

A decorative graphic in the top-left corner consisting of several blue-outlined hexagons of varying sizes arranged in a cluster.

Practice of AI

C2: Machine learning & Data analyze

Jim Xie

Outline



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

Goal



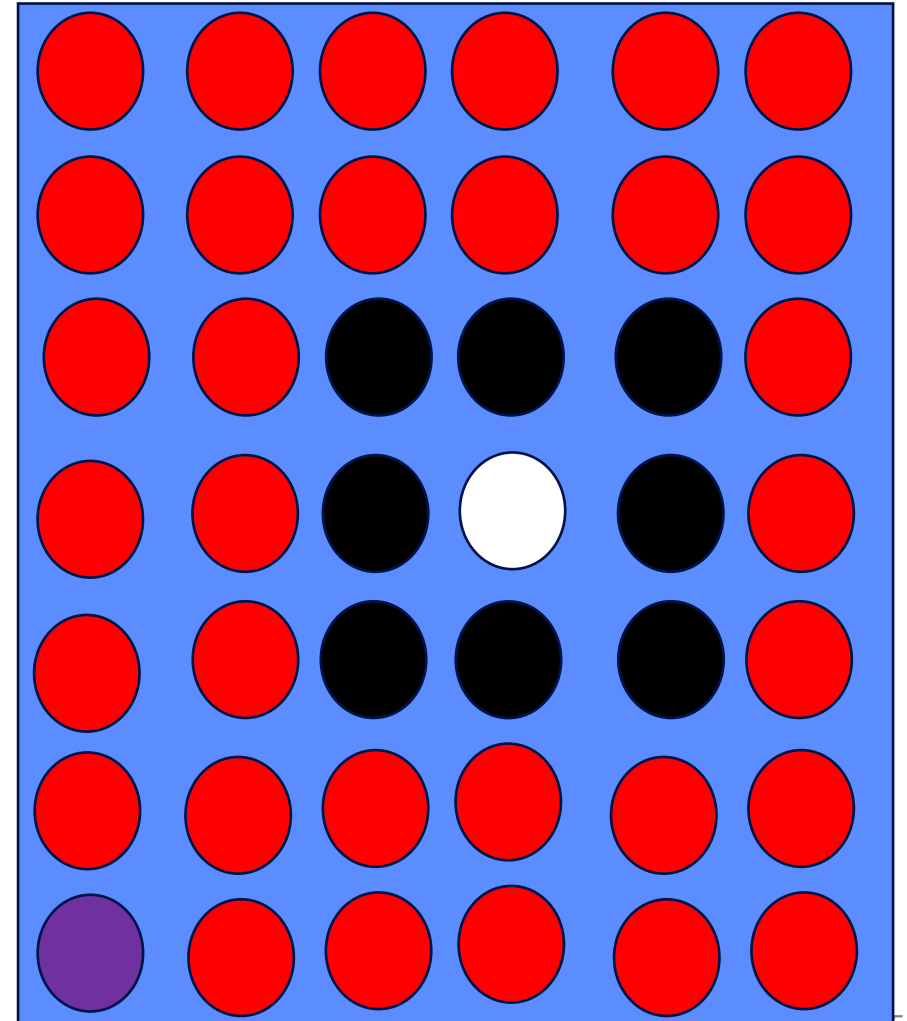
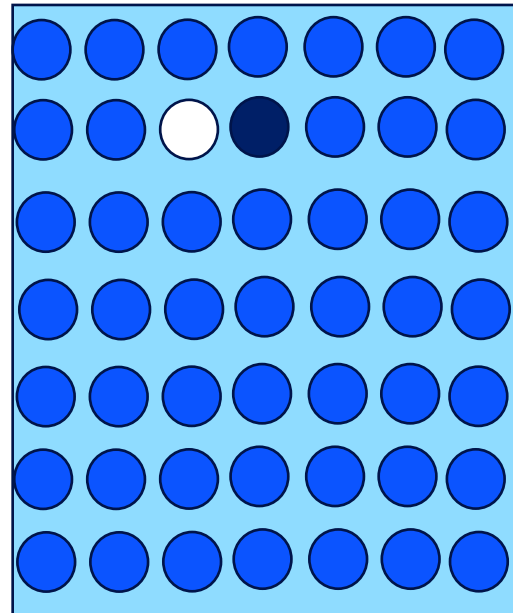
Getting start for data analyze with ML

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**



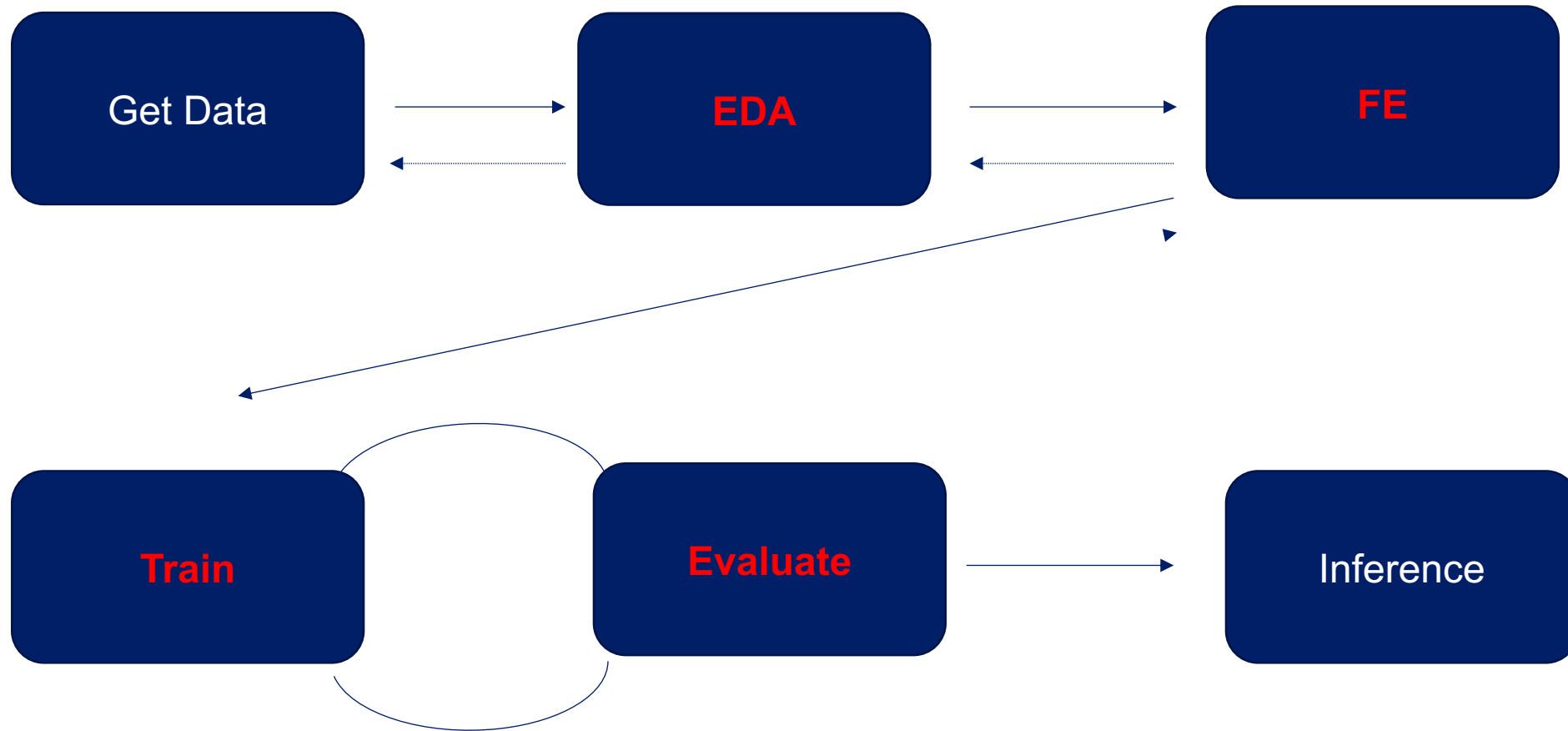
Cat or Dog?

- **Regression**



How much?

Workflow



EDA & preprocess

EDA #1



- 检查样本是否合格？
- 样本量有多大？
- 有多少缺失数据？
-



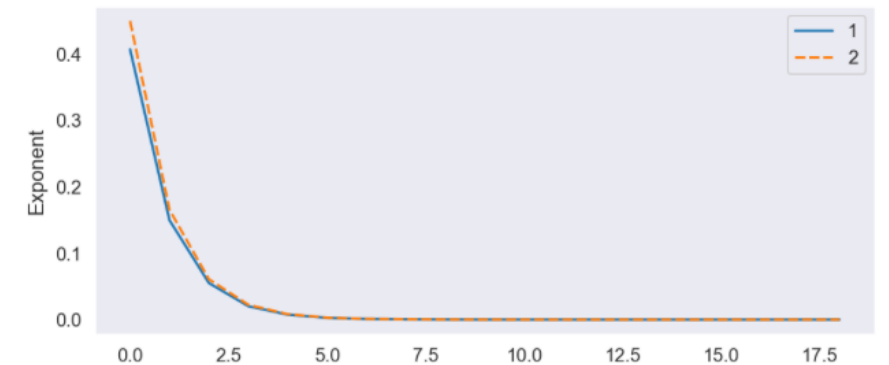
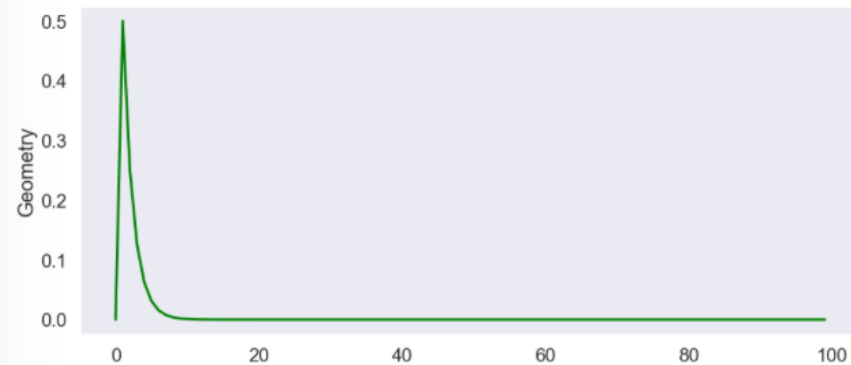
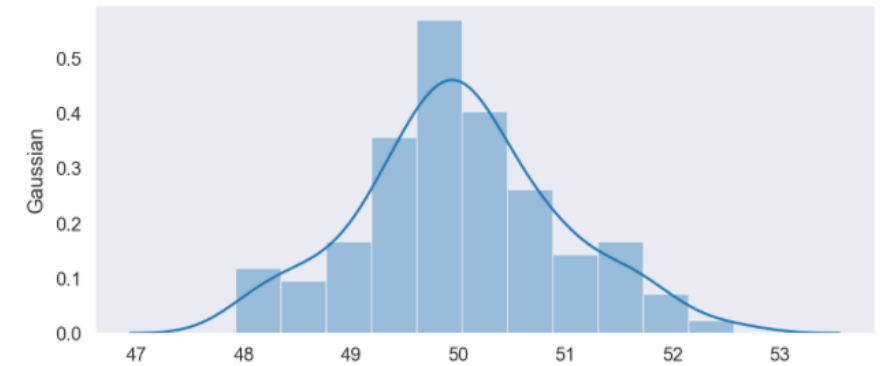
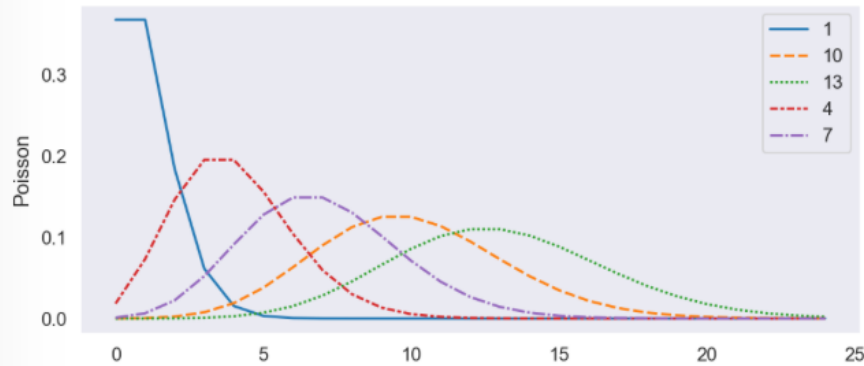
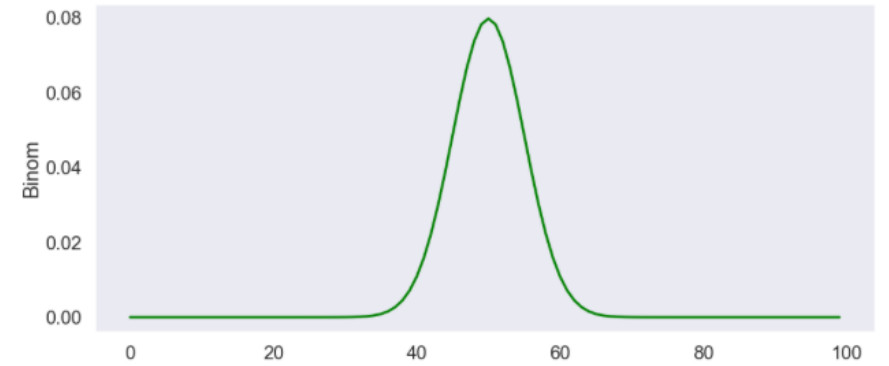
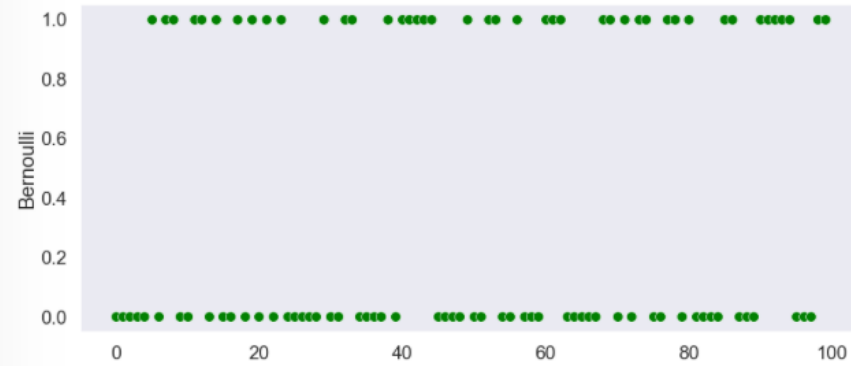
- 回归 or 分类？
- 发现潜在的特征
- 数据如何分布？
-

EDA #2



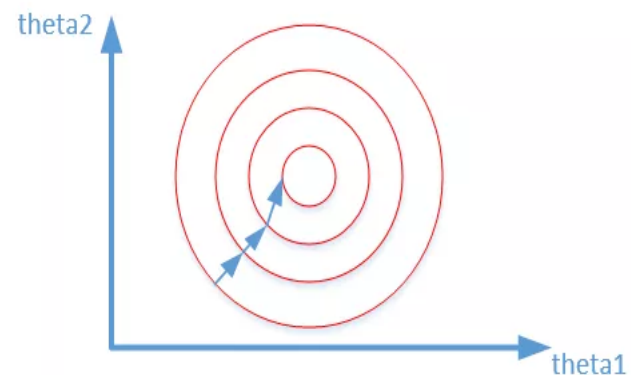
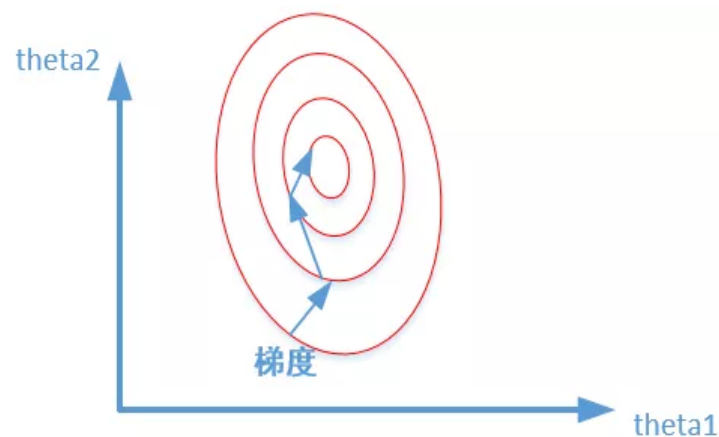
- Insight from graph
- Data distribution
- Normalization

Data distribution



Normalization

- Why ?



- How ?

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

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

FE

Feature Engine

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？

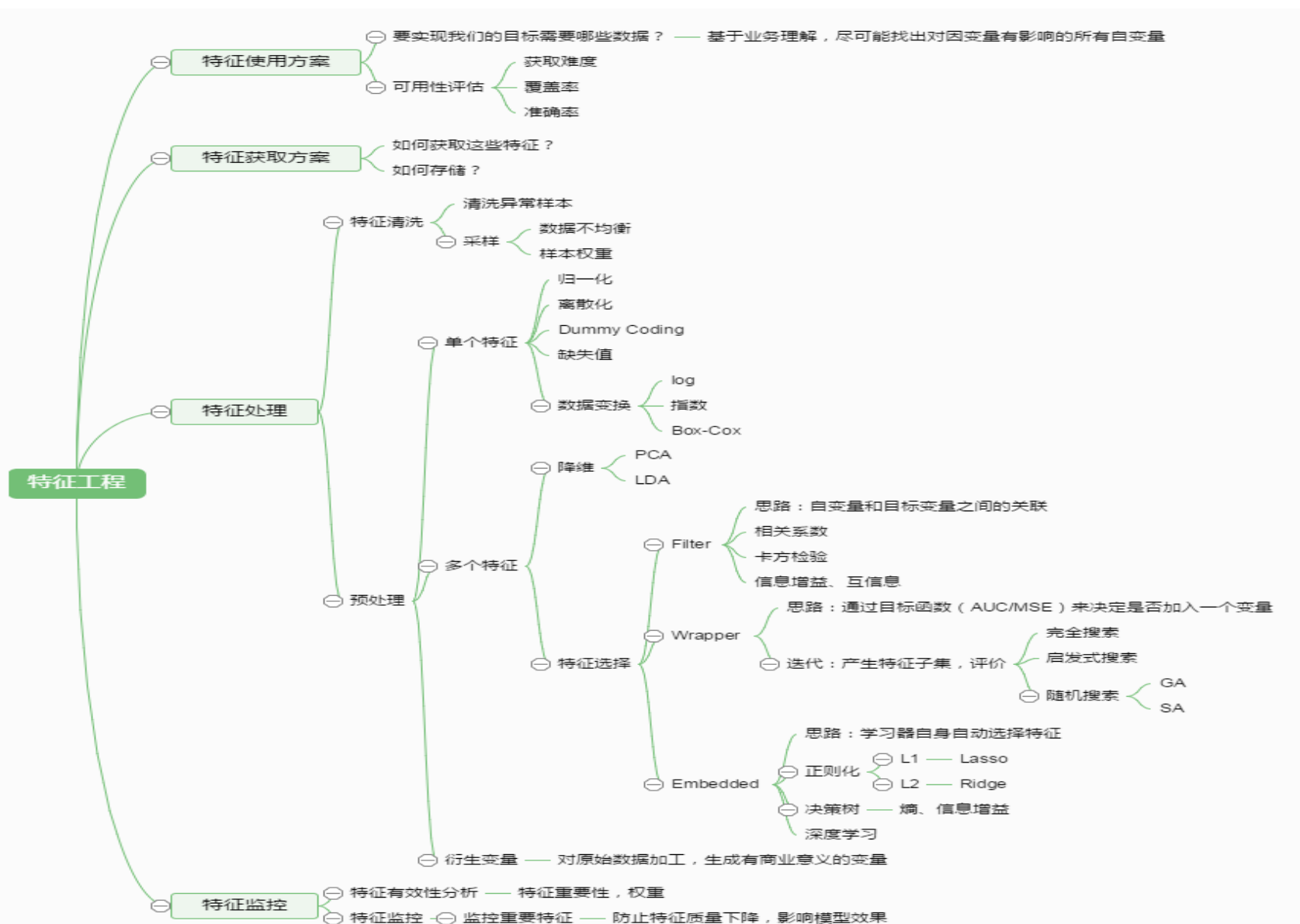


选取
方法



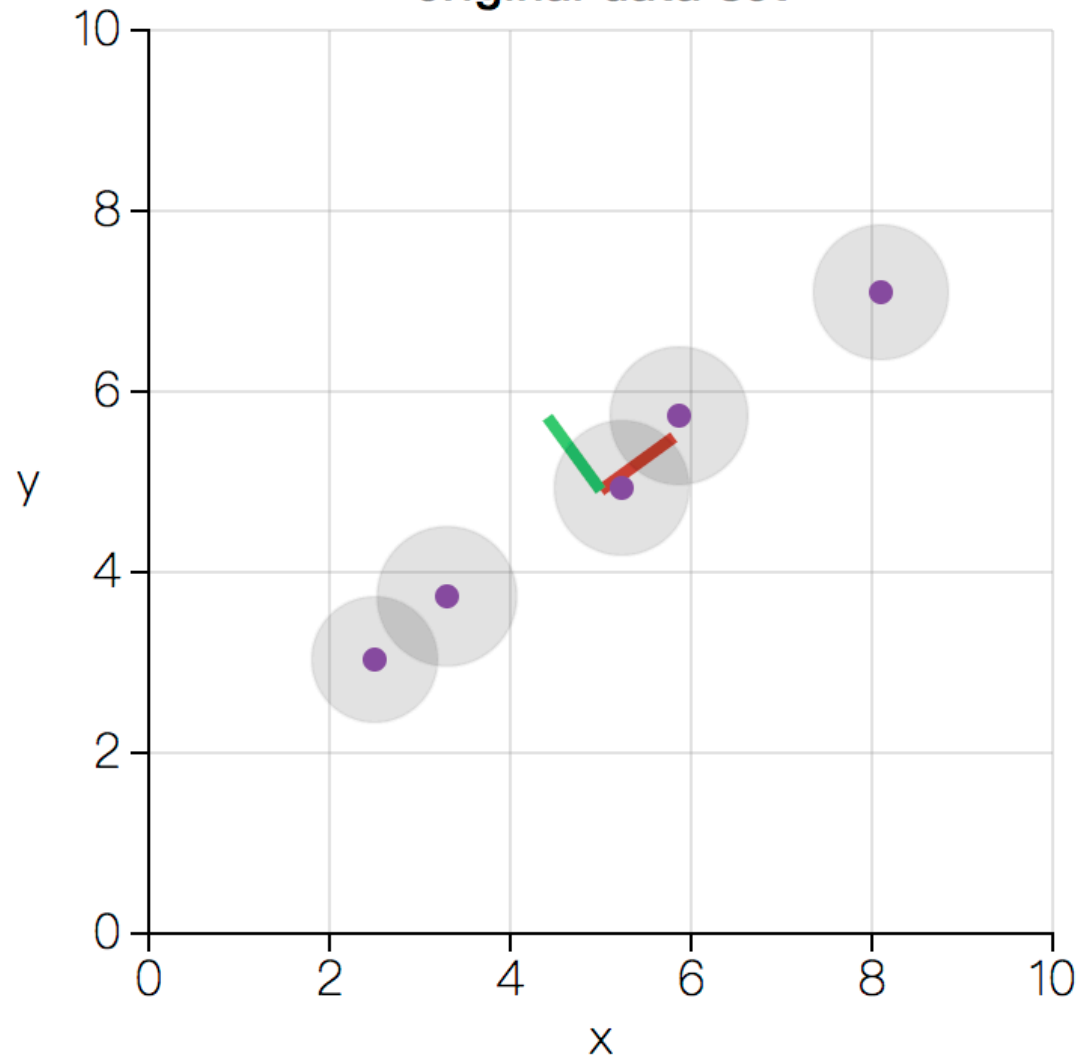
- Filter Methods
 - Correlation
 -
- Embed Methods
 - GA
 -
- Wrap methods
 - CNN
 -

FE #3

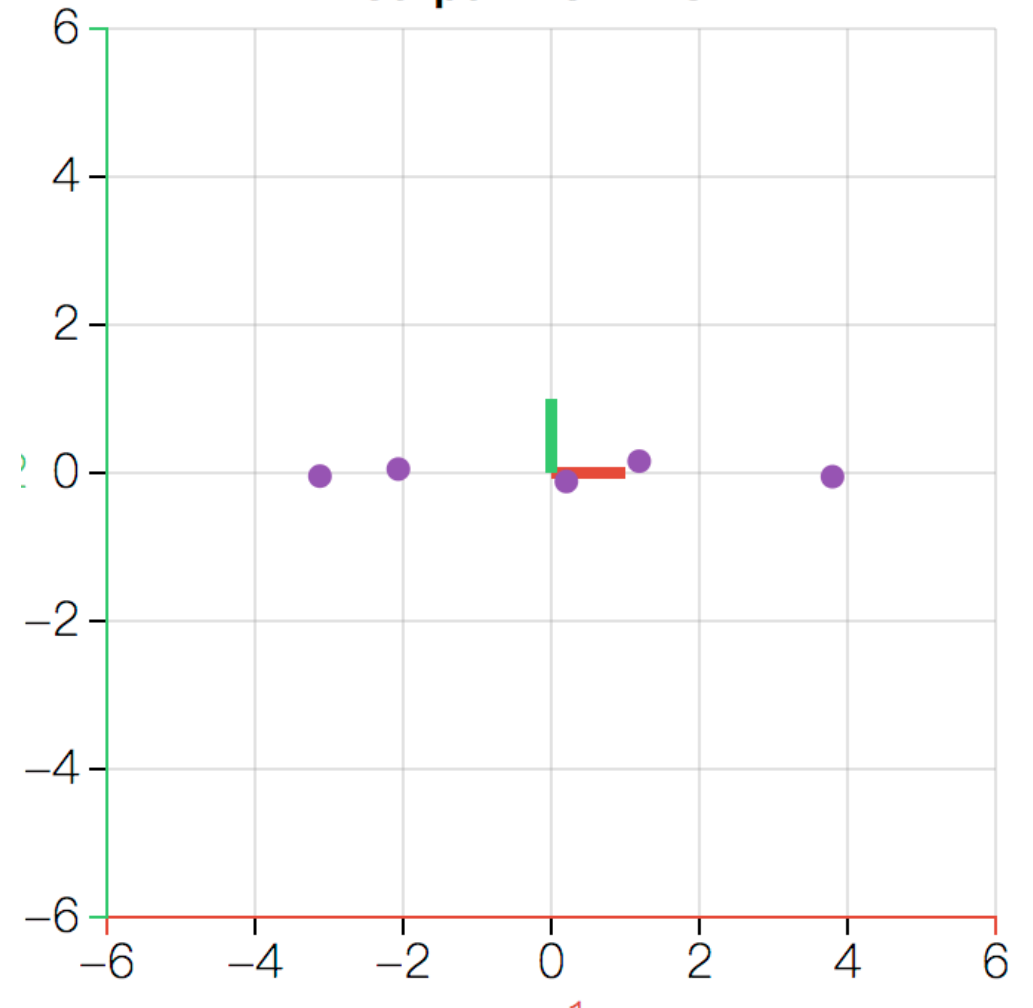


FE #4

original data set



output from PCA



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



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

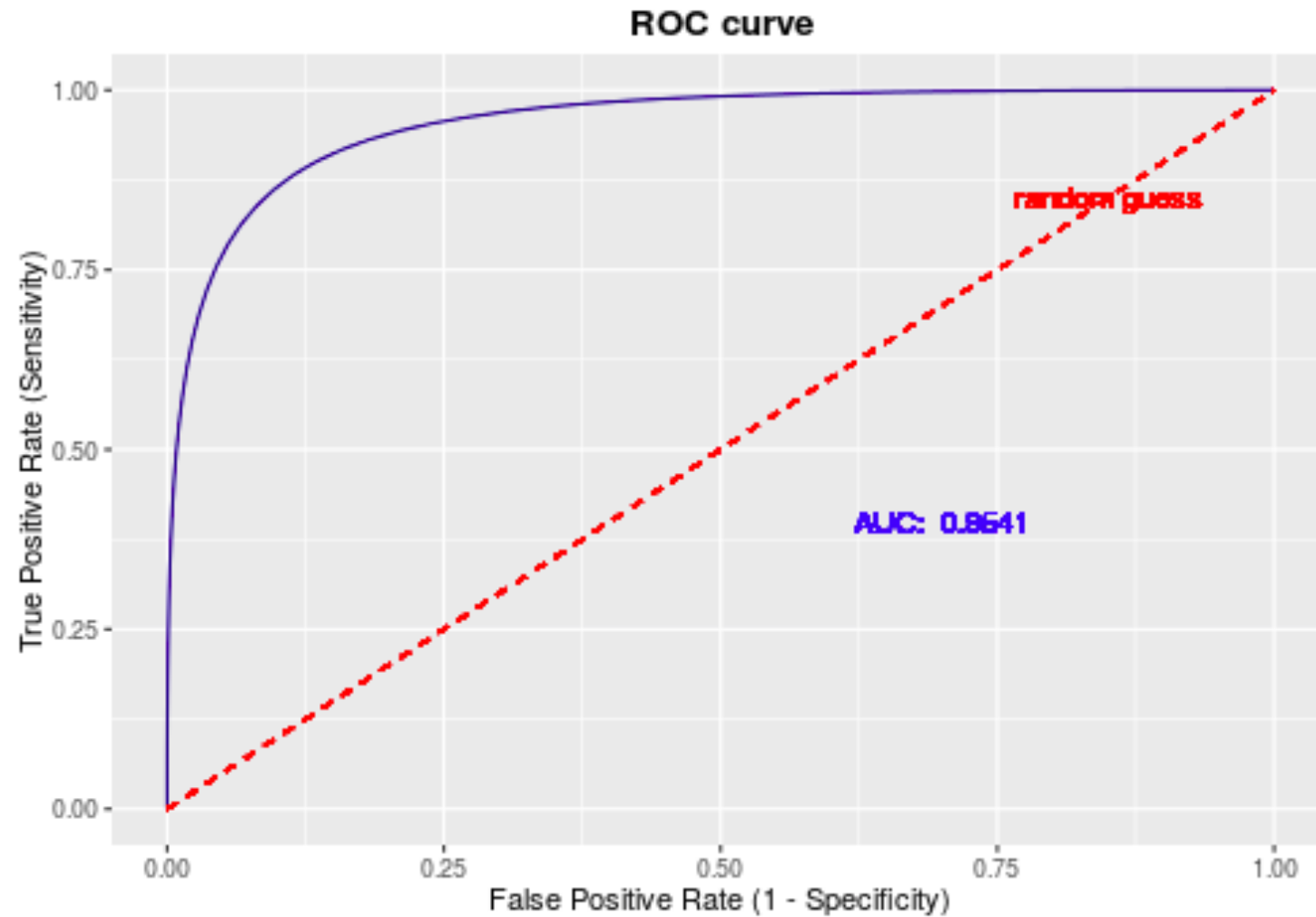
$$\text{Precision} = \frac{TP}{TP + FP}$$

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

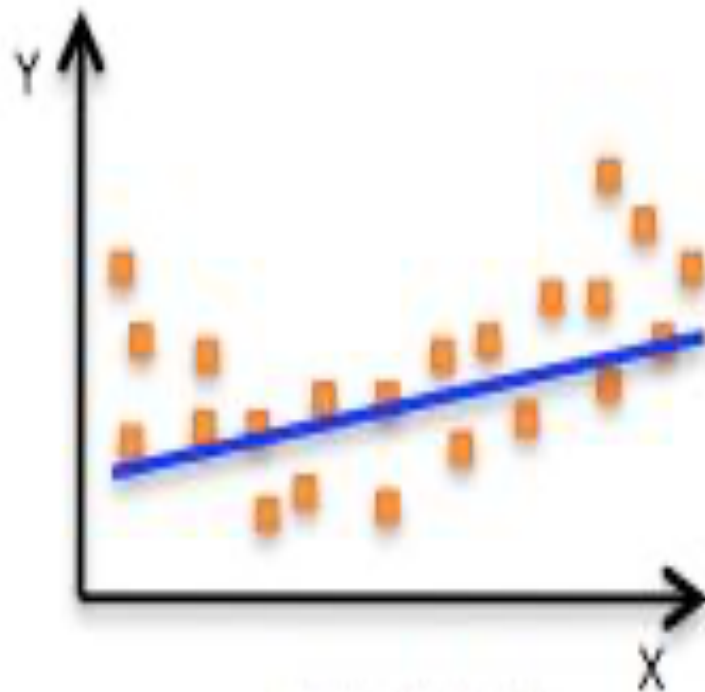
$$\text{Specificity} = \frac{TN}{TN + FP}$$

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

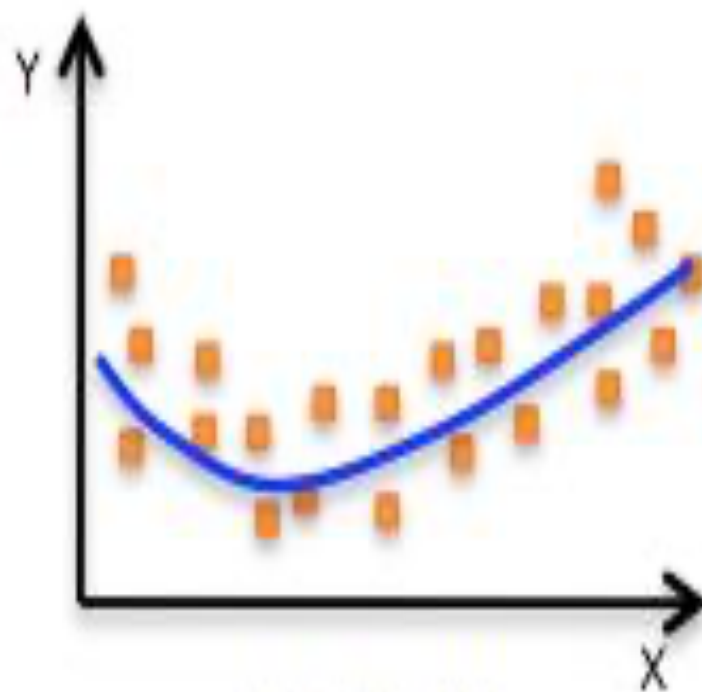
ROC



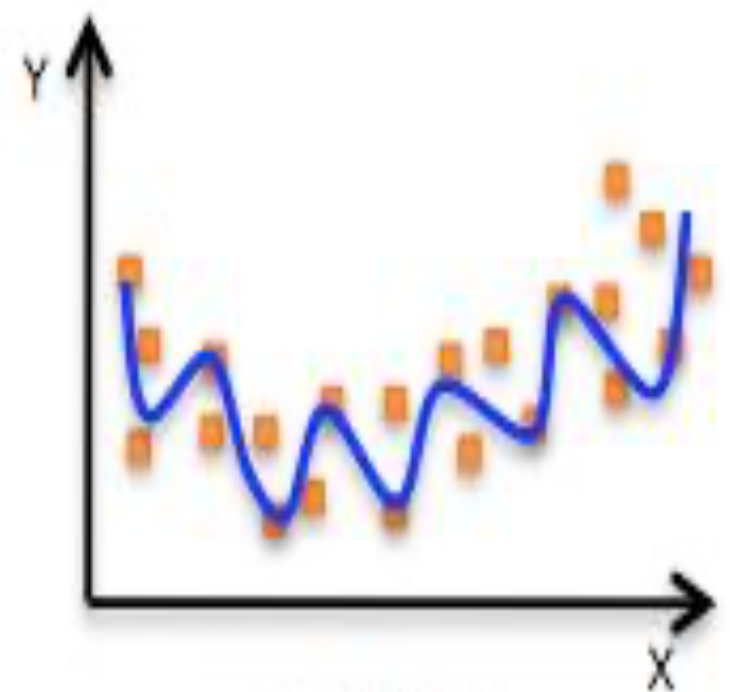
Underfitting/Overfitting



Underfitting



Just right!



overfitting

Q: How to do ?

Backlog

Backlog

