Practice of AI

C2: Machine learning & Data analyze

谢文伟 (Jim Xie)

2020/3/6

Outline

1. Goal
2. ML workflow introduction
3. Basic math knowledge introduction
4. Time series forecasting demo
5. Brief summary

Goal

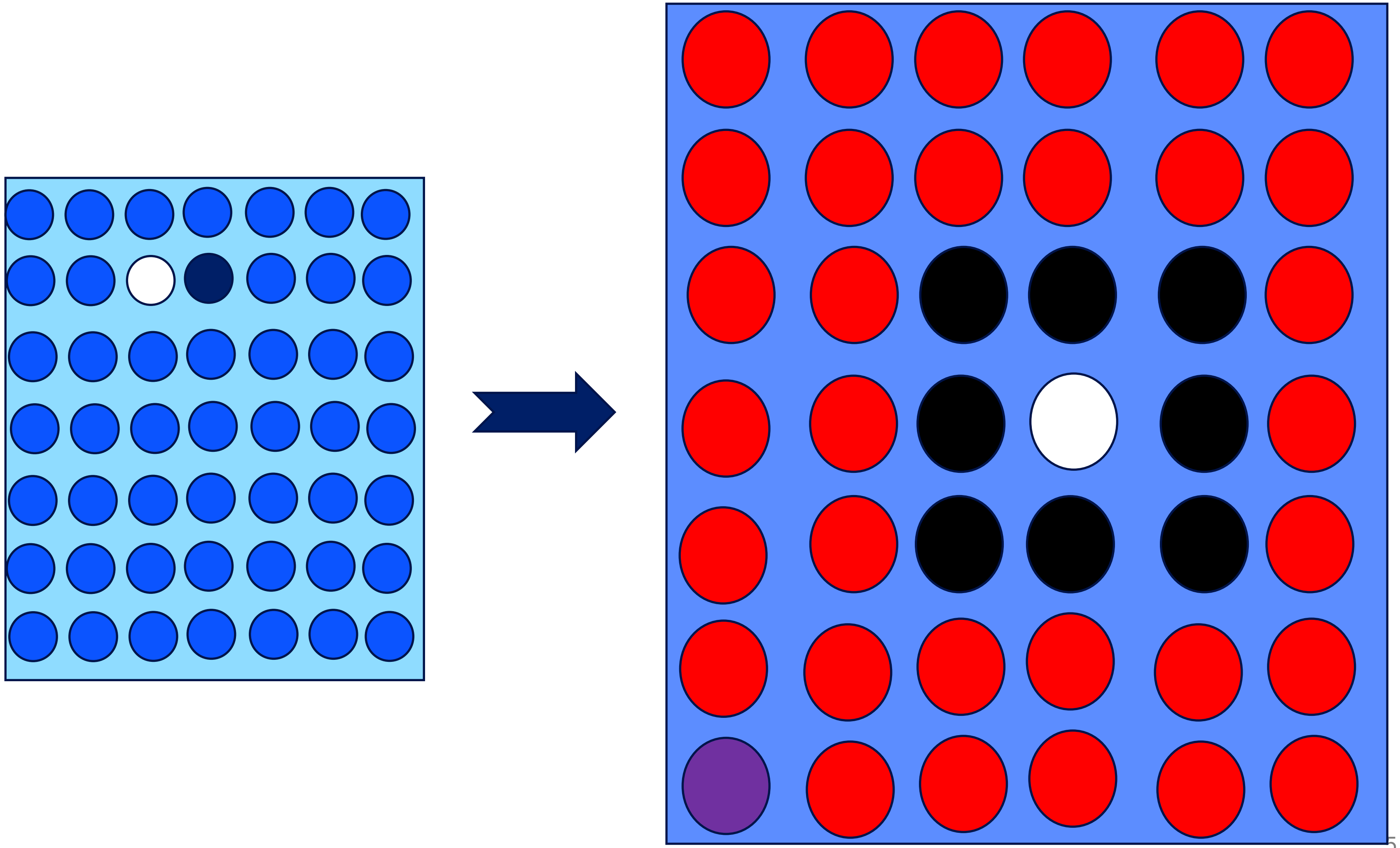
Getting start for data analyze with ML

Demo #1

Is ML Universal ?

Limitations

- NFL
- Smooth
- Boundary



Category by sample

Supervised learning

Unsupervised learning

Reinforcement learning

Q: 样本不平衡怎么办？

Purpose

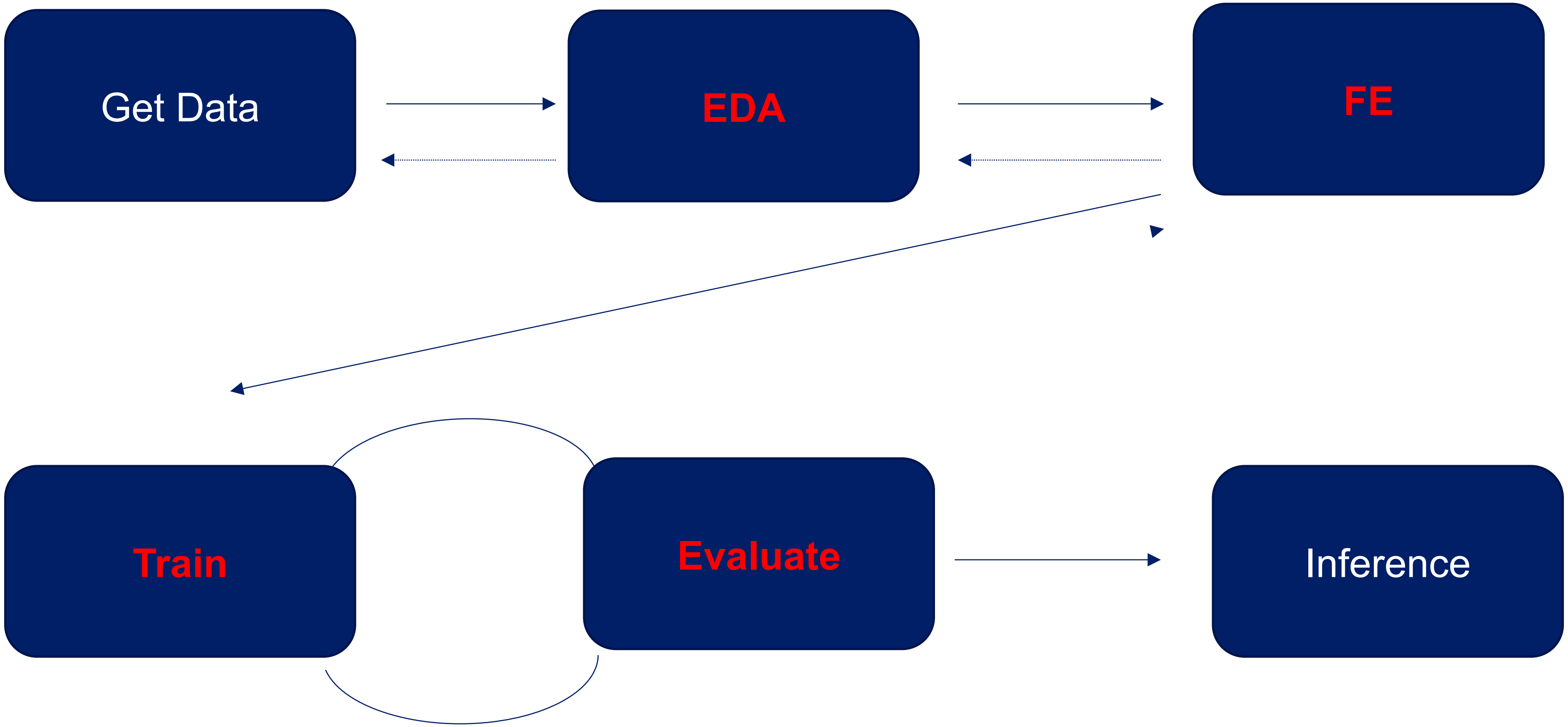
- **Class**

Cat or Dog?

- **Regression**

How much?

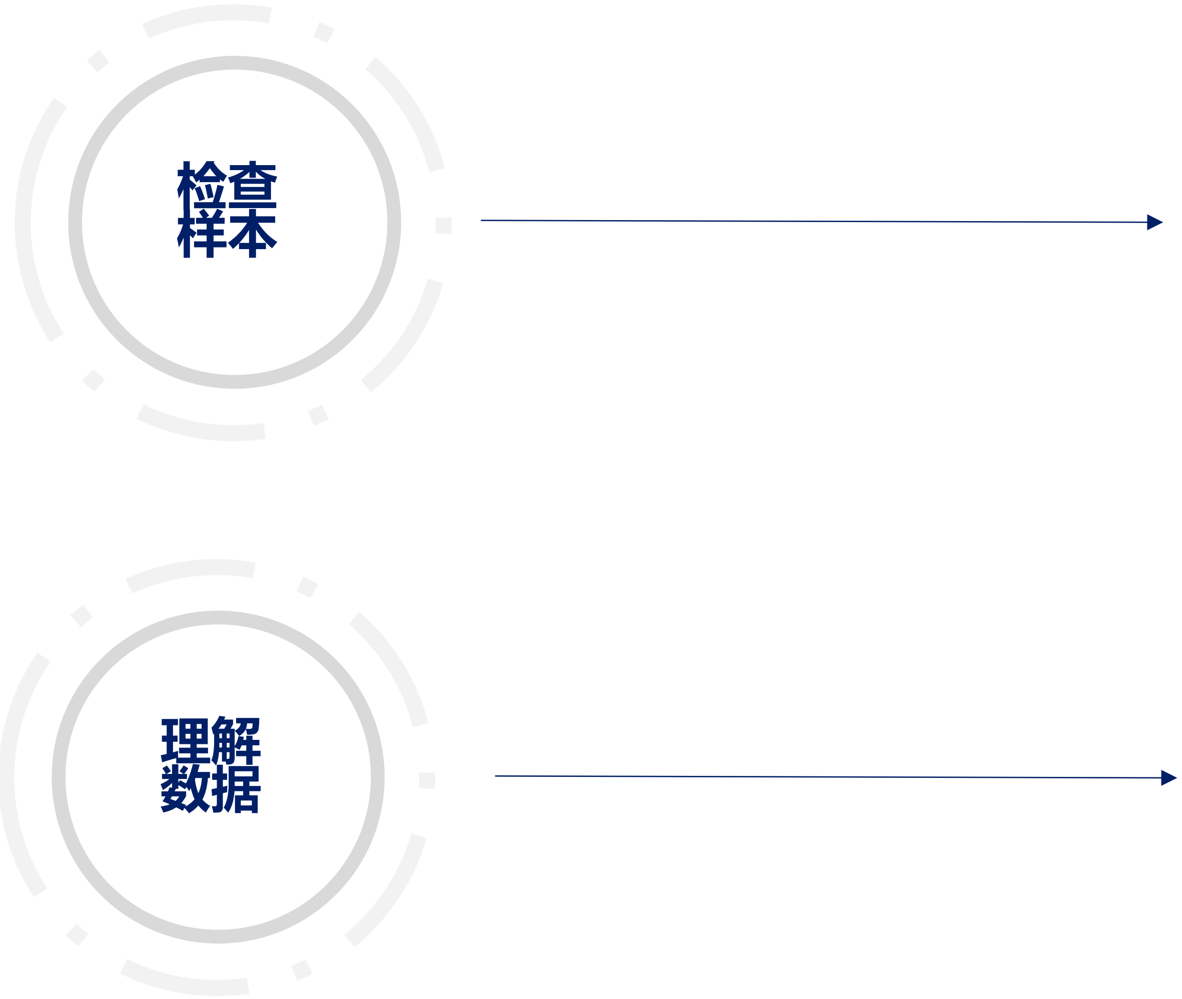
Workflow



EDA

EDA & preprocess

EDA #1



检查样本

- ➢ 检查样本是否合格？
- ➢ 样本量有多大？
- ➢ 有多少缺失数据？
- ➢ …………

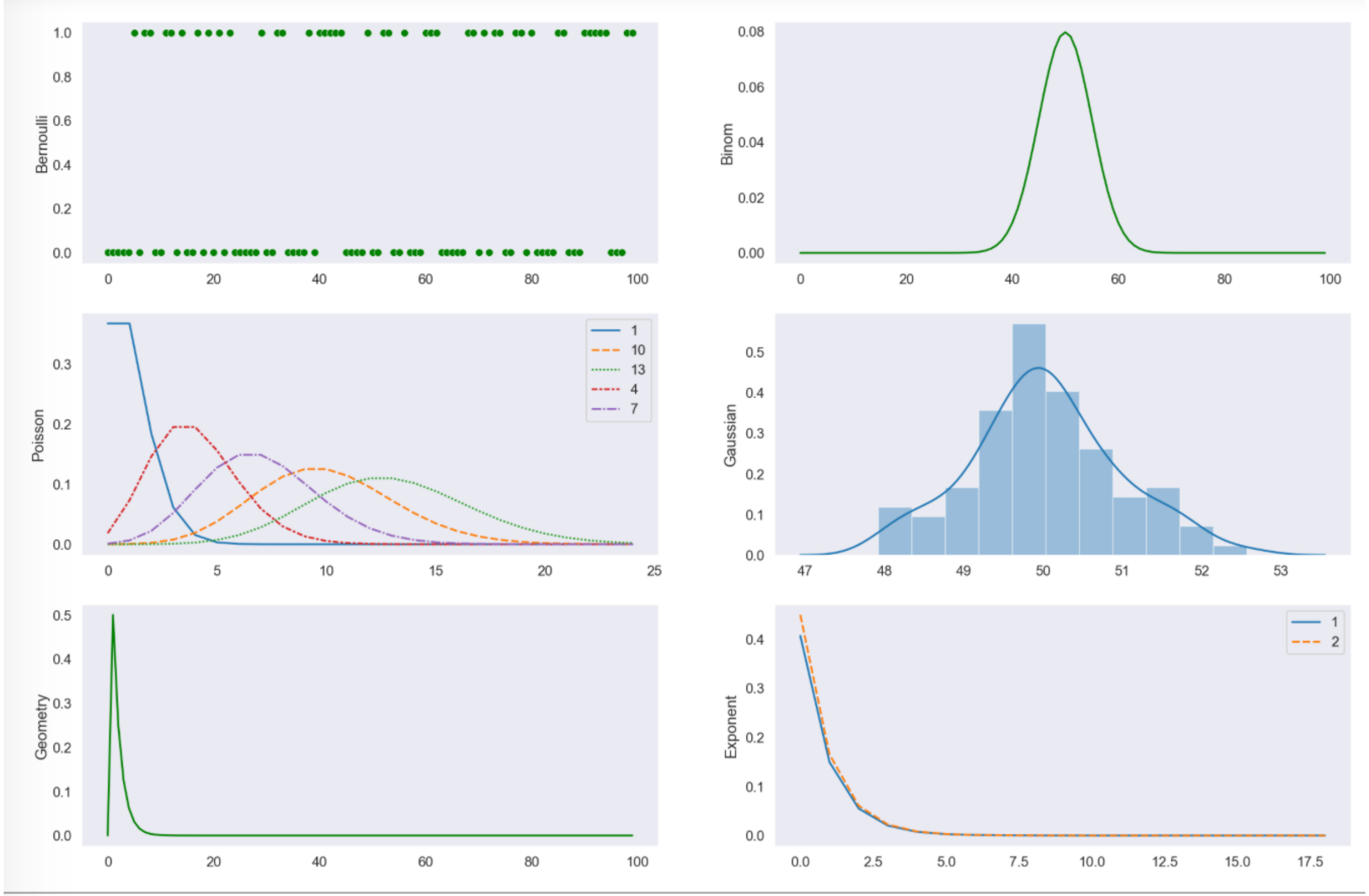
理解数据

- ➢ 回归 or 分类？
- ➢ 发现潜在的特征
- ➢ 数据如何分布？
- ➢ …………

EDA #2

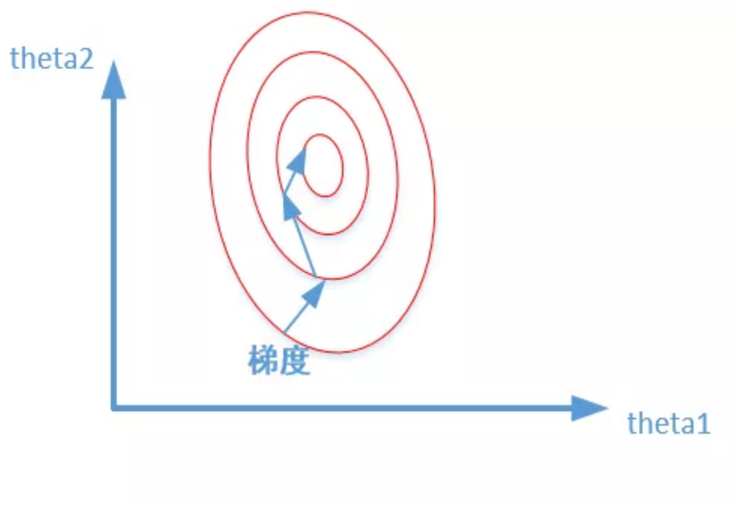
- Insight from graph
- Data distribution
- Normalization

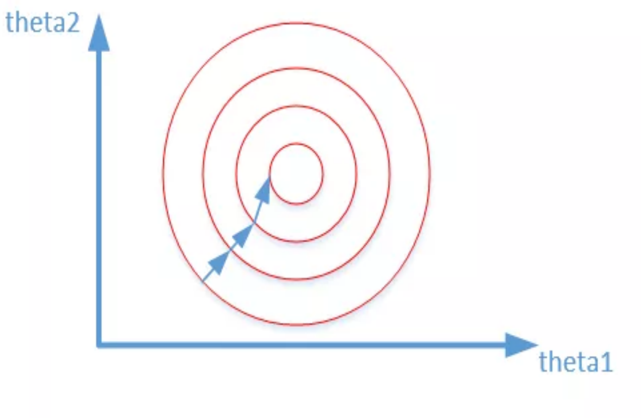
Data distribution



Normalization

- Why ?





- How ?

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

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

FE

Feature Engine

FE #1

Why

01

减少噪音

提高模型性能

02

减少维数

降低运算量

03

降低复杂度

增加可解释性

FE #2

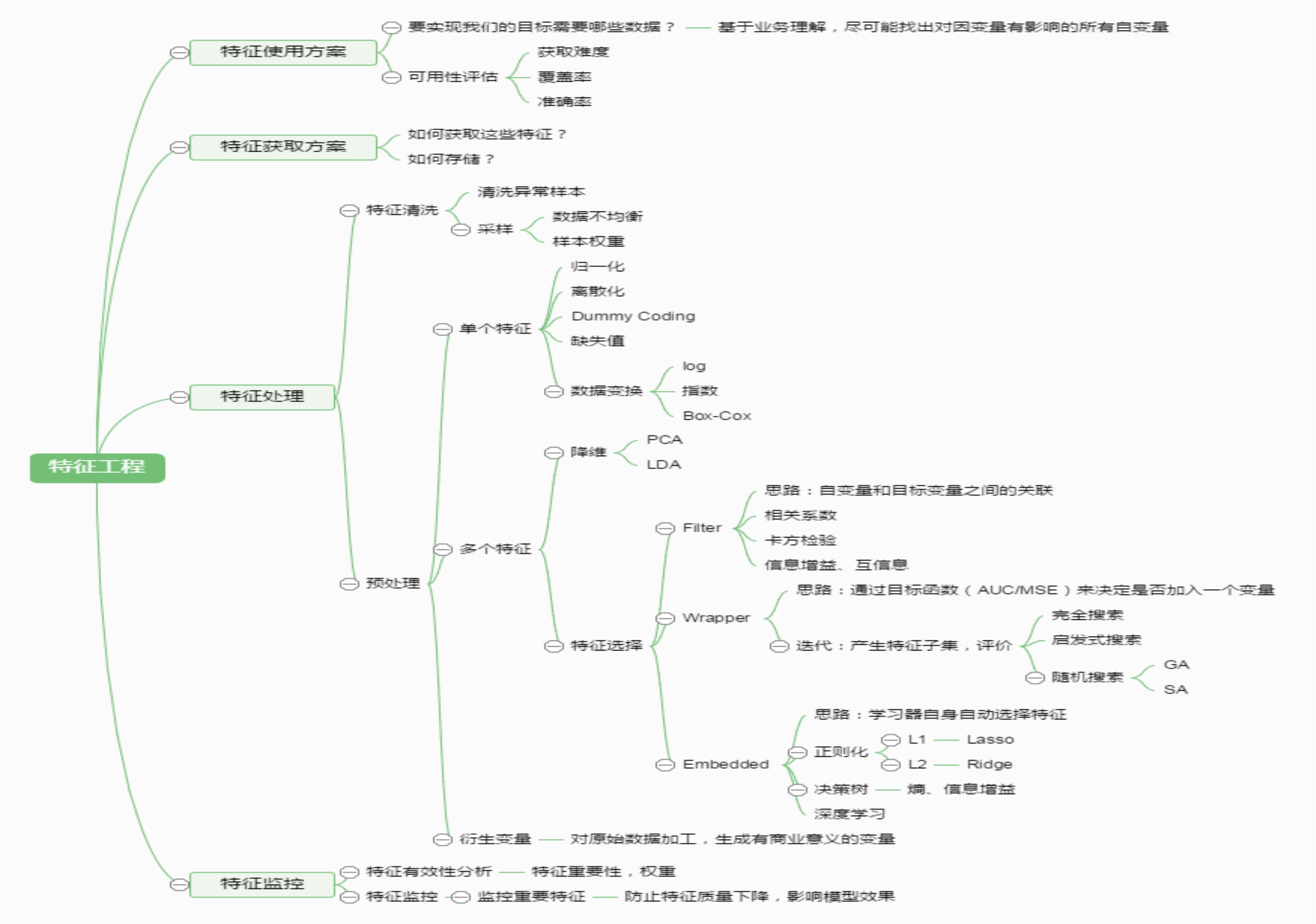
选取依据

- 特征是否发散？
- 特征和目标是否相关？

选取方法

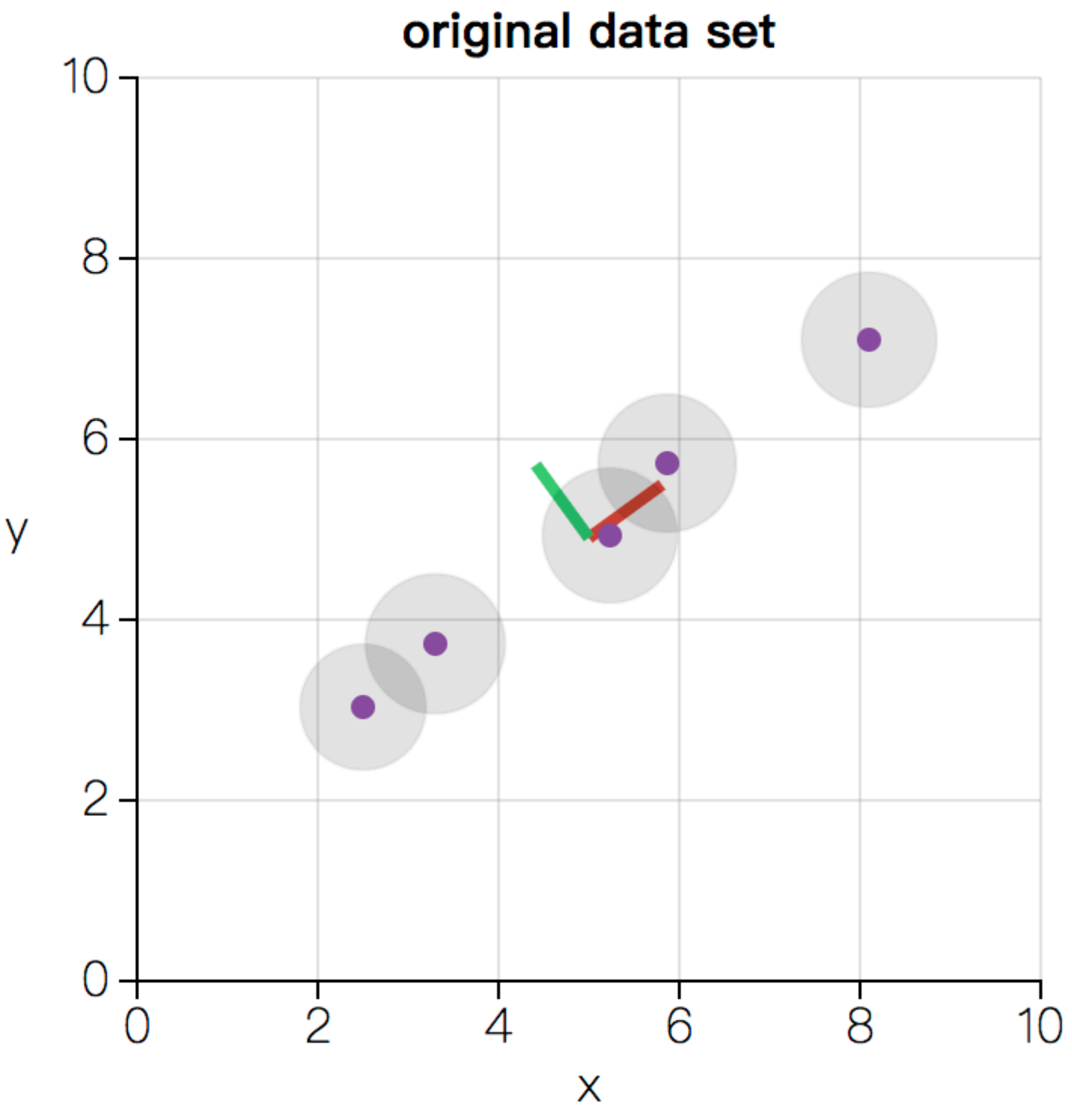
- Filter Methods
  - Correlation
  - ….
- Embed Methods
  - GA
  - ….
- Wrap methods
  - CNN
  - ….

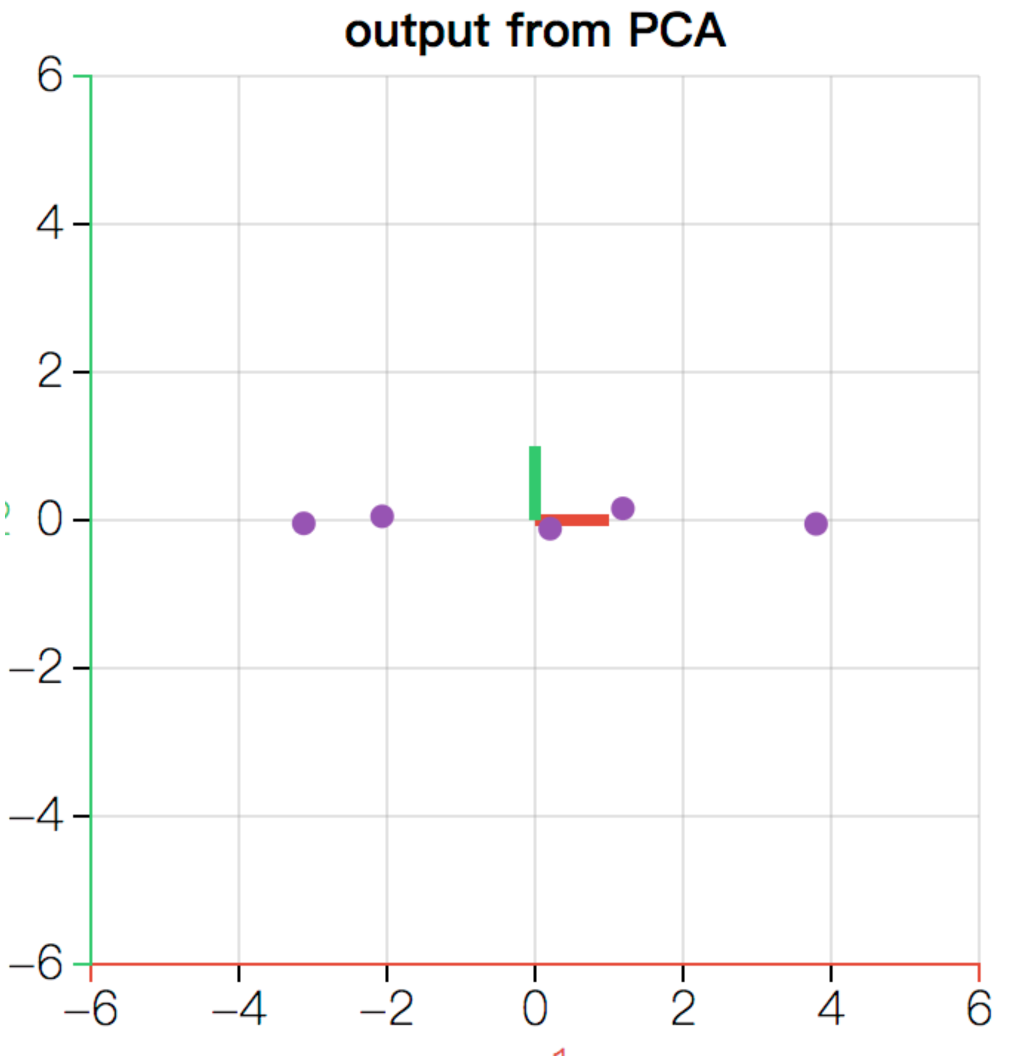
FE #3



- 特征工程
  - 特征使用方案
    - 要实现我们的目标需要哪些数据？ —— 基于业务理解，尽可能找出对因变量有影响的所有自变量
    - 可用性评估
      - 获取难度
      - 覆盖率
      - 准确率
  - 特征获取方案
    - 如何获取这些特征？
    - 如何存储？
  - 特征处理
    - 特征清洗
      - 清洗异常样本
      - 采样
        - 数据不均衡
        - 样本权重
    - 预处理
      - 单个特征
        - 归一化
        - 离散化
        - Dummy Coding
        - 缺失值
        - 数据变换
          - log
          - 指数
          - Box-Cox
      - 多个特征
        - 降维
          - PCA
          - LDA
        - 特征选择
          - Filter
            - 思路：自变量和目标变量之间的关联
            - 相关系数
            - 卡方检验
            - 信息增益、互信息
          - Wrapper
            - 思路：通过目标函数（AUC/MSE）来决定是否加入一个变量
            - 迭代：产生特征子集，评价
              - 完全搜索
              - 启发式搜索
              - 随机搜索
                - GA
                - SA
          - Embedded
            - 思路：学习器自身自动选择特征
            - 正则化
              - L1 —— Lasso
              - L2 —— Ridge
            - 决策树 —— 熵、信息增益
            - 深度学习
      - 衍生变量 —— 对原始数据加工，生成有商业意义的变量
  - 特征监控
    - 特征有效性分析 —— 特征重要性，权重
    - 特征监控 — 监控重要特征 —— 防止特征质量下降，影响模型效果

FE #4


original data set
10
8
6
4
2
0
y
0
2
4
6
8
10
x


output from PCA
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6

Train and Evaluation

Model and Evaluation

Models

- **Linear**
- **Neural network**
- DT/SVM/Bayes
- **XNN**/LSTM/GRU
- Boost/XGBoost
- ……..

Evaluate # Regression

Error = | Real – Predict | →

$$RMSE(X,h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(h(x_i) - y_i)^2}$$

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2$$

$$MAE(X,h) = \frac{1}{m}\sum_{i=1}^{m}|h(x_i) - y_i|$$

$$SD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - avg(x))^2}$$

Evaluate # Class

Confusion Matrix

	Predicted (Positive)	Predicted (Negative)
Actual (Positive)	TP	FN
Actual (Negative)	FP	TN

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

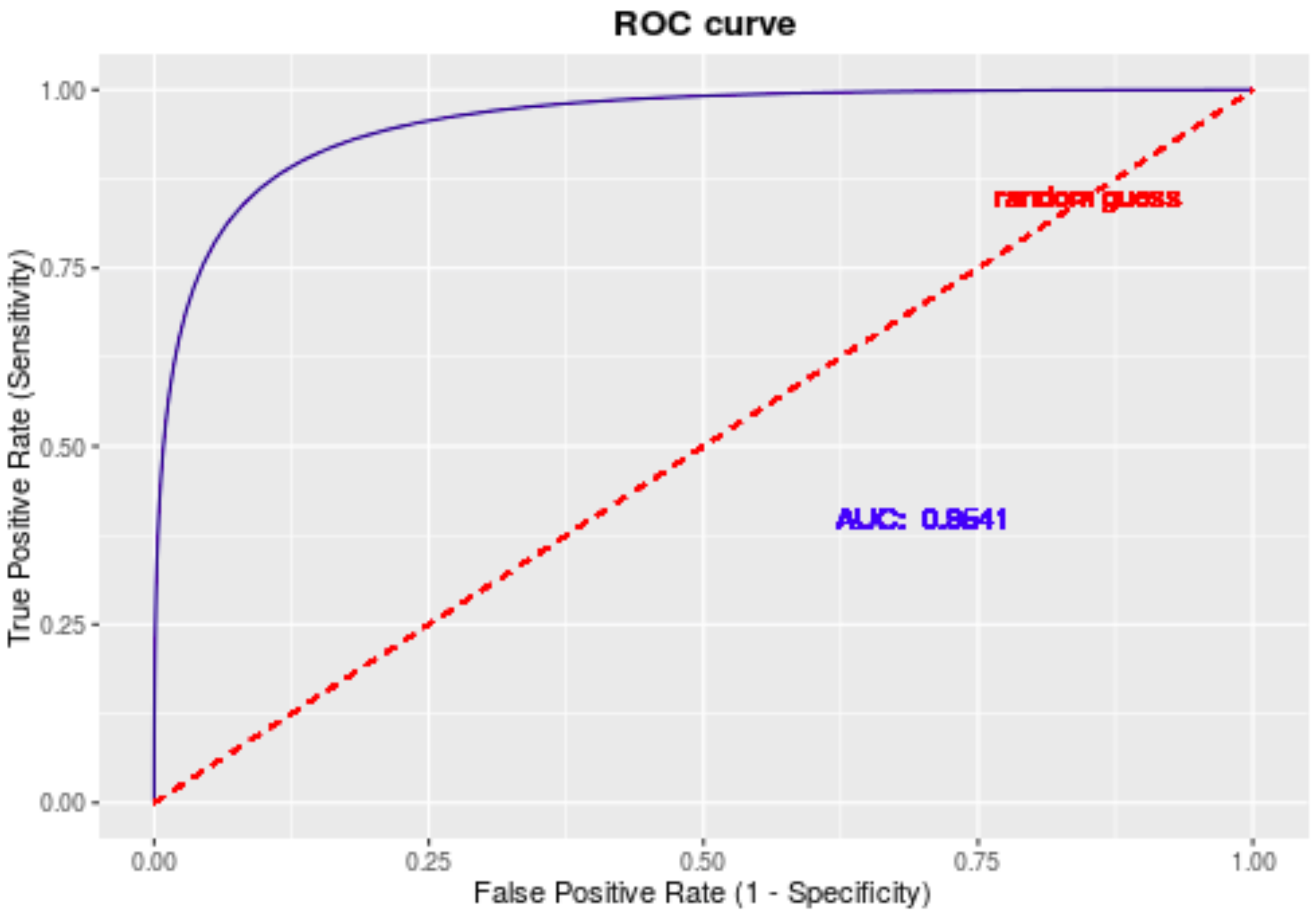
$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Sensitivity} = \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

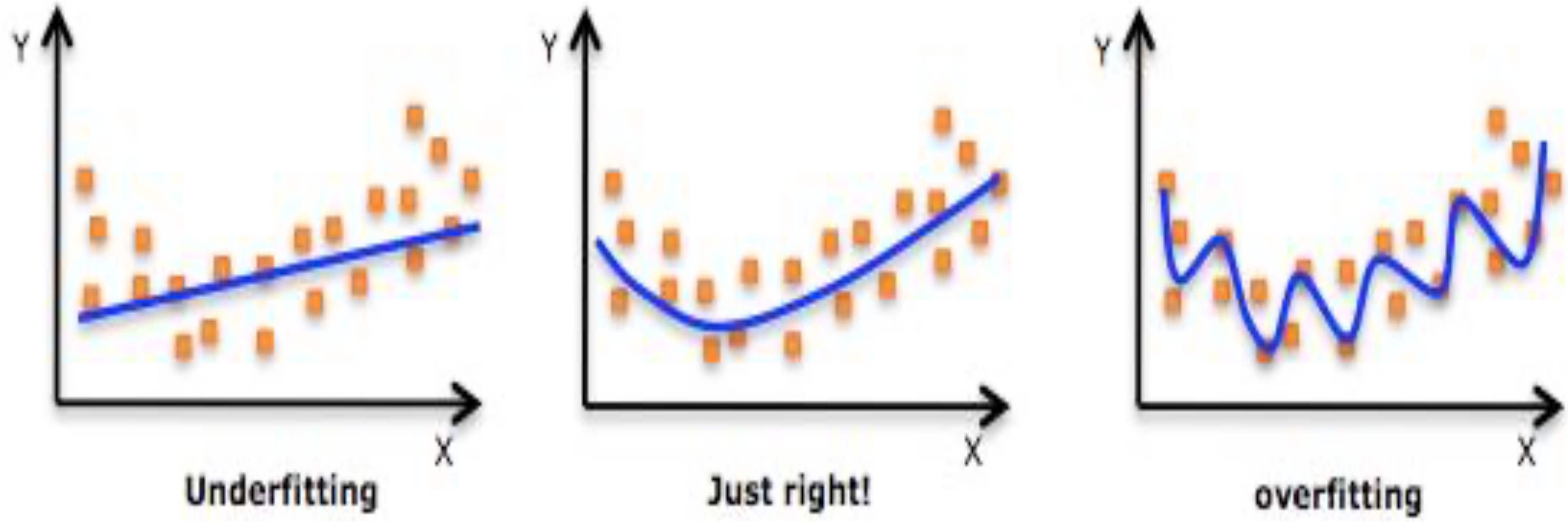
$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

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

ROC


ROC curve
1.00
0.75
0.50
0.25
0.00
True Positive Rate (Sensitivity)
0.00
0.25
0.50
0.75
1.00
False Positive Rate (1 - Specificity)
random guess
AUC: 0.8641

Underfitting/Overfitting



Q: How to do ?

Backlog

Backlog