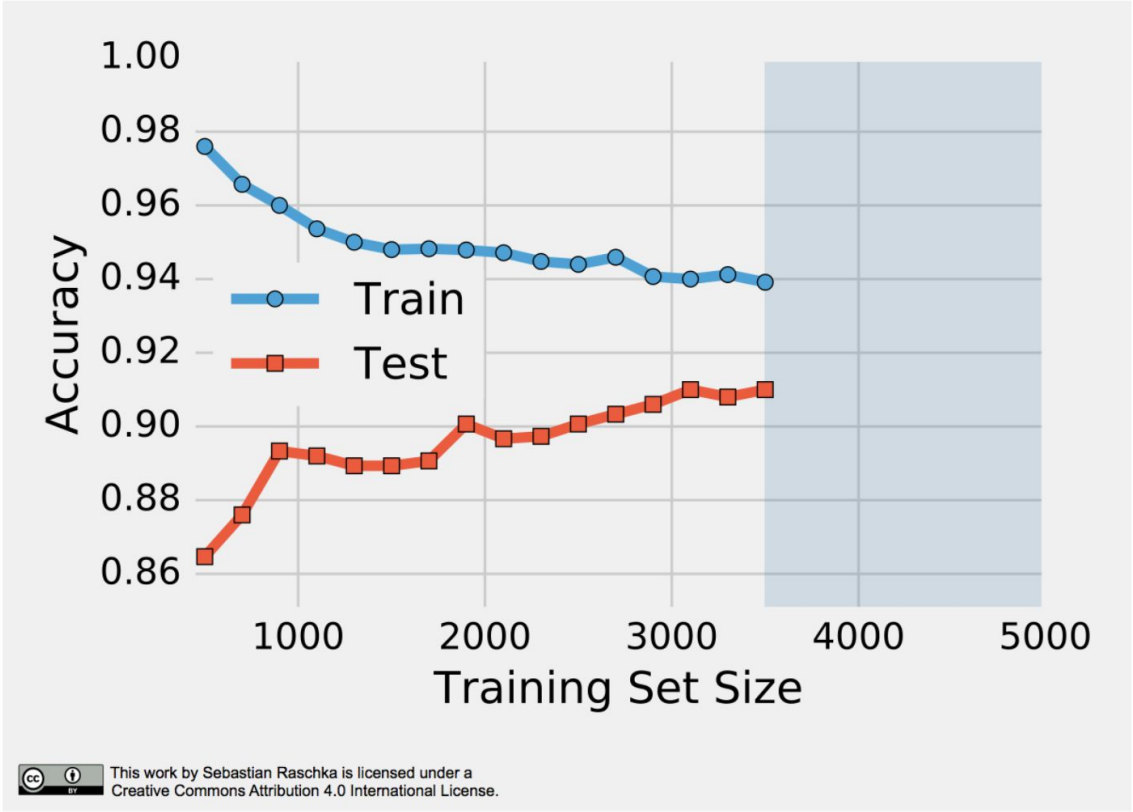
MLDS

Bias and Variance



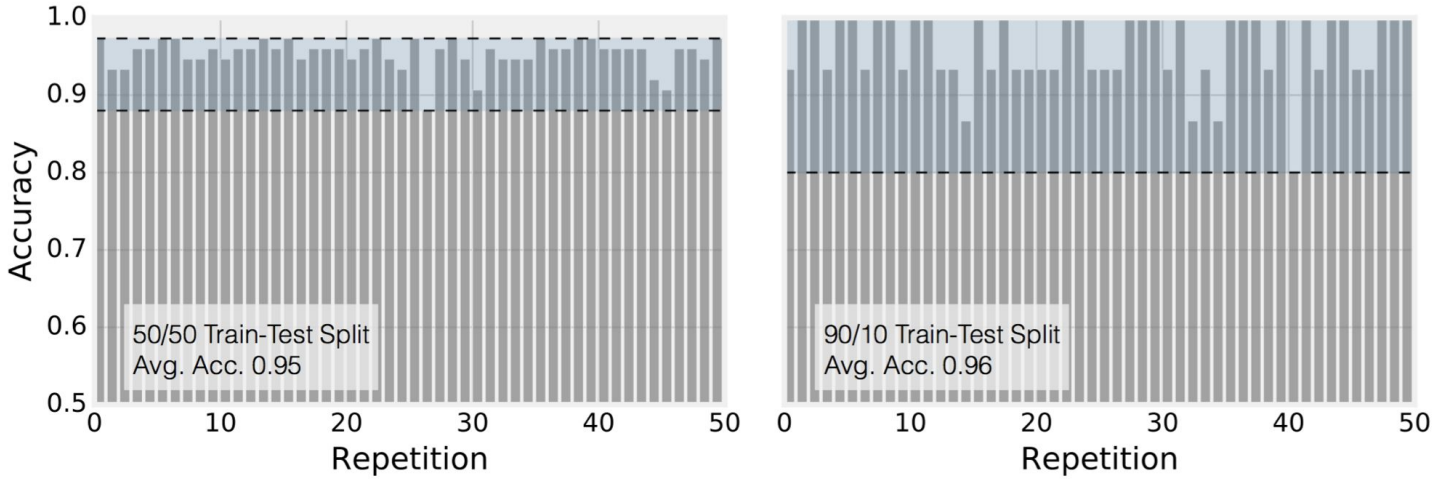
MLDS

Learning Curves



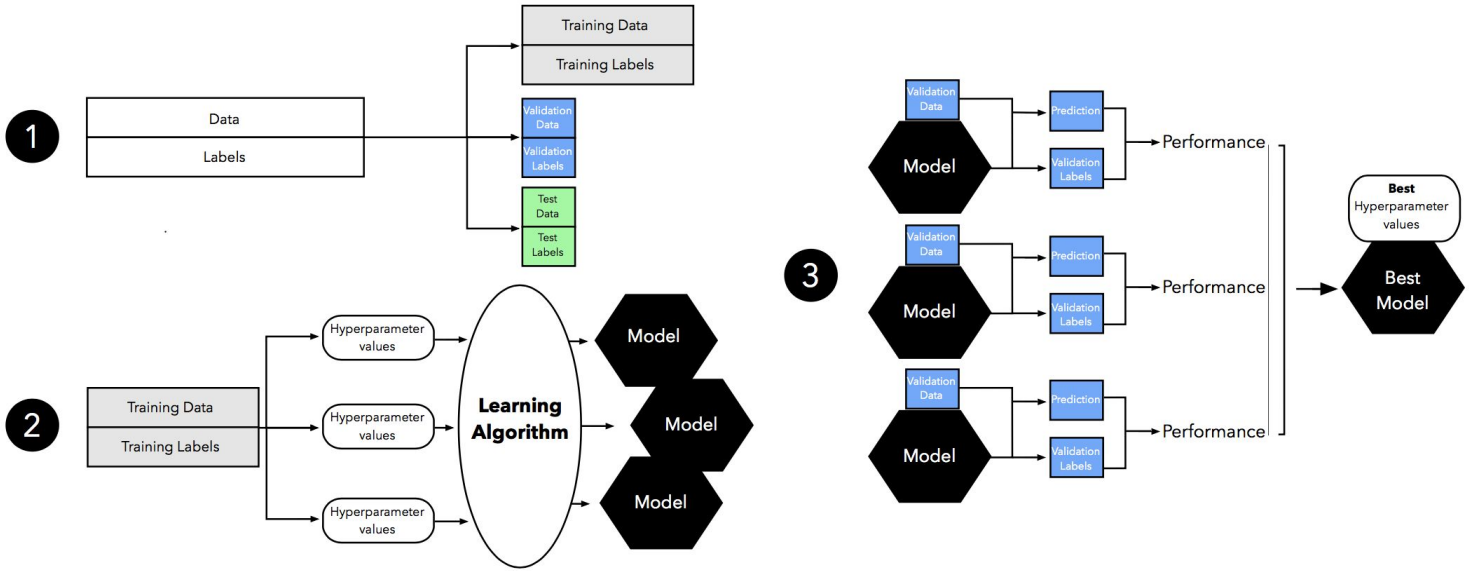
MLDS

Repeated Holdout



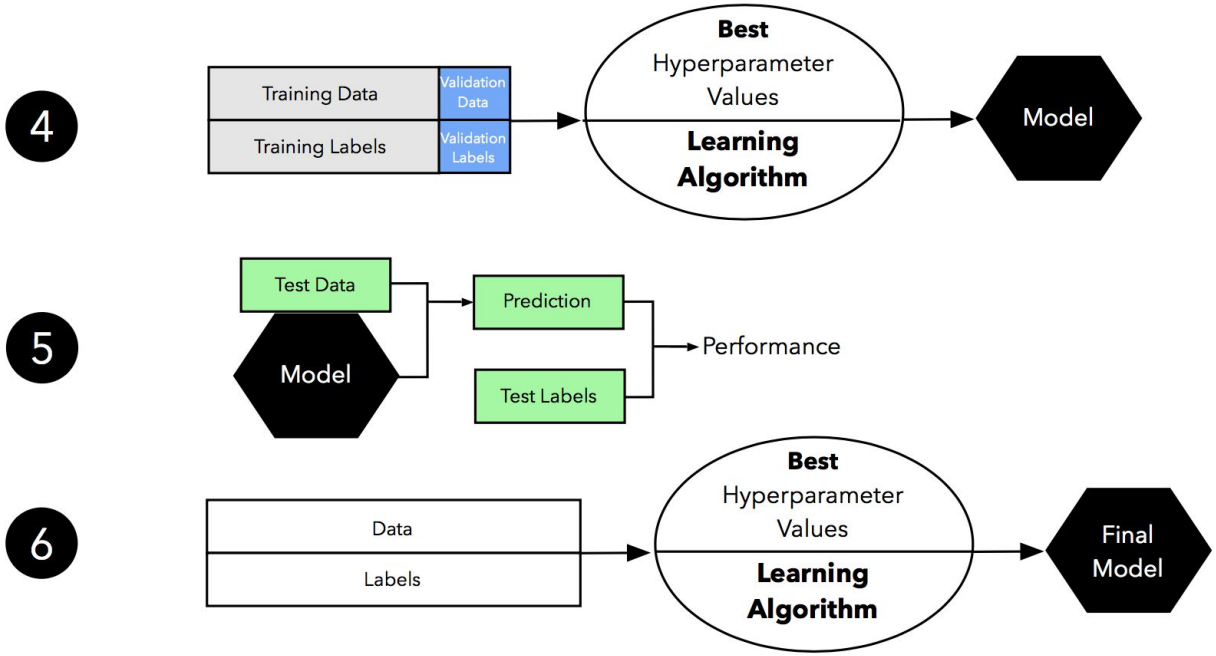
MLDS

Holdout and Hyperparameter Tuning I



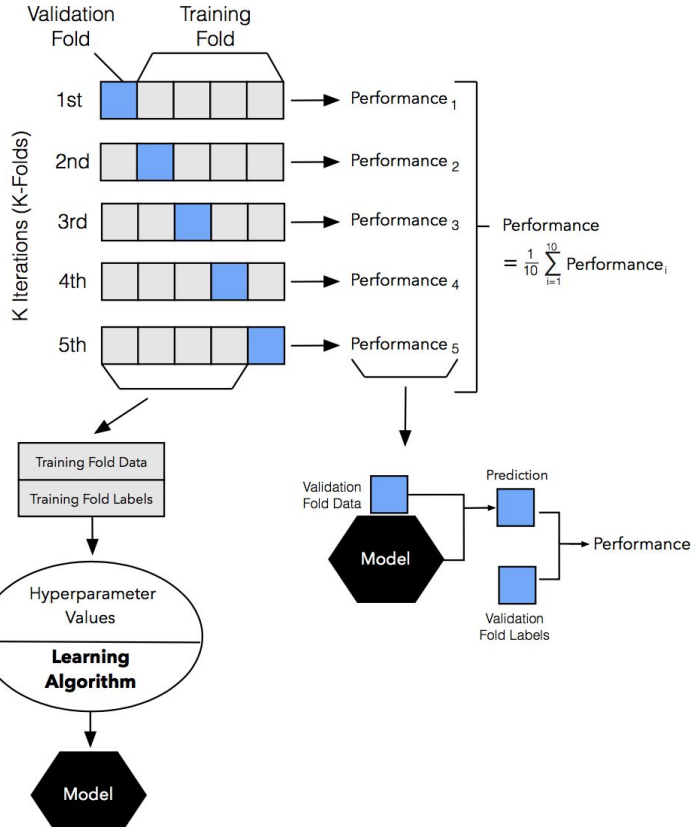
MLDS

Holdout and Hyperparameter Tuning II



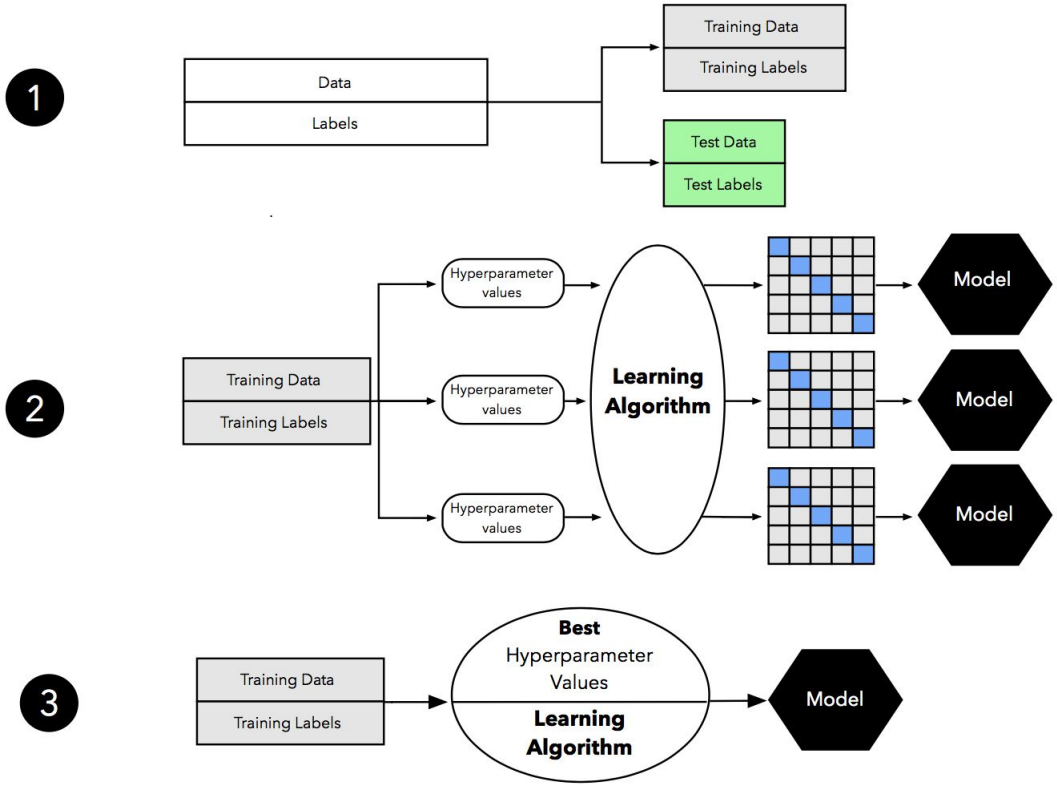
MLDS

K-fold Cross-Validation



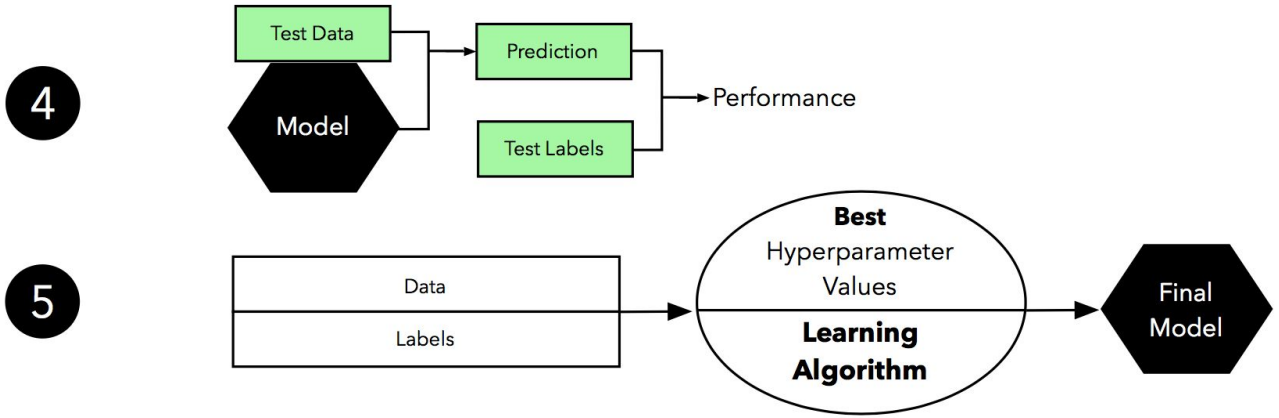
MLDS

K-fold Cross-Validation Workflow I



MLDS

K-fold Cross-Validation Workflow II



MLDS

Coding Time!

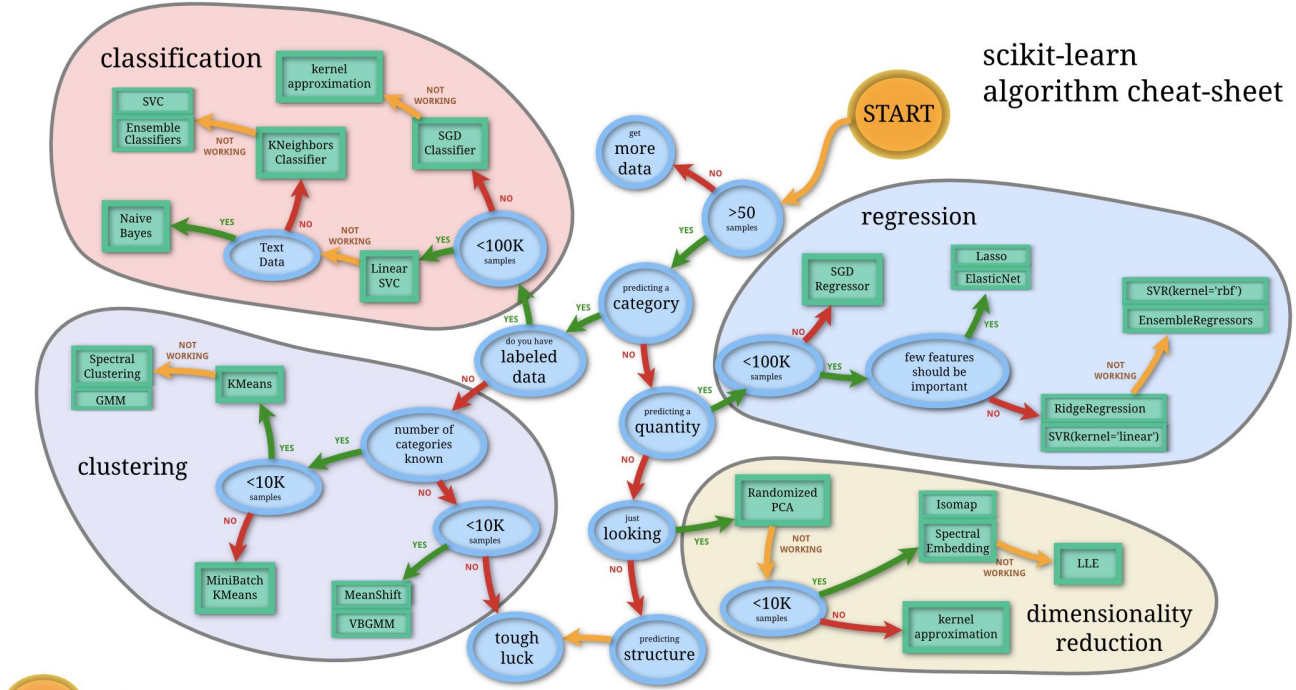
MLDS

Performance Metrics

http://scikit-learn.org/stable/modules/model_evaluation.html

MLDS

Further Resources - Andreas' "cheat sheet"



Thanks!

Machine Learning for Data Science Interest Group
Advanced Informatics School
Universiti Teknologi Malaysia

@utmmlds
ais.utm.my/mlds

May 2017