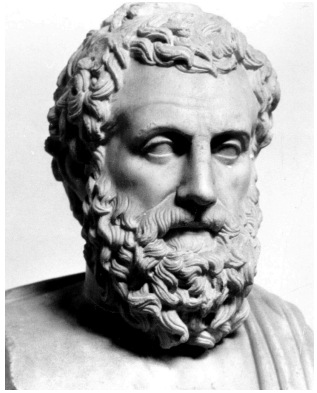
Artificial Intelligence For NLP Lesson-03



人工智能与自然语言处理课程组

2019.10. 18

Outline

Machine Learning

- Background
- Main Methodologies
- Current Trends

Underfitting and Overfitting

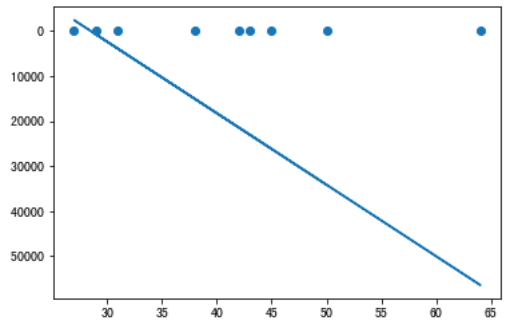
- Bias and Variance
- Model Capacity
- Underfitting and Overfitting

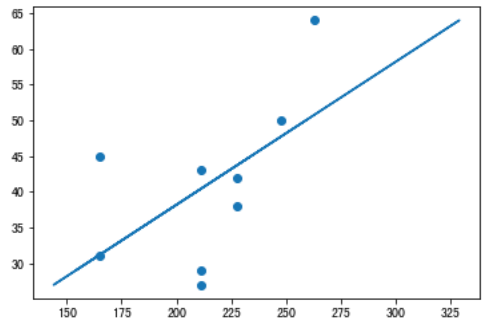
Train set, test set, validation set

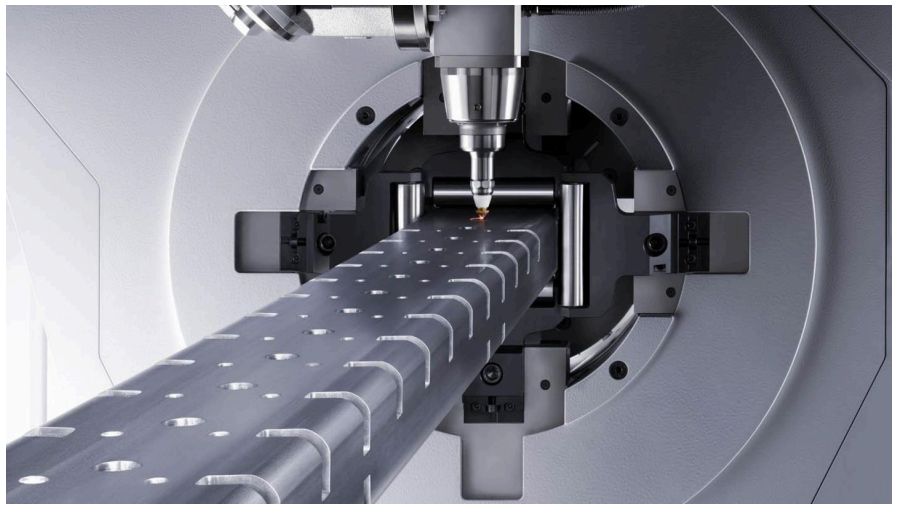
- Influence of dataset
- The relation of train, test and validation.

Review

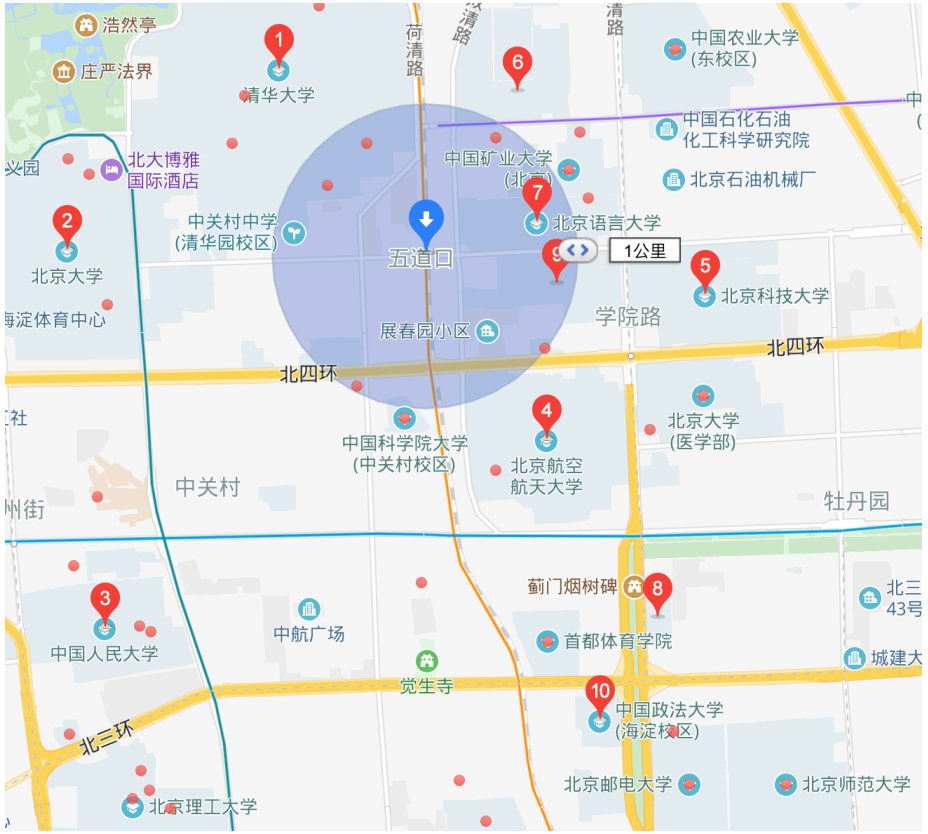
```
while True:
    if loss(y_true=fare, yhats=yhats) < eps: break

    indices = np.random.choice(range(len(age)), size=10)

    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(x, _a, _b) for x 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.

2018

Your source for Calendar
www.jfcoolgame.com

January

Sun	Mon	Tue	Wed	Thu	Fri	Sat
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

February

Sun	Mon	Tue	Wed	Thu	Fri	Sat
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28			

March

Sun	Mon	Tue	Wed	Thu	Fri	Sat
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

April

Sun	Mon	Tue	Wed	Thu	Fri	Sat
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30					

May

Sun	Mon	Tue	Wed	Thu	Fri	Sat
		1	2	3	4	5
6	7	8	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31		

June

Sun	Mon	Tue	Wed	Thu	Fri	Sat
					1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30

July

Sun	Mon	Tue	Wed	Thu	Fri	Sat
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

August

Sun	Mon	Tue	Wed	Thu	Fri	Sat
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

September

Sun	Mon	Tue	Wed	Thu	Fri	Sat
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

October

Sun	Mon	Tue	Wed	Thu	Fri	Sat
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

November

Sun	Mon	Tue	Wed	Thu	Fri	Sat
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	

December

Sun	Mon	Tue	Wed	Thu	Fri	Sat
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31					

01-Jan-18 New Year's Day
15-Jan-18 B'day of M Luther King, Jr
19-Feb-18 Washington's Birthday
13-May-18 Mother's Day
28-May-18 Memorial Day
17-Jun-18 Father's Day
04-Jul-18 Independence Day
03-Sep-18 Labor Day
08-Oct-18 Columbus Day
12-Nov-18 Veterans Day
22-Nov-18 Thanksgiving Day
25-Dec-18 Christmas Day

Decision

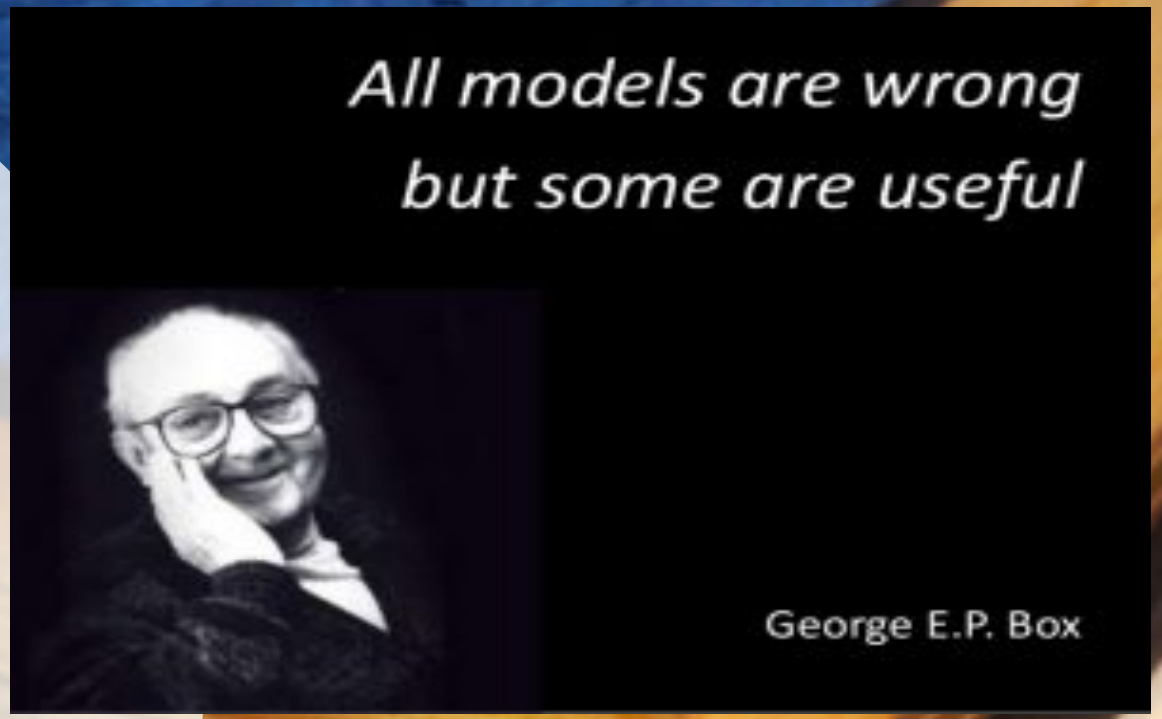
- 1. Looking for a book
  - if there is holding a ceremony
  - how may days
  - female vs male
  - The closet college
  - ..

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



Model

2. Target: Which college should I go to?

y

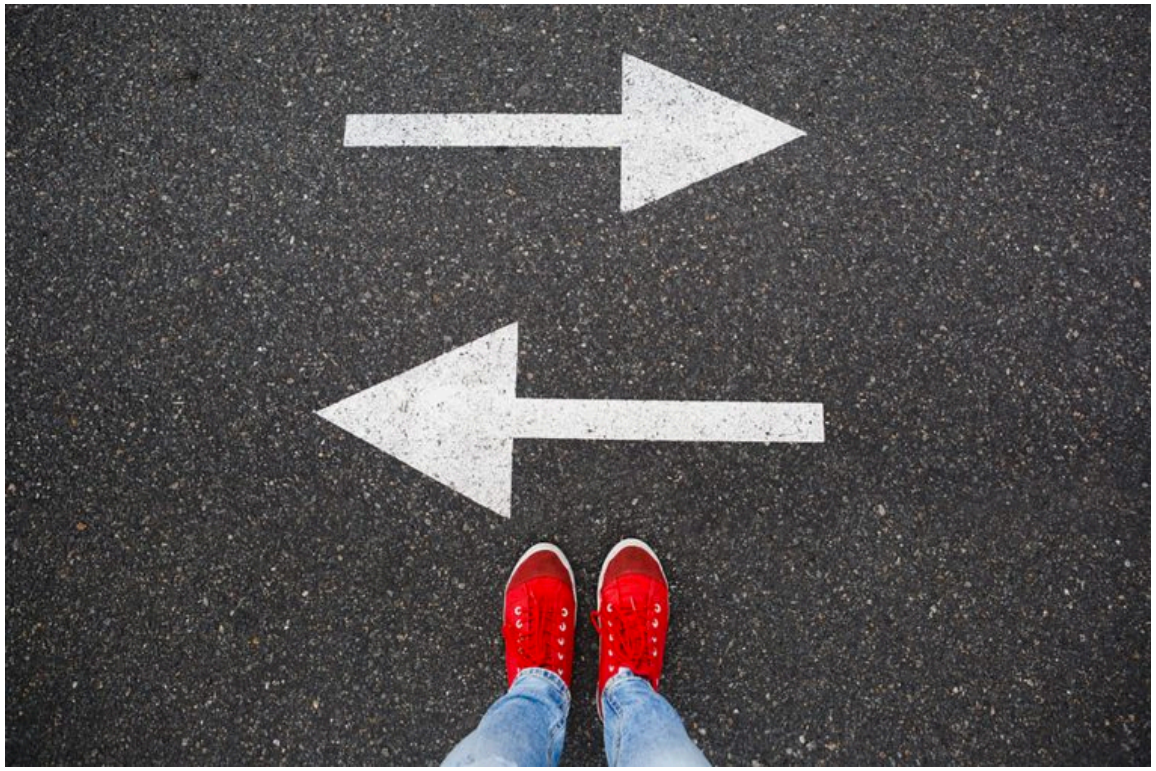
The First Book

- 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
16	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
16	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	最近的学校
		客流明显变少	不下雨	北京师范大学
			下雨	
	时间小于7天	客流没有明显变少		
			XXXXXX	北京大学
无活动			XXXXXX	北京语言大学
			XXXXXX	
		XXXXXX	XXXXXX	
			XXXXXX	清华大学
			XXXXXX	
			XXXXXX	
			XXXXXX	
			XXXXXX	
			XXXXXX	
	时间大于7天	XXXXXX	XXXXXX	
			XXXXXX	
			XXXXXX	
			XXXXXX	
	时间小于7天			
有活动				



The First Book

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

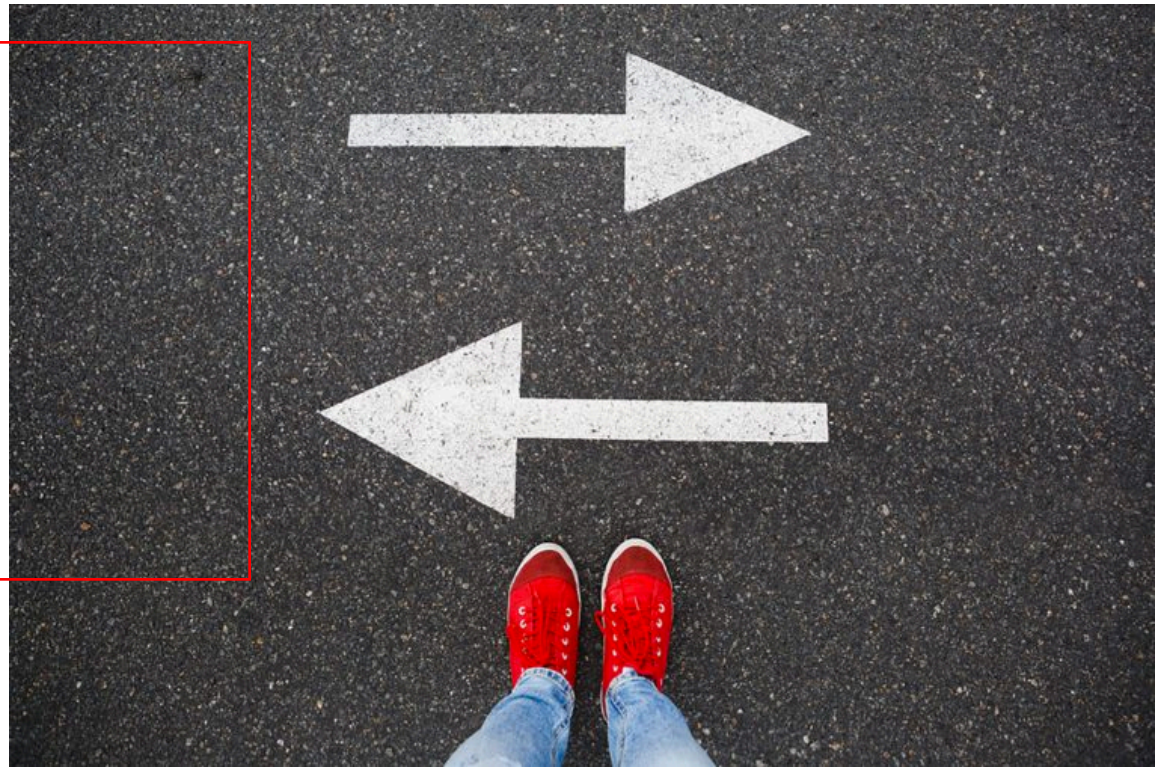
K-nearest neighbors

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
16	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
16	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	最近的学校
		客流明显变少	不下雨	北京师范大学
			下雨	
	时间小于7天	客流没有明显变少		
			XXXXXX	北京大学
无活动			XXXXXX	北京语言大学
			XXXXXX	
		XXXXXX	XXXXXX	
			XXXXXX	清华大学
			XXXXXX	
			XXXXXX	
			XXXXXX	
			XXXXXX	
			XXXXXX	
	时间大于7天	XXXXXX	XXXXXX	
			XXXXXX	
			XXXXXX	
			XXXXXX	
	时间小于7天			
有活动				

Decision Tree



The third book

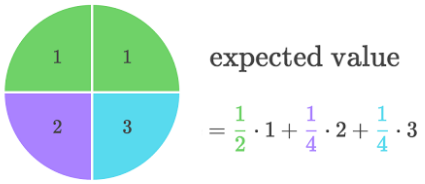
$$\text{expected value} = \frac{1}{2}\cdot 1 + \frac{1}{4}\cdot 2 + \frac{1}{4}\cdot 3$$

平均纯收入	时间3天	时间5天	时间7天	下雨	晴天	沙城暴	在中关村	在元大都遗址
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	379	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	884	163	934	317

The third book

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

- Naïve Bayesian Classification



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

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

平均纯收入	时间3天	时间5天	时间7天	下雨	晴天	沙城暴	在中关村	在元大都遗址
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	379	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	884	163	934	317

The Forth book

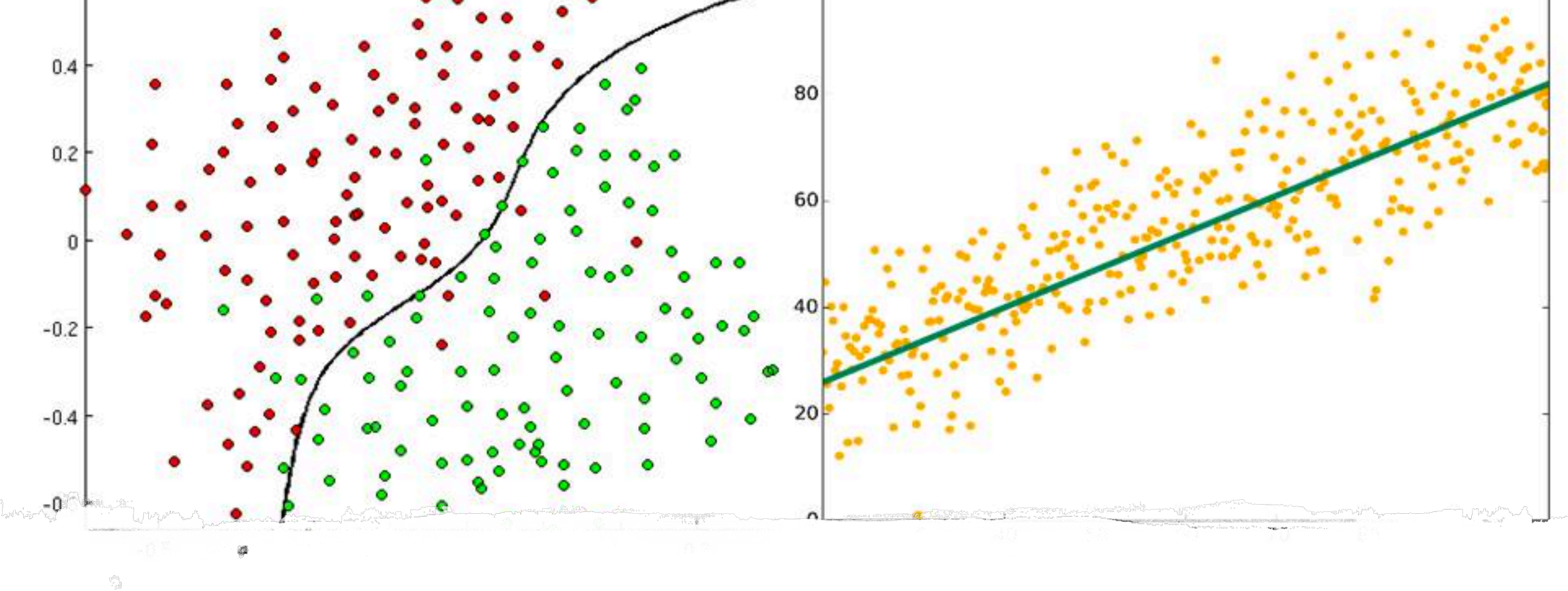
- 食堂打饭 不标价钱， 问， 每种菜多少钱?
-
- 第一次 茄子*2 + 西瓜*1 + 馒头 * 1 : 7.8元
- 第二次 茄子*1 + 西瓜*1 + 馒头 * 2 : 6.5元
- 第三次 玉米*1 + 菠菜*2 + 馒头 * 1: 5.6元
- …

- 每种价格?

- Neural Network

Chat

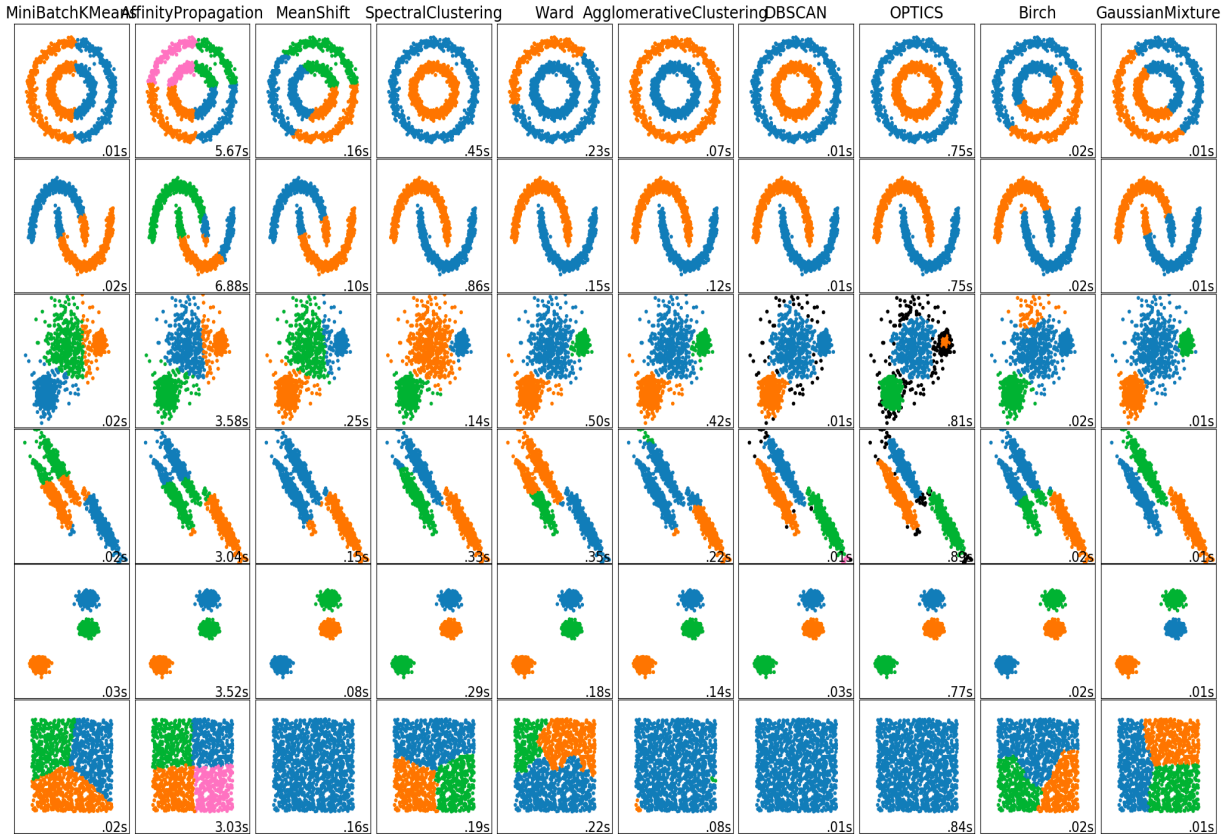
- Determine if is *a valuable customer* in Wechat?
- What Features do we need?
- How to predicate it?

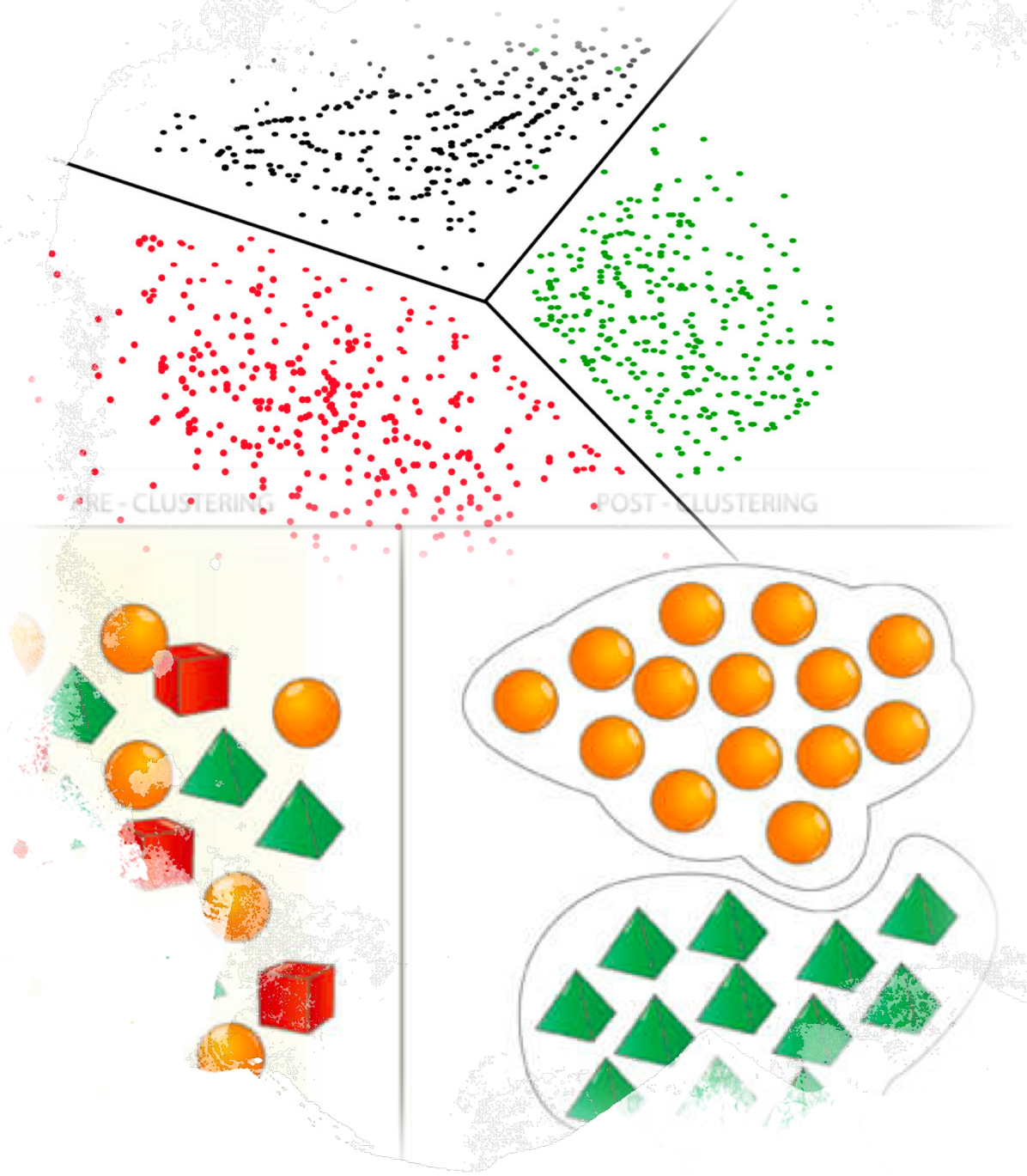


Two Type

- 1. Classification
- 2. Regression

Cluster


MiniBatchKMeans
AffinityPropagation
MeanShift
SpectralClustering
Ward
AgglomerativeClustering
DBSCAN
OPTICS
Birch
GaussianMixture
.01s
5.67s
.16s
.45s
.23s
.07s
.01s
.75s
.02s
.01s
.02s
6.88s
.10s
.86s
.15s
.12s
.01s
.75s
.02s
.01s
.02s
3.58s
.25s
.14s
.50s
.42s
.01s
.81s
.02s
.01s
.02s
3.04s
.15s
.33s
.35s
.22s
.01s
.89s
.02s
.01s
.03s
3.52s
.08s
.29s
.18s
.14s
.03s
.77s
.02s
.01s
.02s
3.03s
.16s
.19s
.22s
.08s
.01s
.84s
.02s
.01s



How to evaluate?

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

- Generation

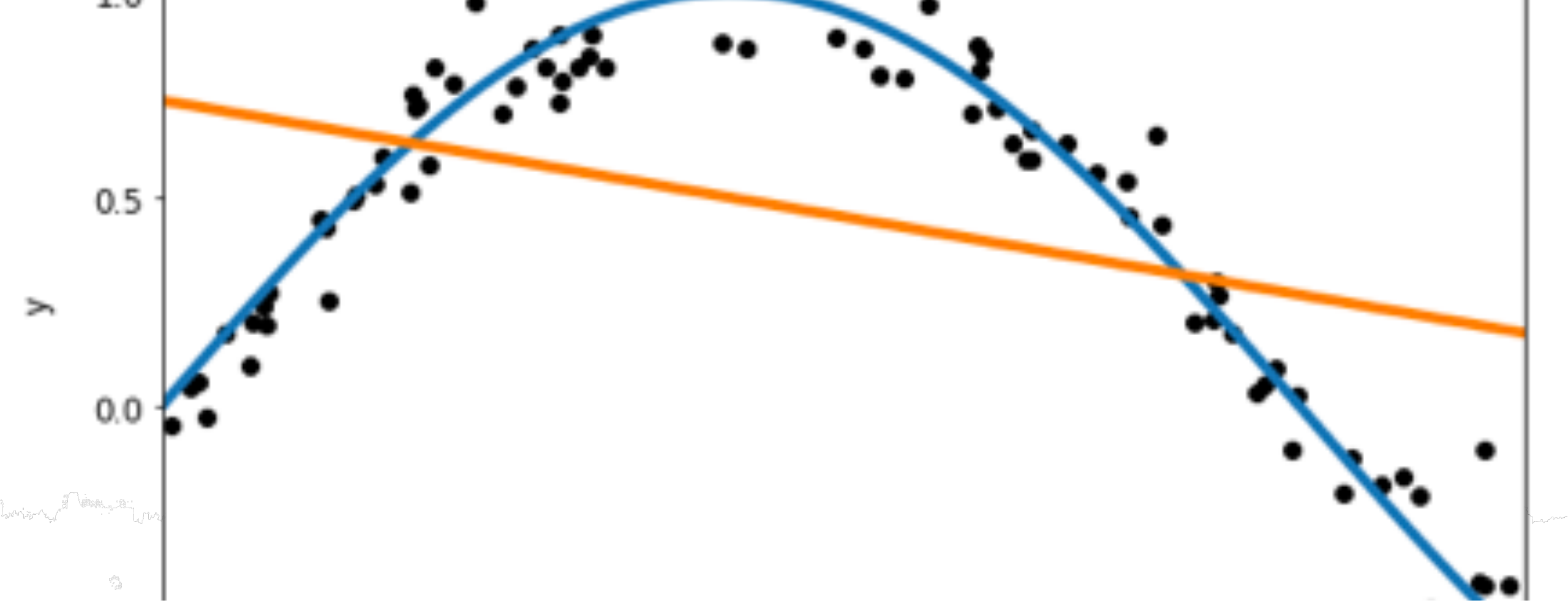
$$\text{Precision} = \frac{tp}{tp + fp}$$

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

		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?

Lazy-Learning and Eager Learning

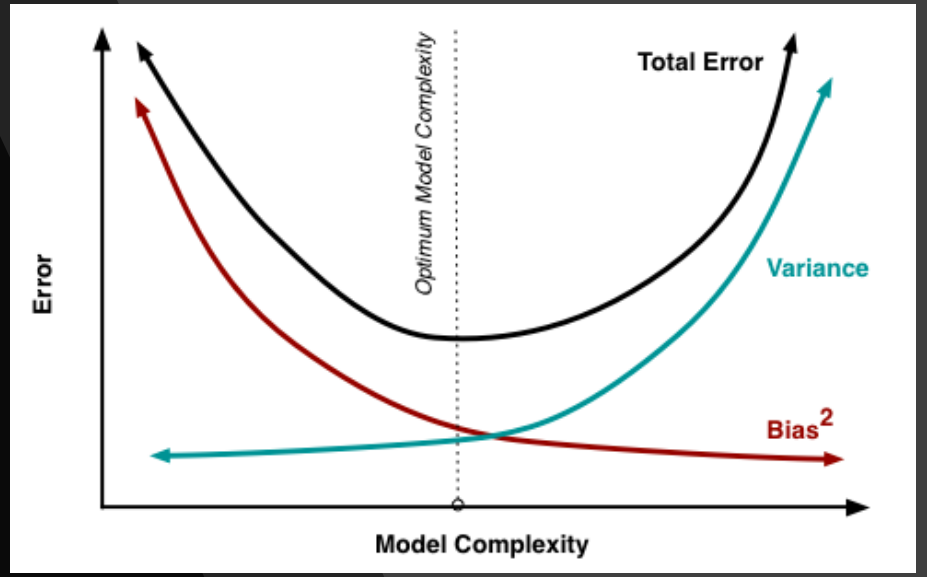
- Lazy Learning: Target function will be approximately locally;
- Dataset with few attributes.

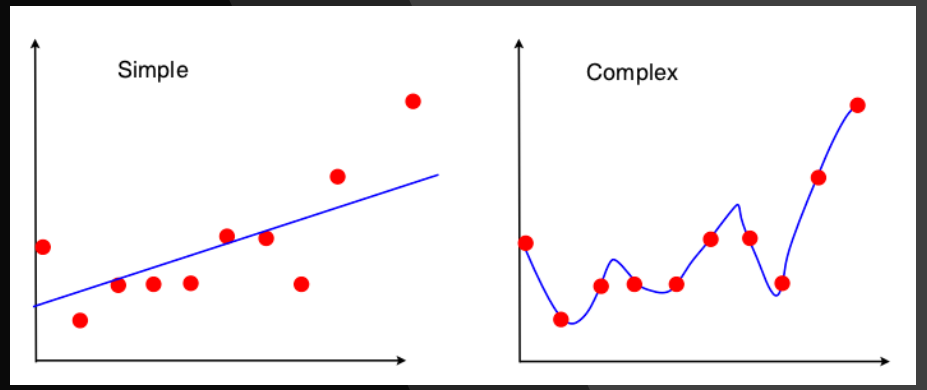
Outliner

- 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).



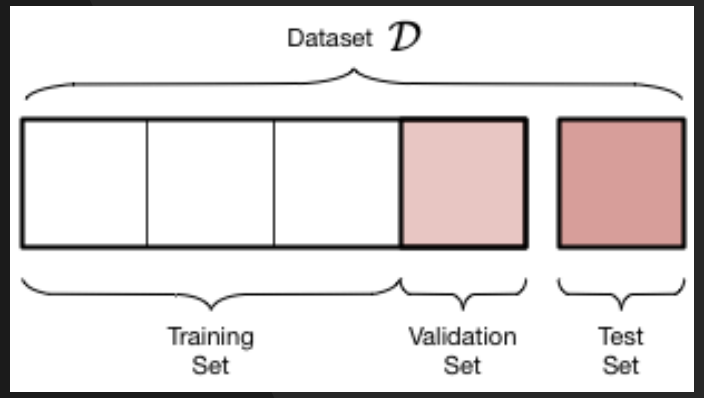


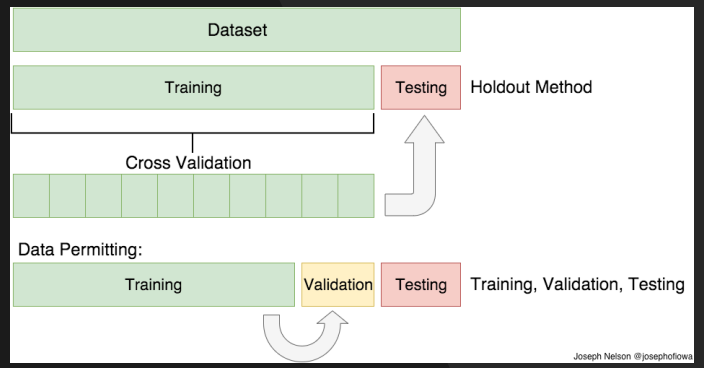
Train, validation, test

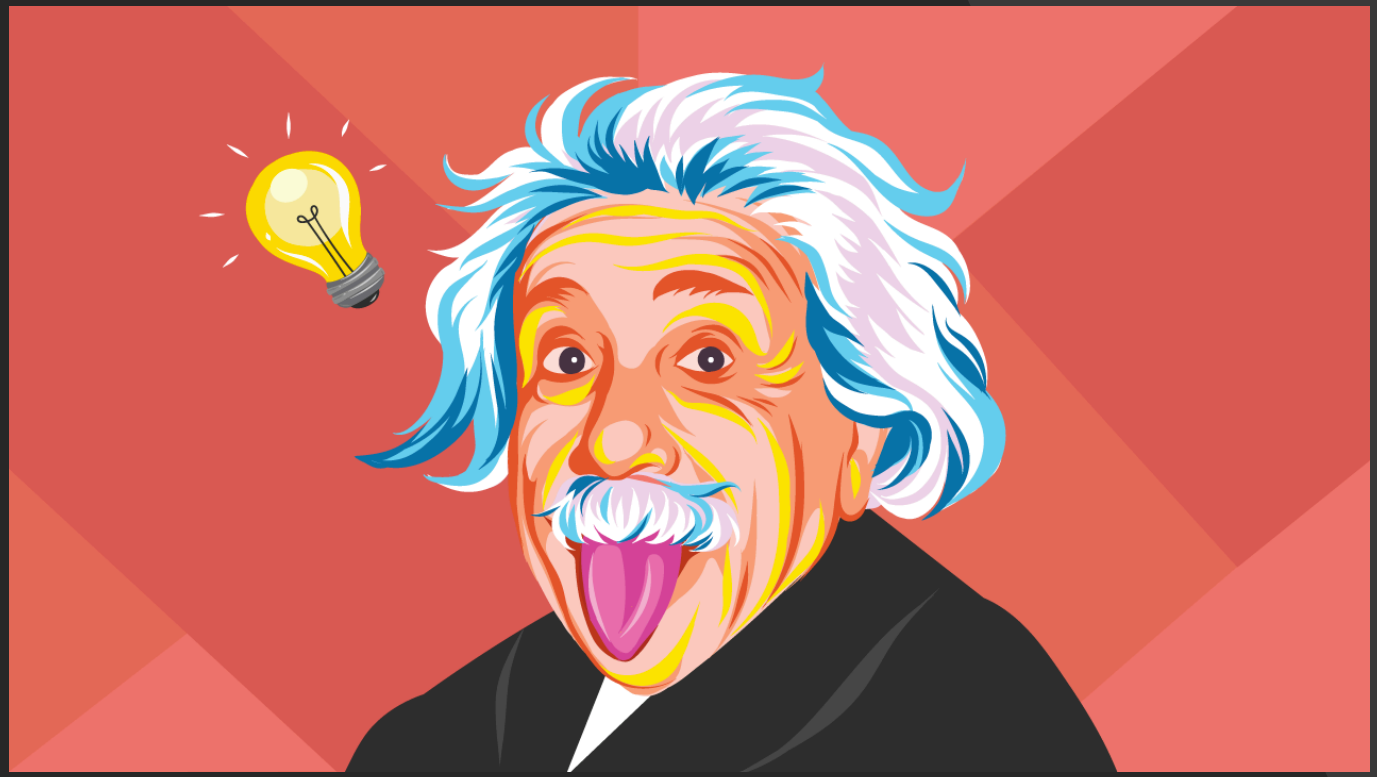
- Good Memory V.S Good Learning

Train, validation, test

- Good Memory V.S Good Learning







Assignment

- 1. Summarize the reasons of overfitting and underfitting. Put them in github repository.
- 2. Implement the absolute value loss function in Linear Regression Model;
- 3. Implement the text cluster using Kmeans Model;
- 4. (Optional) Implement the Bayes text classifier Model.