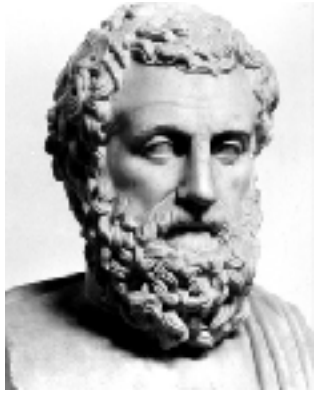
Artificial Intelligence For NLP Lesson-06



人工智能与自然语言处理课程组

2019.August.10

Outline

Machine Learning

- Background
- Main Methodologies
- Current Trends

Underfitting and Overfitting

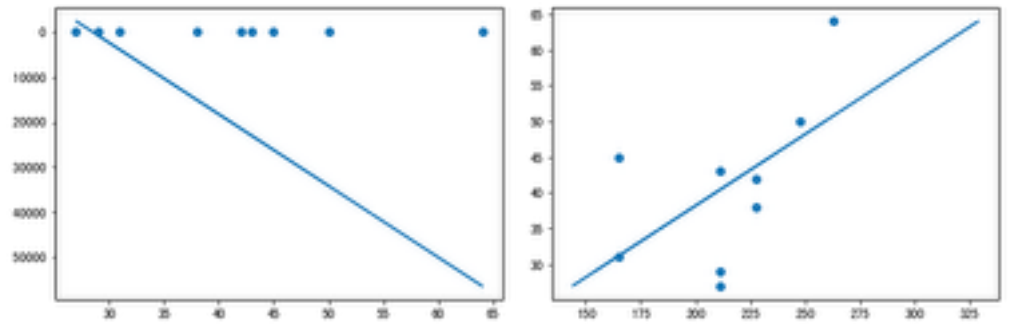
- Bias and Variance
- Model Capacity
- Underfitting and Overfitting

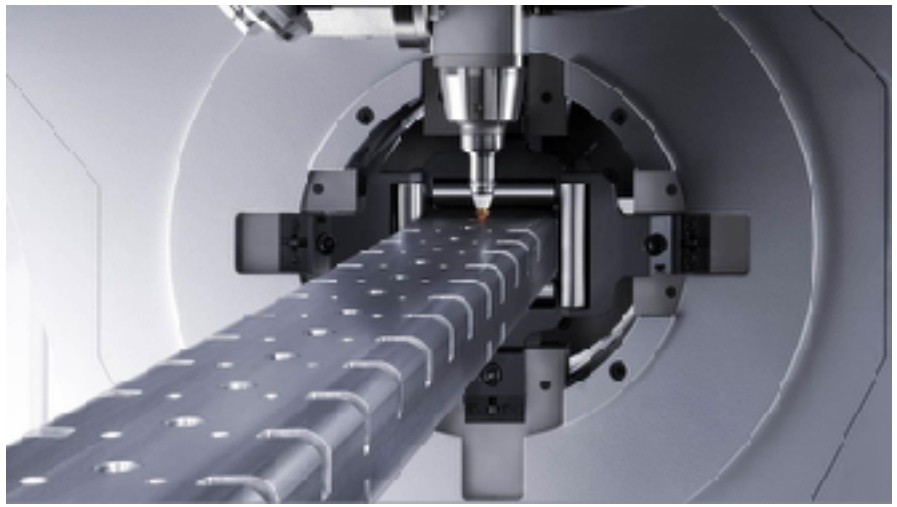
Train set, test set, validation set

- Influence of dataset
- The relation of train, test and validation.

Review

```
while True:
    if loss(y_true=fare, yhats=yhats) < eps: break

    indices = np.random.choice(range(len(age)), size=100

    sample_x = age[indices]
    sample_y = fare[indices]

    new_a, new_b = a, b

    for d in directions:
        da, db = d

        if min_loss != float('inf'):
            _a = a + da * min_loss * learning_rate
            _b = b + db * min_loss * learning_rate
        else:
            _a, _b = a + db, b + db

        y_hats = [model(i, _a, _b) for i in sample_x]
        l = loss(sample_y, np.array([model(x, a + da, b + db) for x in sample_x]))

        if l < min_loss:
            min_loss = l
            new_a, new_b = _a, _b

    if batch % 10 == 0:
        print('batch {}/ {} fare with {} * age + {}, with loss: {}'.format(batch, total, a, b, l))

    if batch > total: break

    batch += 1
```





Example Driven

Target

- More sales, More money.

Decision

- 1. Looking for a book
  - if there is holding a ceremony
  - how may days
  - female vs male
  - The closet college
  - ..

2. Target: Which college should I go to?

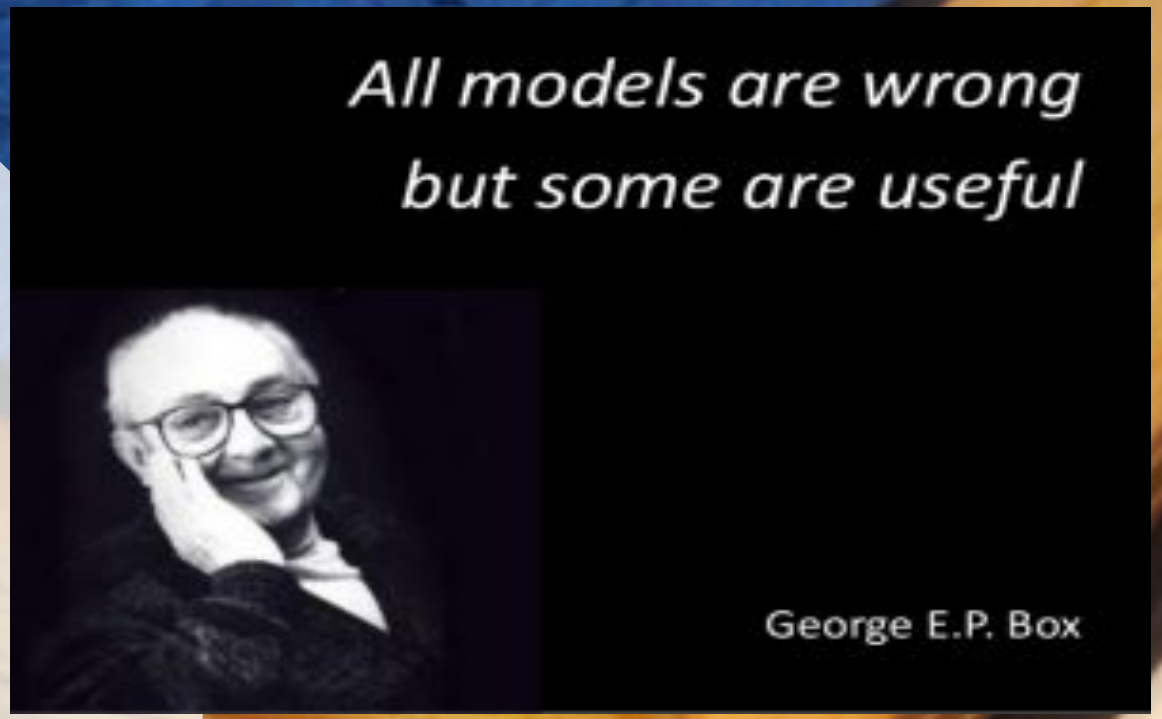
Decision

- 1. Looking for a book
  - if there is holding a ceremony
  - how may days
  - female vs male
  - The closet college
  - ..

Features

2. Target: Which college should I go to?

y



Model

The First Book

- Advantage?
- Disadvantage?

气温	沙尘暴	PM2.5	日期	买了多久	地点	学校
10	强	20-30	一月1日	1天	清华大学	北京大学
11	强	20-30	一月1日	1天	北京大学	北京大学
13	强	20-30	一月1日	1天	师范大学	师范大学
15	强	20-30	一月1日	1天	地质大学	地质大学
15	强	20-30	一月1日	1天	语言大学	语言大学
17	强	20-30	一月1日	1天	林业大学	林业大学
18	强	20-30	一月1日	1天	农业大学	农业大学
19	强	20-30	一月1日	1天	信息科技	信息科技
20	强	20-30	一月1日	1天	城市学院	城市学院
11	强	20-30	一月1日	2天	北京大学	北京大学
13	强	20-30	一月1日	2天	师范大学	师范大学
15	强	20-30	一月1日	2天	地质大学	地质大学
15	强	20-30	一月1日	2天	语言大学	语言大学
17	强	20-30	一月1日	2天	林业大学	林业大学
18	强	20-30	一月1日	2天	农业大学	农业大学
19	强	20-30	一月1日	2天	信息科技	信息科技
20	强	20-30	一月1日	2天	城市学院	城市学院

The second book

			下雨	最后的学校
		幸福明显变少	下下雨	北京师范大学
			下雨	
	时间小于7天	幸福感有点变少		
无运动			XXXXX	北京大学
			XXXXX	北京体育大学
			XXXXX	
		XXXXX	XXXXX	
			XXXXX	清华大学
			XXXXX	
			XXXXX	
			XXXXX	
			XXXXX	
			XXXXX	
	时间大于7天	XXXXX	XXXXX	
			XXXXX	
			XXXXX	
			XXXXX	
	时间小于7天			
有运动				

The First Book

- Advantage?
  - Quick Inference
  - Easy Computing
  - ….
- Disadvantage?
  - Heavy
  - …

K-nearest neighbors

气温	沙尘暴	PM2.5	日期	买了多久	地点	学校
10	强	20-30	一月1日	1天	清华大学	北京大学
11	强	20-30	一月1日	1天	北京大学	北京大学
13	强	20-30	一月1日	1天	师范大学	师范大学
15	强	20-30	一月1日	1天	地质大学	地质大学
15	强	20-30	一月1日	1天	语言大学	语言大学
17	强	20-30	一月1日	1天	林业大学	林业大学
18	强	20-30	一月1日	1天	农业大学	农业大学
19	强	20-30	一月1日	1天	信息科技	信息科技
20	强	20-30	一月1日	1天	城市学院	城市学院
11	强	20-30	一月1日	2天	北京大学	北京大学
13	强	20-30	一月1日	2天	师范大学	师范大学
15	强	20-30	一月1日	2天	地质大学	地质大学
15	强	20-30	一月1日	2天	语言大学	语言大学
17	强	20-30	一月1日	2天	林业大学	林业大学
18	强	20-30	一月1日	2天	农业大学	农业大学
19	强	20-30	一月1日	2天	信息科技	信息科技
20	强	20-30	一月1日	2天	城市学院	城市学院

The second book

Decision Tree

The third book

expected value

$$-\frac{1}{2}\cdot 1+\frac{1}{4}\cdot 2+\frac{1}{4}\cdot 3$$

平均薪资入	时间3天	时间5天	时间7天	下雨	晴天	沙城暴	在中资材	在元大部退址
北京大学	400	250	200	468	352	154	1014	925
北京城市学院	500	200	1165	316	877	138	925	372
中国音乐学院	350	300	386	1022	450	226	317	543
中国地质大学	400	400	388	1054	1069	179	378	925
北京信息科技大学	500	500	1022	638	1023	127	925	1014
北京师范大学	800	800	827	313	1078	200	317	1156
北京林业大学	800	700	939	1042	499	266	997	543
目前的学校	1200	1100	818	792	854	163	934	317

The third book

- P（U|A1, A2, A3）
  - ~ P(A1|U) * P(A2|U) * P(A3|U)

- Naïve Bayesian Classification

$$\text{Posterior}=\frac{\text{Likelihood}*\text{Prior}}{\text{Evidence}}$$

$$P(C_j \mid A_1, A_2, \ldots, A_n)=\frac{\left(\prod_{i=1}^{n} P(A_i \mid C_j)\right) P(C_j)}{P(A_1, A_2, \ldots, A_n)}$$

平均预收入	时间3天	时间5天	时间7天	下雨	晴天	沙城暴	在中资州	在三大都市
北京大学	400	250	200	468	352	154	1014	925
北京航空学院	500	200	1165	316	877	138	925	377
中国传媒学院	350	300	386	1022	450	226	317	543
中医药大学	400	400	388	1054	1069	179	378	925
北京邮电科技大学	500	500	1022	638	1023	127	925	1014
北京师范大学	800	800	827	313	1078	200	317	1156
北京体育大学	800	700	959	1042	499	266	987	543
目前的学校	1200	1100	818	792	854	163	934	317

The Forth book

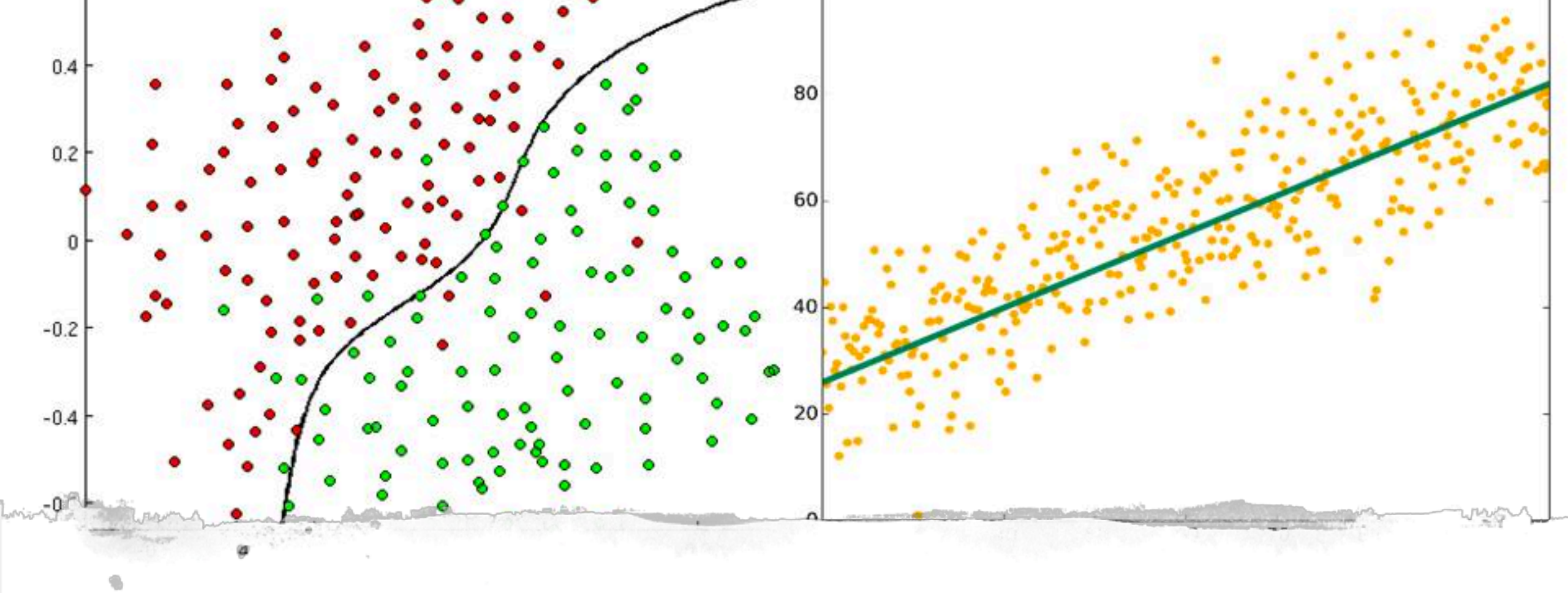
- 食堂打饭 不标价钱， 问， 每种菜多少钱?
-
- 第一次 茄子*2 + 西瓜*1 + 馒头 * 1 : 7.8元
- 第二次 茄子*1 + 西瓜*1 + 馒头 * 2 : 6.5元
- 第三次 玉米*1 + 菠菜*2 + 馒头 * 1: 5.6元
- ...

- 每种价格?

- Neural Network

Chat

- Determine if is *a valuable customer* in Wechat?
- What Features do we need?
- How to predicate it?



Two Type

- 1. Classification
- 2. Regression

How to evaluate?

- Accuracy
- Precision
- Recall
- AOC/AUC
- F1_score, F2_score
- MSE
- Loss Function

- Generation

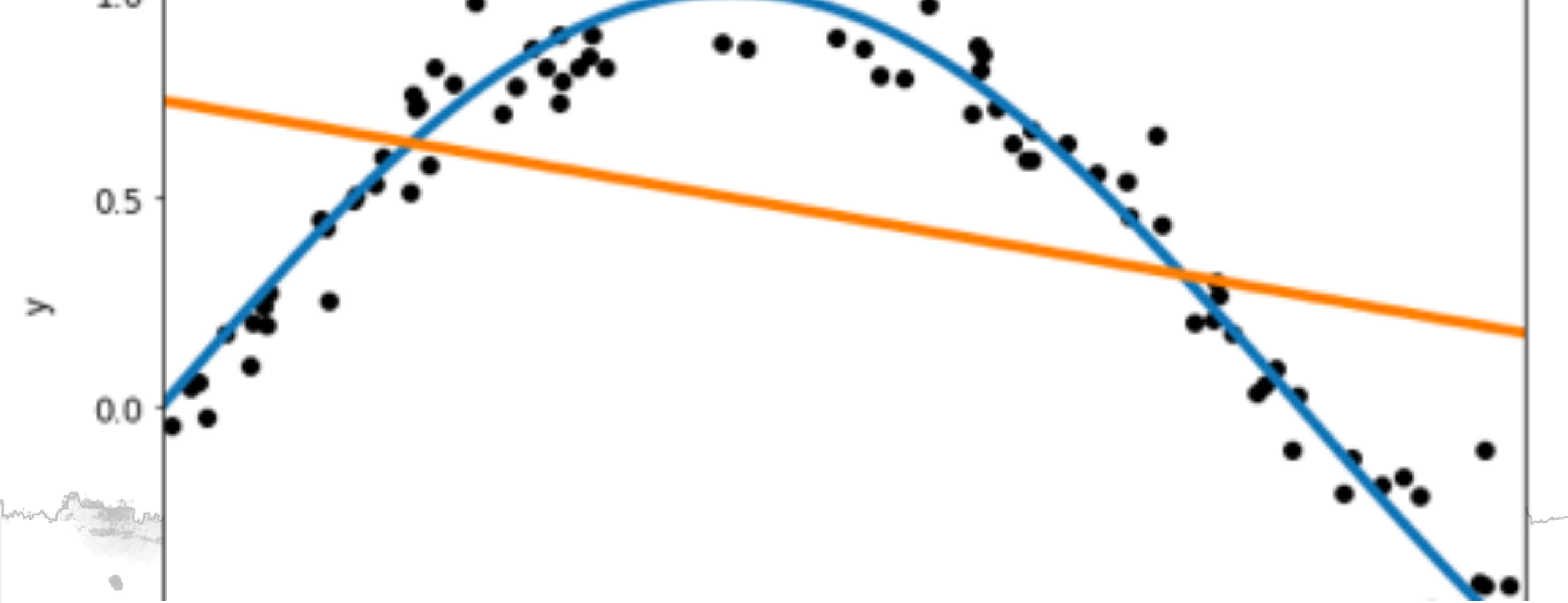
$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

		True condition	
	Total population	Condition positive	Condition negative
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error
	Predicted condition negative	**False negative**, Type II error	**True negative**

F1 = 2 * (precision * recall) / (precision + recall)

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$



Overfitting & Underfitting

- Why?

Assignment

- Summarize the reasons of overfitting and underfitting. Put them in github repository.

Lazy-Learning and Eager Learning

- Lazy Learning: Target function will be approximately locally;
- Dataset with few attributes.

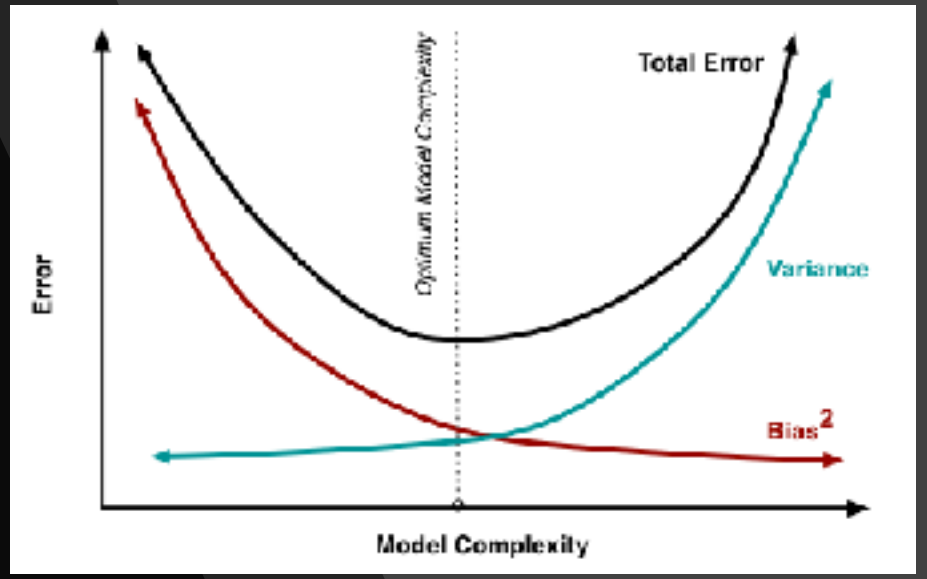
Outliner

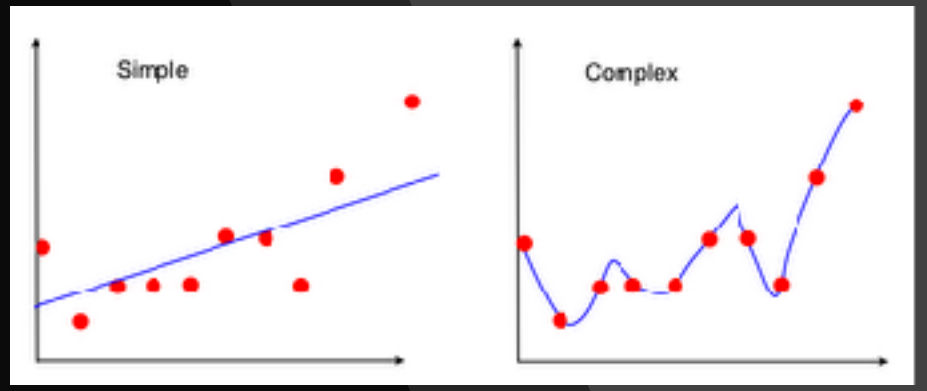
- What’s outliner and how to detect?
- Percentile

Bias and Variance

- dilemma
- The **bias** is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between feature and target outputs. (underfitting);

- The **variance** is an error from sensitivity to small fluctuations in the training set. High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs (overfitting).



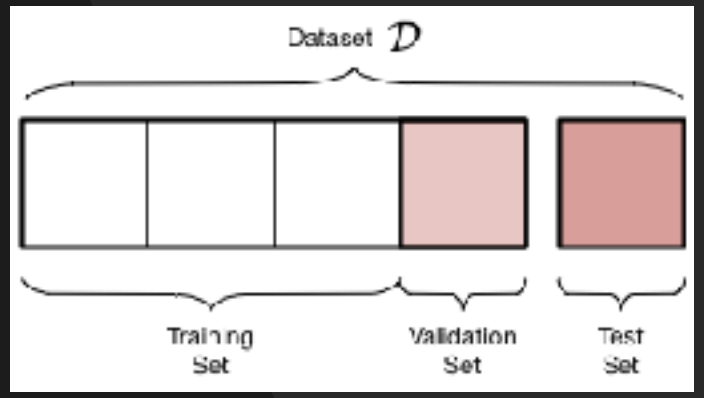


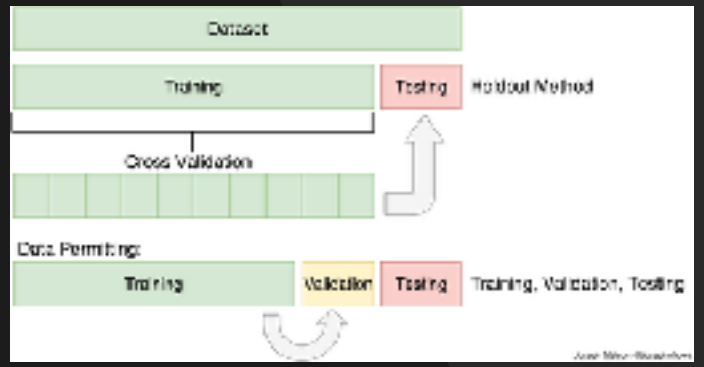
Train, validation, test

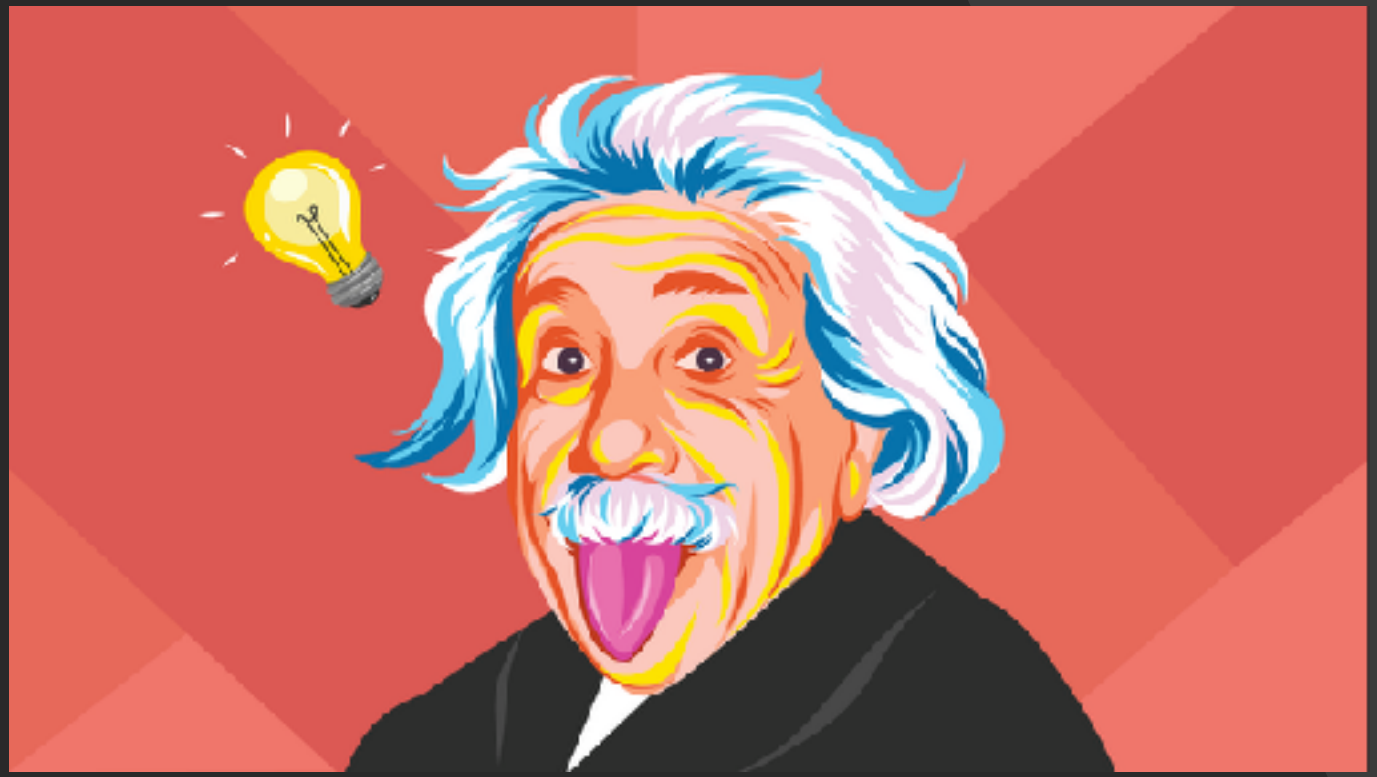
- Good Memory V.S Good Learning

Train, validation, test

- Good Memory V.S Good Learning







Assignment

- 1. Summarize the reasons of overfitting and underfitting. Put them in github repository.
- 2. install the numpy, scikit-learning, keras, tensorflow
- 3. Writing down three sceneries that machine learning has been used now.
- 4. Come out with three new sceneries with which machine learning may be applied.