به نام خدا

کارگاه علم داده با پایتون پیشرفته

جلسه اول : مقدمه ای بر یادگیری ماشین و مروری بر کتابخانه های Numpy و Pandas

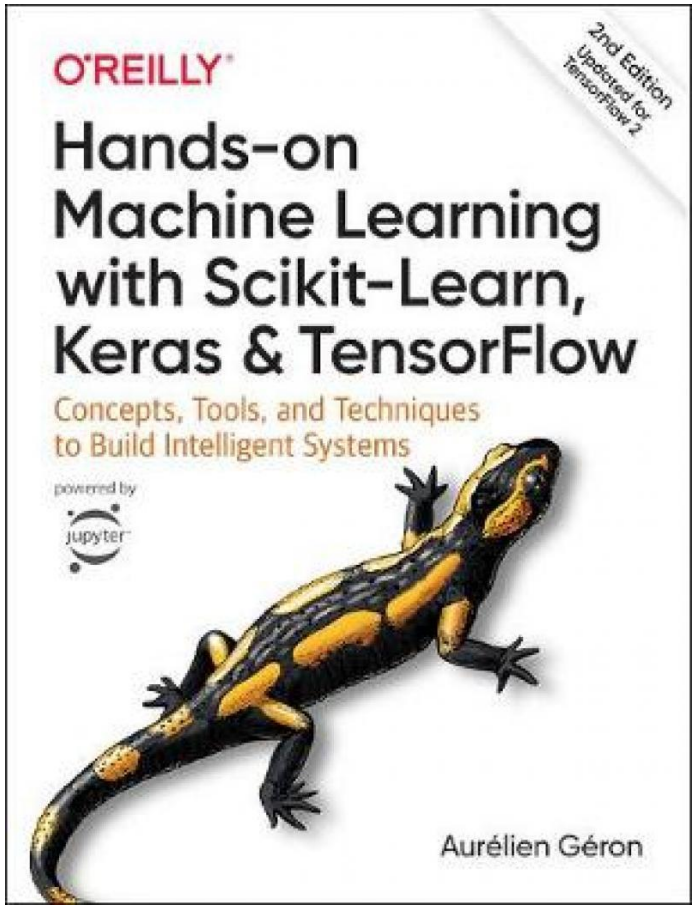
مدرس :

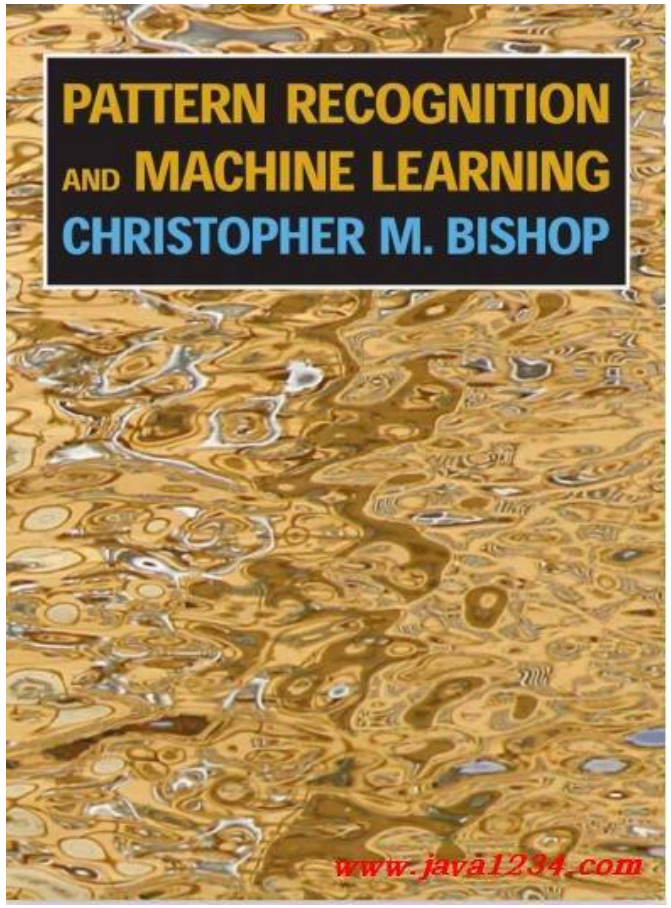
مهرناز جلیلی

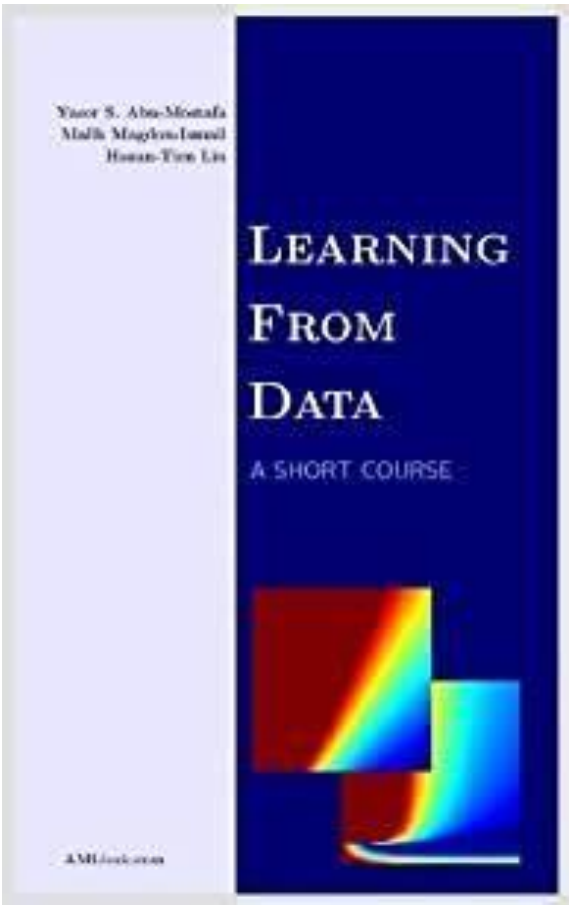
دانشجو کارشناسی ارشد علم داده ها

دانشگاه شهید بهشتی

منابع


O'REILLY
2nd Edition
Updated for TensorFlow 2
Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow
Concepts, Tools, and Techniques to Build Intelligent Systems
powered by
jupyter
Aurélien Géron


PATTERN RECOGNITION AND MACHINE LEARNING
CHRISTOPHER M. BISHOP
www.java1234.com


Yaser S. Abu-Mostafa
Malik Magdon-Ismail
Hsuan-Tien Lin
LEARNING FROM DATA
A SHORT COURSE

سرفصل ها

درس ۱: مقدمه ای بر یادگیری ماشین (تئوری)
مروری بر کتابخانه های Numpy و Pandas

درس ۲: کار با دیتا فریم ها در پکیج Pandas (عملی)
پاک سازی داده ها

درس ۳: رگرسیون خطی ساده و چندگانه (تئوری و عملی)

درس ۴: طبقه بندی یا classification (تئوری)
K نزدیک ترین همسایه (KNN) (تئوری و عملی)

درس ۵: رگرسیون لجستیک (تئوری و عملی)

درس۶: درخت تصمیم (تئوری و عملی)

درس ۷: ماشین بردار پشتیبان (SVM) (تئوری و عملی)

درس۸: تقلیل ابعاد (تئوری و عملی)

درس ۹: کشف داده های پرت (تئوری و عملی)

درس ۱۰: Case study

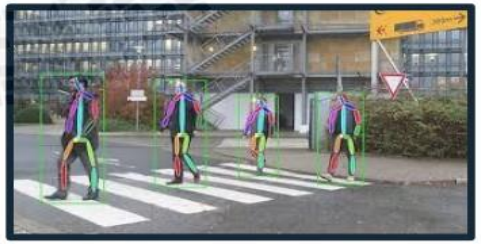
مقدمه ای بر یادگیری ماشین

یادگیری ماشین در ابتدا زیر شاخه ای از هوش مصنوعی محسوب میشد ولی در حال حاضر بهترین رویکرد برای حل مسائل هوش مصنوعی است.

بینایی ماشین

(computer vision)

تشخیص اشیاء داخل تصویر



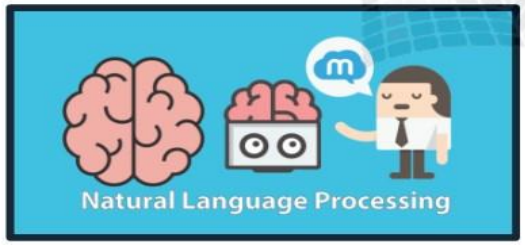
تشخیص صوت

(Speech Recognition)

تشخیص کلمات سیگنال صوتی

Robotics

پردازش زبان طبیعی



برای این که بتوان یک مساله را با تکنیک های یادگیری ماشین حل کرد:

۱- الگویی باید وجود داشته باشد

۲- این الگو توسط ریاضی شناخته شده باشد

۳- داده ای داشته باشیم

Clump	UnifSize	UnifShape	MargAdh	SingEpiSize	BareNuc	BlandChrom	NormNucl	Mit	Class
5	1	1	1	2	1	3	1	1	benign
5	4	4	5	7	10	3	2	1	benign
3	1	1	1	2	2	3	1	1	malignant
6	8	8	1	3	4	3	7	1	benign
4	1	1	3	2	1	3	1	1	benign
8	10	10	8	7	10		7	1	malignant
1	1	1	1	2	10	3	1	1	benign
2	1	2	H	2	1	3	1	1	benign
2	1	1	1	2	1	1	1	5	benign
4	2	1	1	2	1	2	1	1	benign

انواع الگوریتم های یادگیری

۱- یادگیری با ناظر (Supervised Learning)

۲- یادگیری بدون ناظر (Unsupervised Learning)

۳- یادگیری نیمه نظارتی (Semi_supervised Learning)

۴- یادگیری تقویتی (Reinforcement Learning)

Supervised vs. Unsupervised

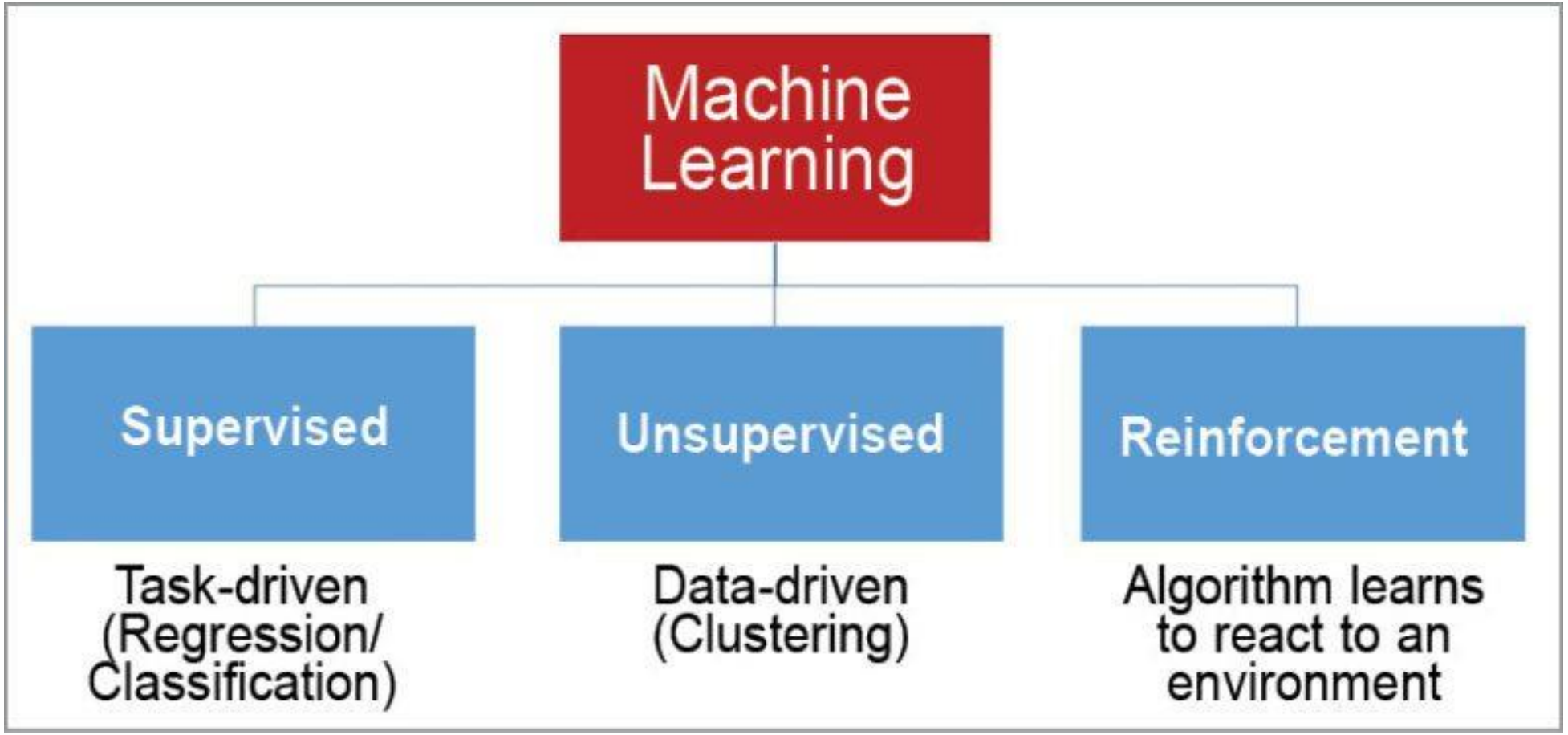
Supervised Learning

- **Classification:**
  Classifies labeled data
- **Regression:**
  Predicts trends using previous labeled data
- Has more evaluation methods than unsupervised learning
- Controlled environment

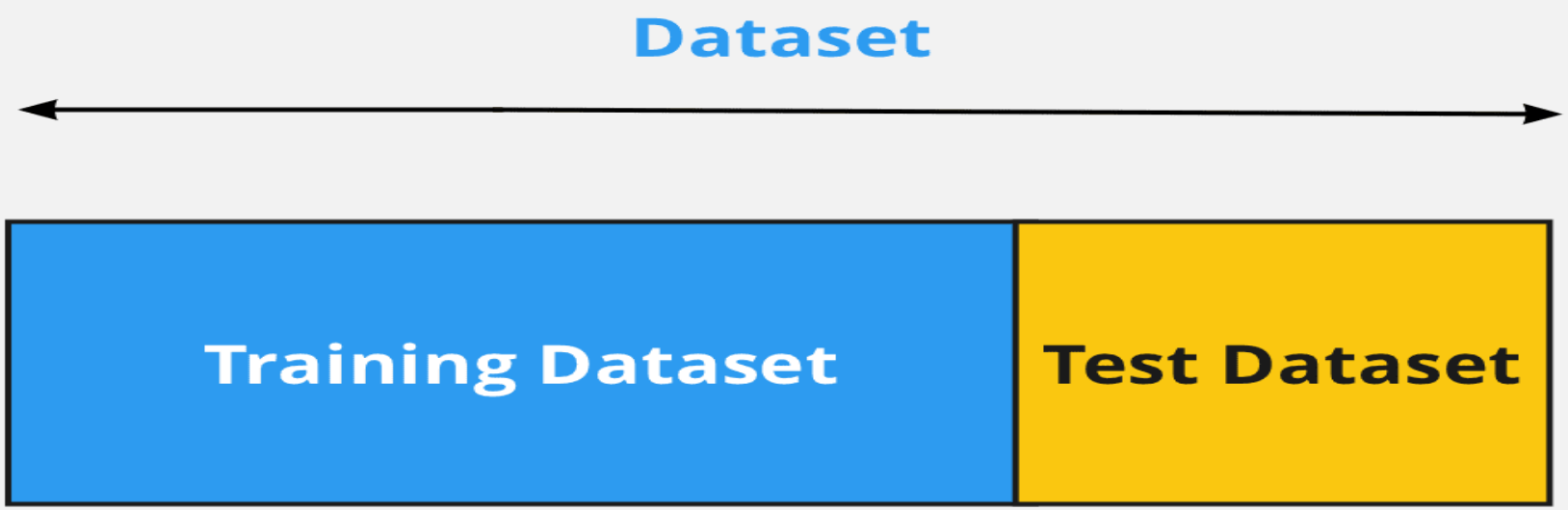
Unsupervised Learning

- **Clustering:**
  Finds patterns and groupings from unlabeled data
- Has fewer evaluation methods than supervised learning
- Less controlled environment

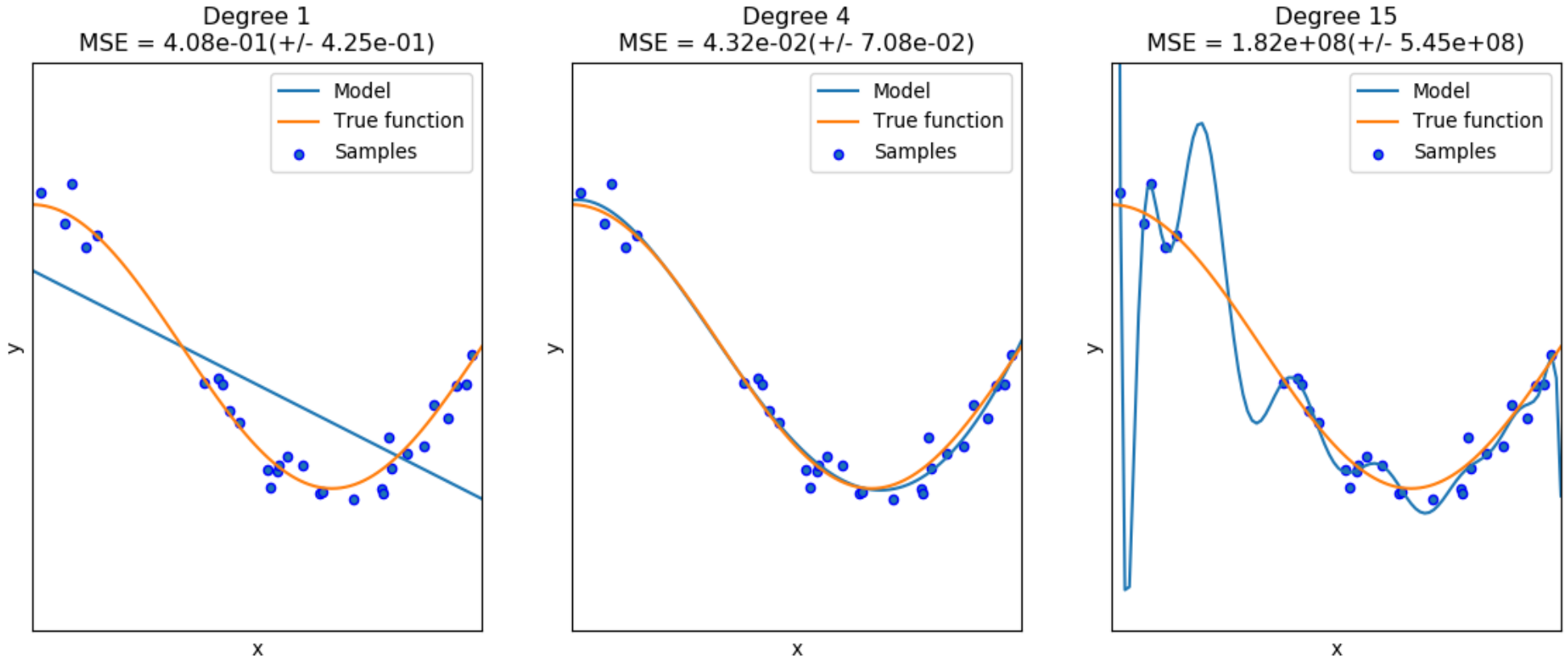
اصطلاحات و مفاهیم آن ها


Machine Learning
Supervised
Unsupervised
Reinforcement
Task-driven (Regression/ Classification)
Data-driven (Clustering)
Algorithm learns to react to an environment

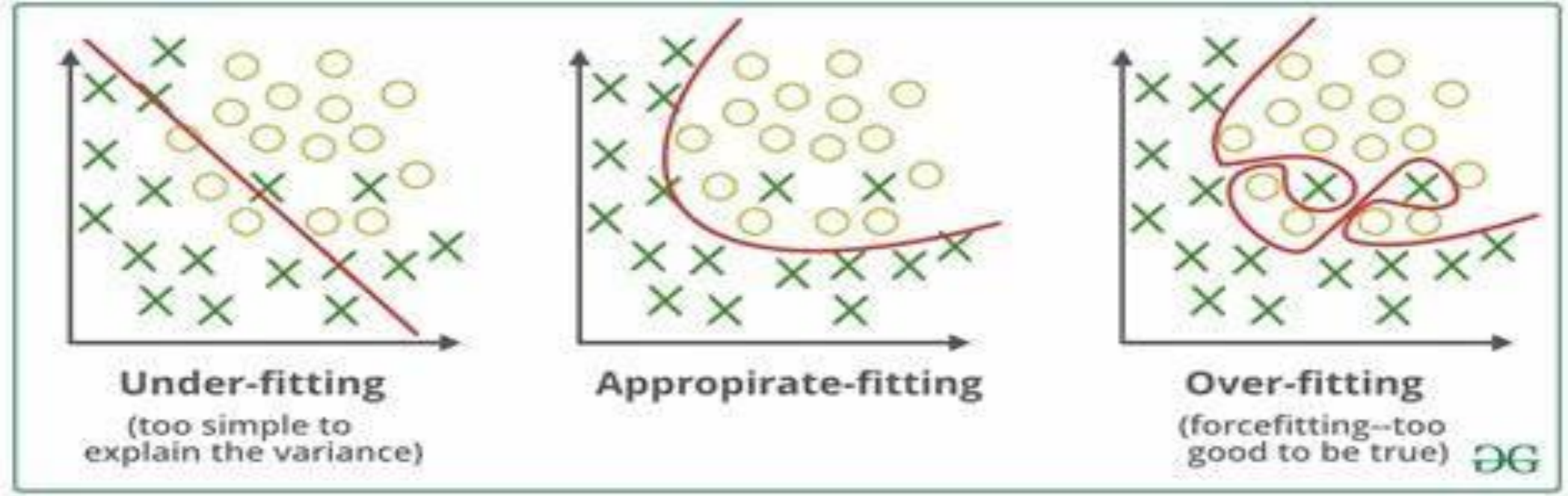
Train & Test



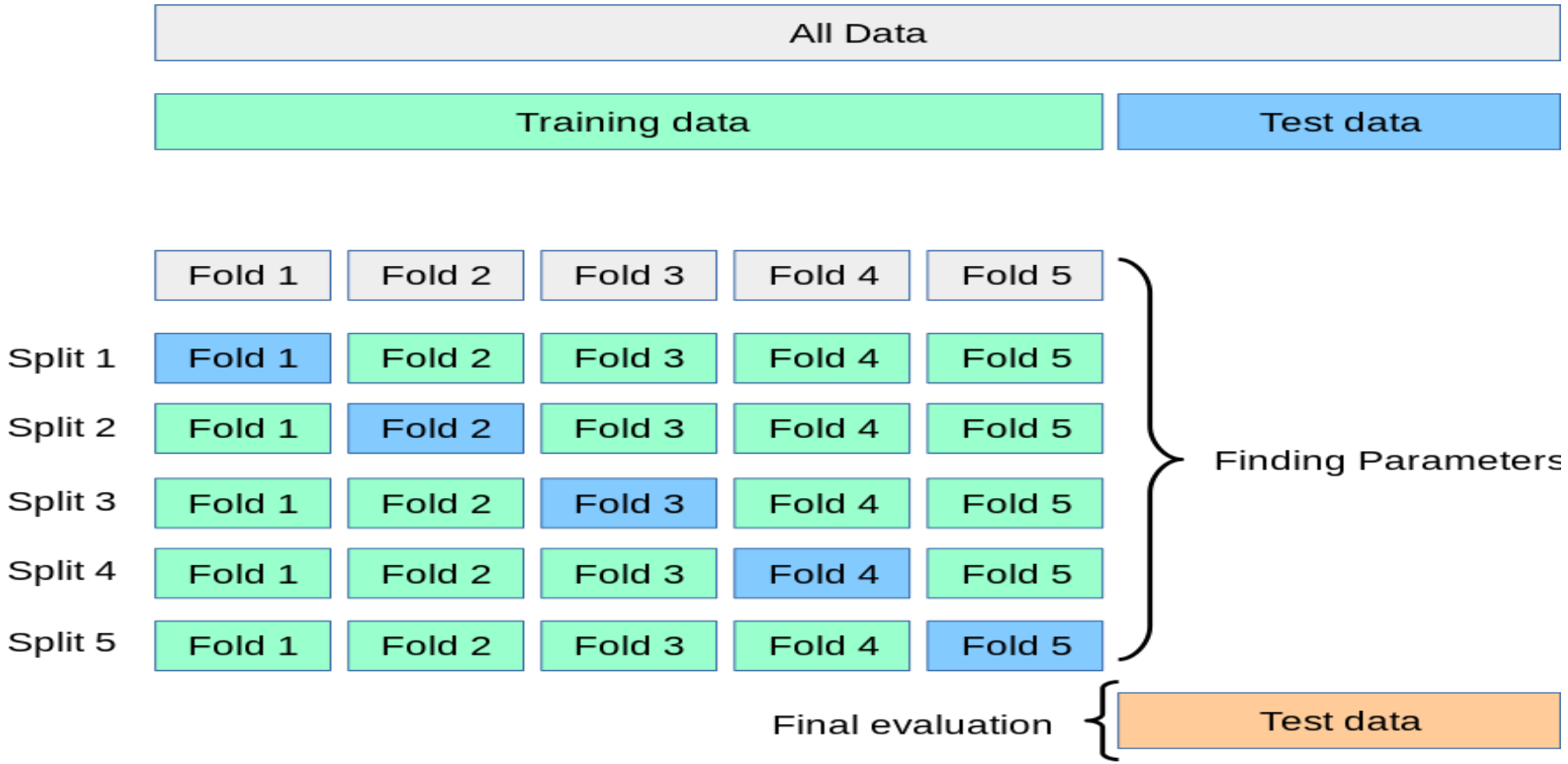
Overfit vs. Underfit



Overfit vs. Underfit

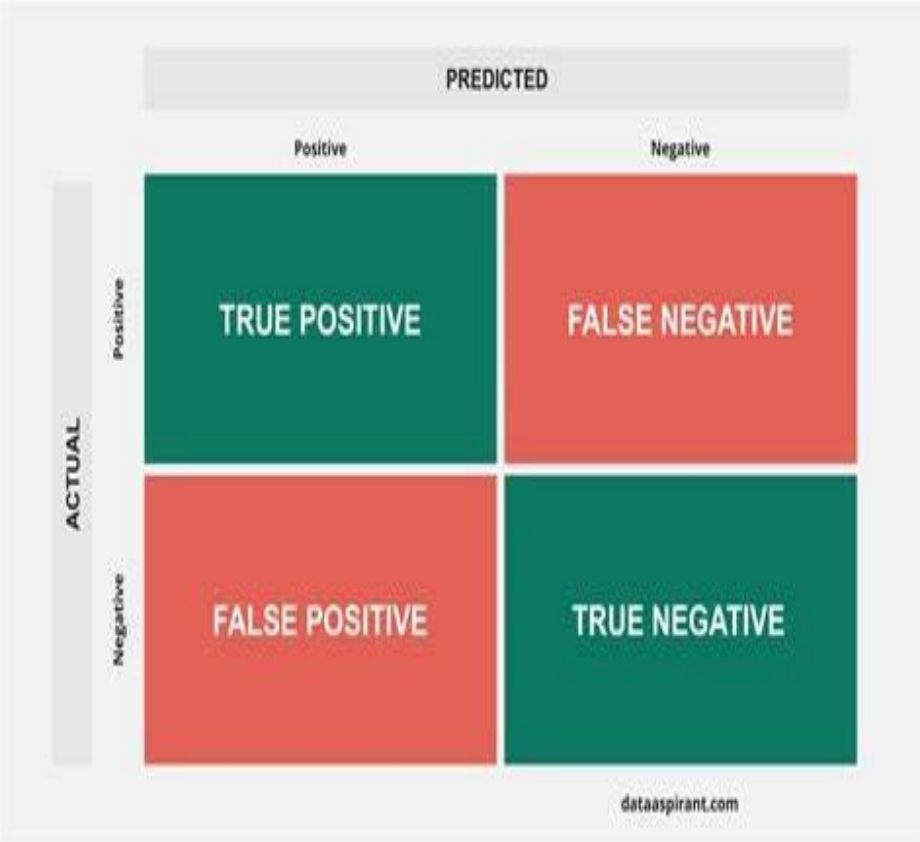


Cross-validation



ارزیابی سیستم های یادگیری

ماتریس درهم ریختگی (confusion matrix)



$$Accuracy = \frac{T_p + T_n}{T_p + T_n + F_p + F_n}$$

$$Precision = \frac{T_p}{T_p + F_p}$$

$$Recall = \frac{T_p}{T_p + T_n}$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

محاسبه کنید

Actual \Predicted	Spam +	Real −
Spam +	1558	255
Real −	125	2633

Accuracy = 92%

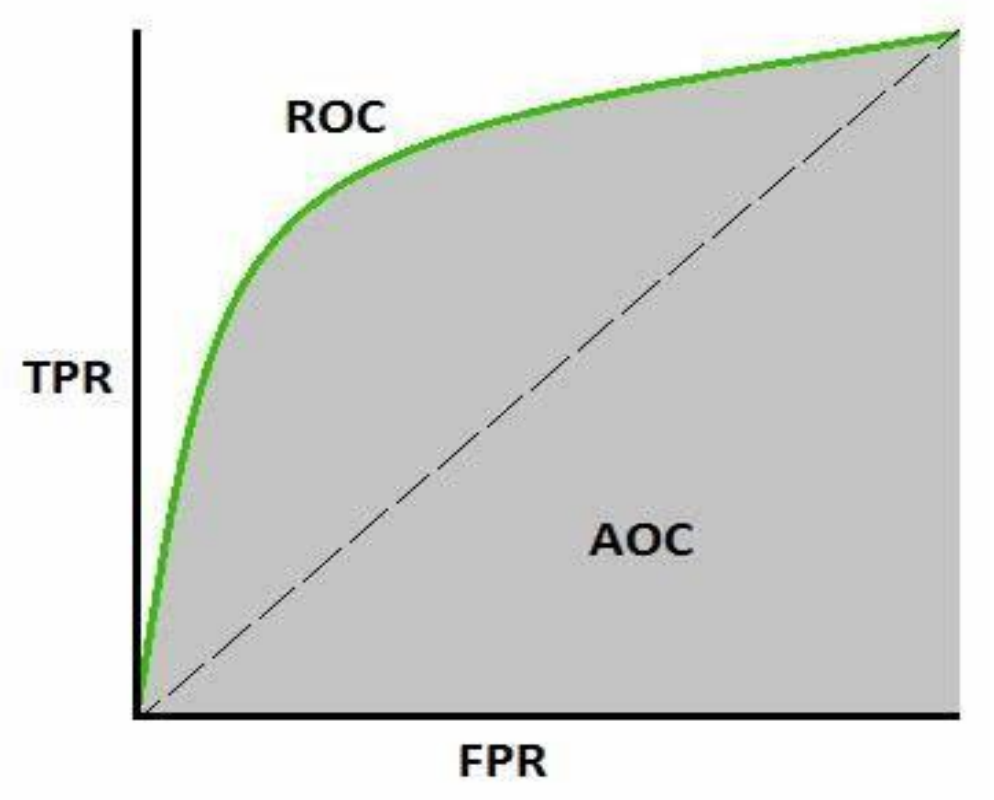
Precision = 92.5%

Actual \Predicted	Cancer = yes	Cancer = no
Cancer = yes	90	210
Cancer = no	140	9560

Accuracy = 96%

recall = 30%

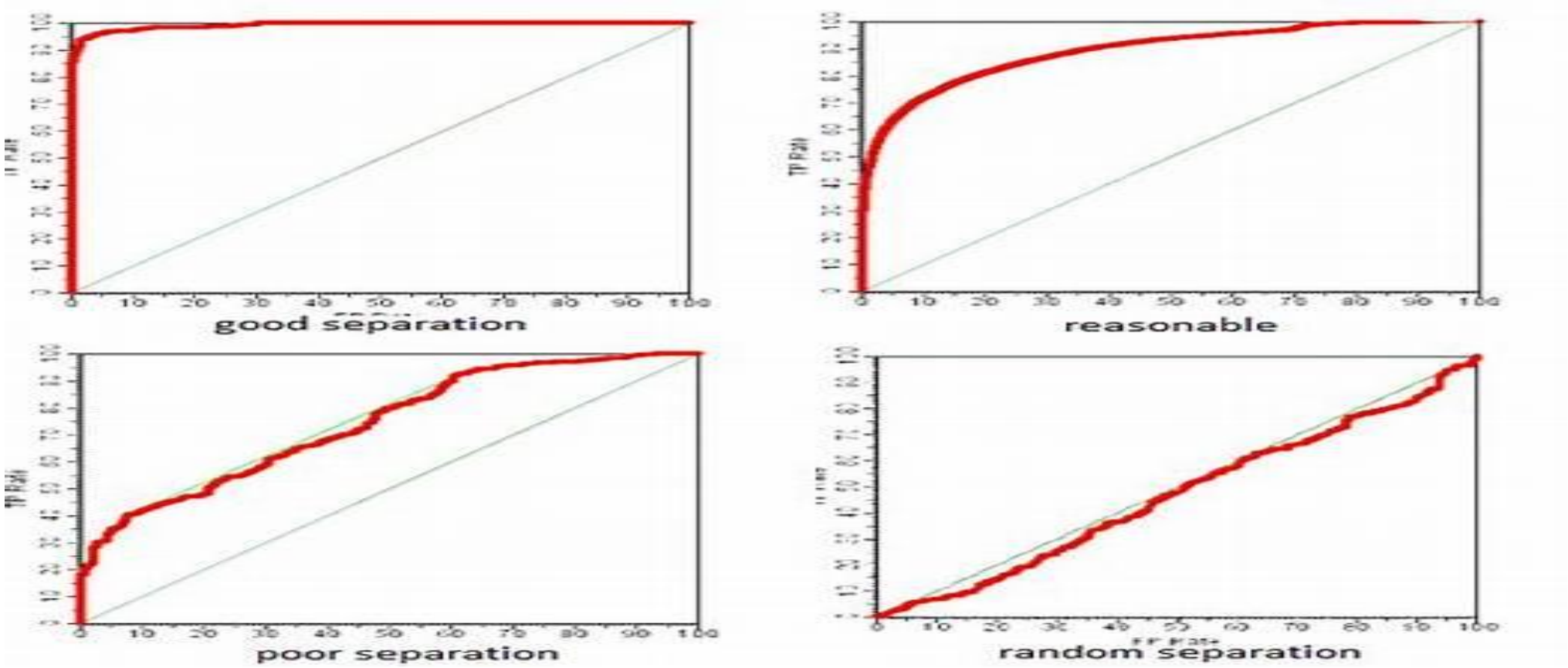
Roc منحنی
Receiver Operating Characteristic Curve



در نمودار ROCهر دوشاخص «حساسیت» (Sensitivity) یا «بازیابی» (Recall) ترکیب شده و به صورت یک منحنی نمایش داده می‌شوند.

اغلب برای بررسی کارایی الگوریتم‌های دسته‌بندی از منحنی ROC استفاده می‌کنند. این موضوع در شاخه یادگیری ماشین با نظارت بیشتر مورد توجه قرار گرفته است

مقایسه


good separation
reasonable
poor separation
random separation

Numpy & Pandas

برای دانلود و مشاهده کدها و اسلاید ها به لینک گیتهاب زیر مراجعه کنید:

mehrjy/Advanced-python-course (github.com)