

Artificial Intelligence For NLP Lesson-08

人工智能与自然语言处理课程组
2019.May. 19



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

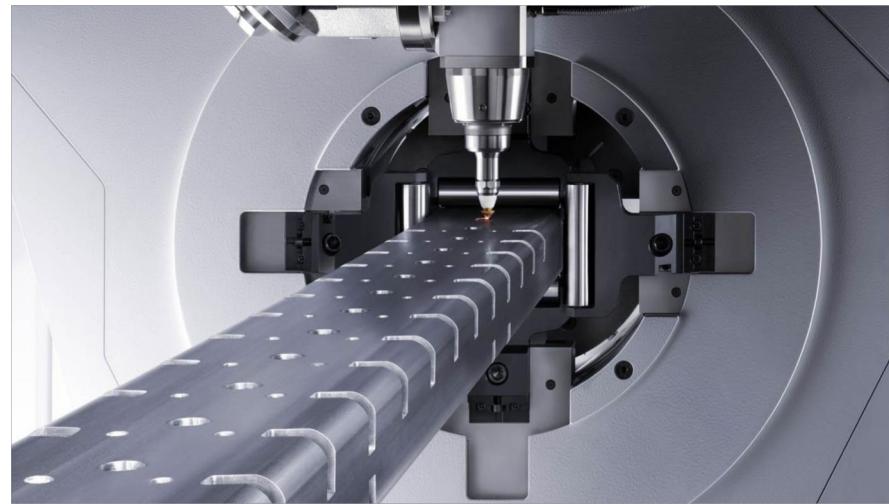
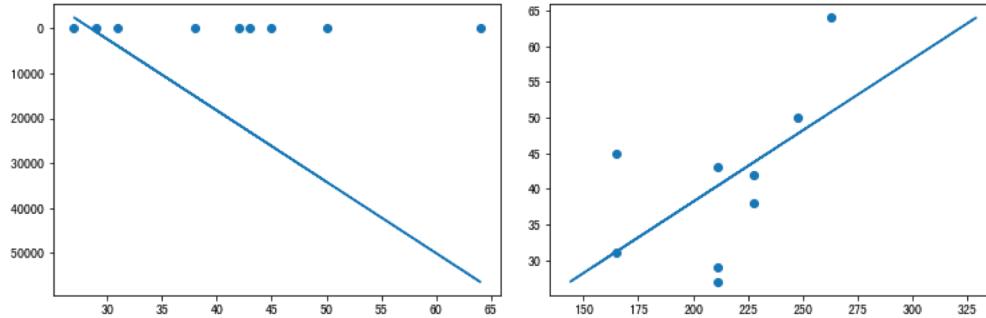
        y_hats = [model(x, _a, _b) for x in sample_x]
        l = loss(sample_y, np.array([model(x, a + da, b + db) for x in sample_x]))

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

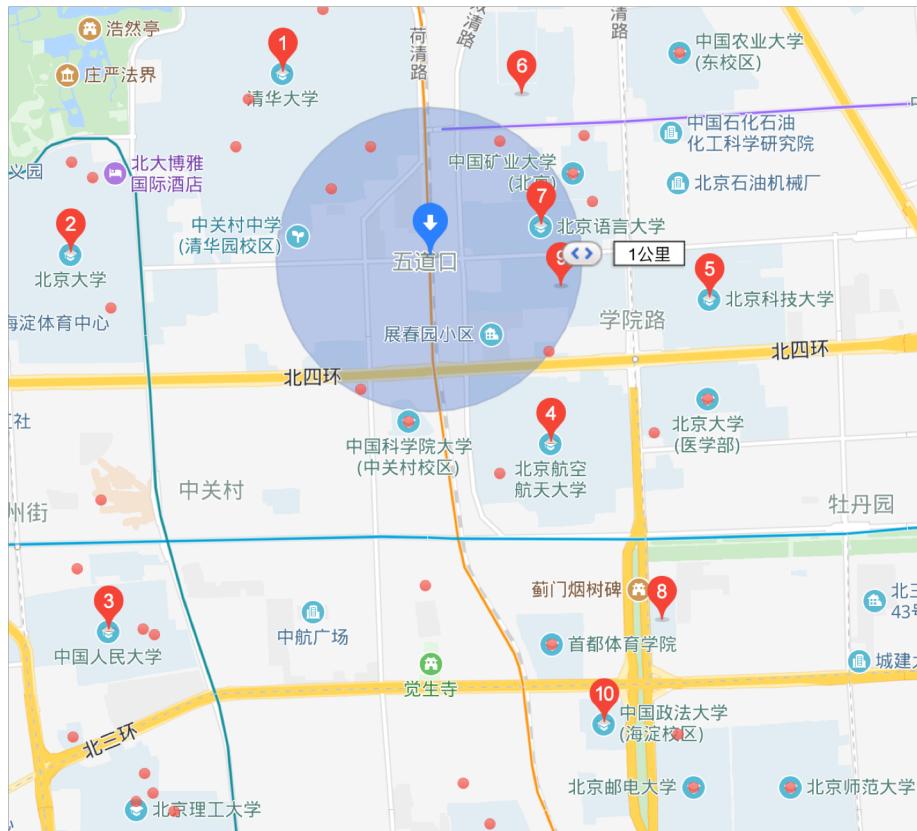
    if batch % 10 == 0:
        print('batch {} / {} fare with {} * age + {}, with loss: {}'.format(batch, total, a, b, l))

    if batch > total: break

    batch += 1
```



Example Driven



Target

- More sales, More money.



2018													
January				February				March					
Sun	Mon	Tue	Wed	Thu	Fri	Sat	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	
28	29	30	31				4	5	6	7	8	9	
				11	12	13	14	15	16	17	18	19	
21	22	23	24	25	26	27	18	19	20	21	22	23	
24	25	26	27	28	29	30	25	26	27	28	29	30	
April				May				June					
Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat
1	2	3	4	5	6	7	8	9	10	11	12	13	
9	10	11	12	13	14	15	16	17	18	19	20	21	
22	23	24	25	26	27	28	29	30	31				
				1	2	3	4	5	6	7	8	9	
15	16	17	18	19	20	21	13	14	15	16	17	18	
22	23	24	25	26	27	28	20	21	22	23	24	25	
29	30	31		27	28	29	30	31					
July				August				September					
Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat
1	2	3	4	5	6	7	8	9	10	11	12	13	
9	10	11	12	13	14	15	16	17	18	19	20	21	
22	23	24	25	26	27	28	19	20	21	22	23	24	
29	30	31		26	27	28	29	30	31				
				1	2	3	4	5	6	7	8	9	
15	16	17	18	19	20	21	12	13	14	15	16	17	
22	23	24	25	26	27	28	19	20	21	22	23	24	
29	30	31		26	27	28	29	30	31				
October				November				December					
Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat
1	2	3	4	5	6	7	8	9	10	11	12	13	
7	8	9	10	11	12	13	14	15	16	17	18	19	
14	15	16	17	18	19	20	11	12	13	14	15	16	
21	22	23	24	25	26	27	18	19	20	21	22	23	
28	29	30	31		25	26	27	28	29	30	31		
					1	2	3	4	5	6	7	8	
15-Jan-18 New Year's Day	28-May-18 Memorial Day	08-Jun-18 Columbus Day	08-Nov-18 Veterans Day	15-Jan-18 Day of M Luhur King	19-Jun-18 Father's Day	12-Nov-18 Thanksgiving Day	15-Feb-18 Washington's Birthday	04-Jul-18 Independence Day	22-Nov-18 Thanksgiving Day	13-May-18 Mother's Day	03-Sep-18 Labor Day	25-Dec-18 Christmas Day	



Decision

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

Features

- 2. Target: Which college should I go to?

y

*All models are wrong
but some are useful*



George E.P. Box

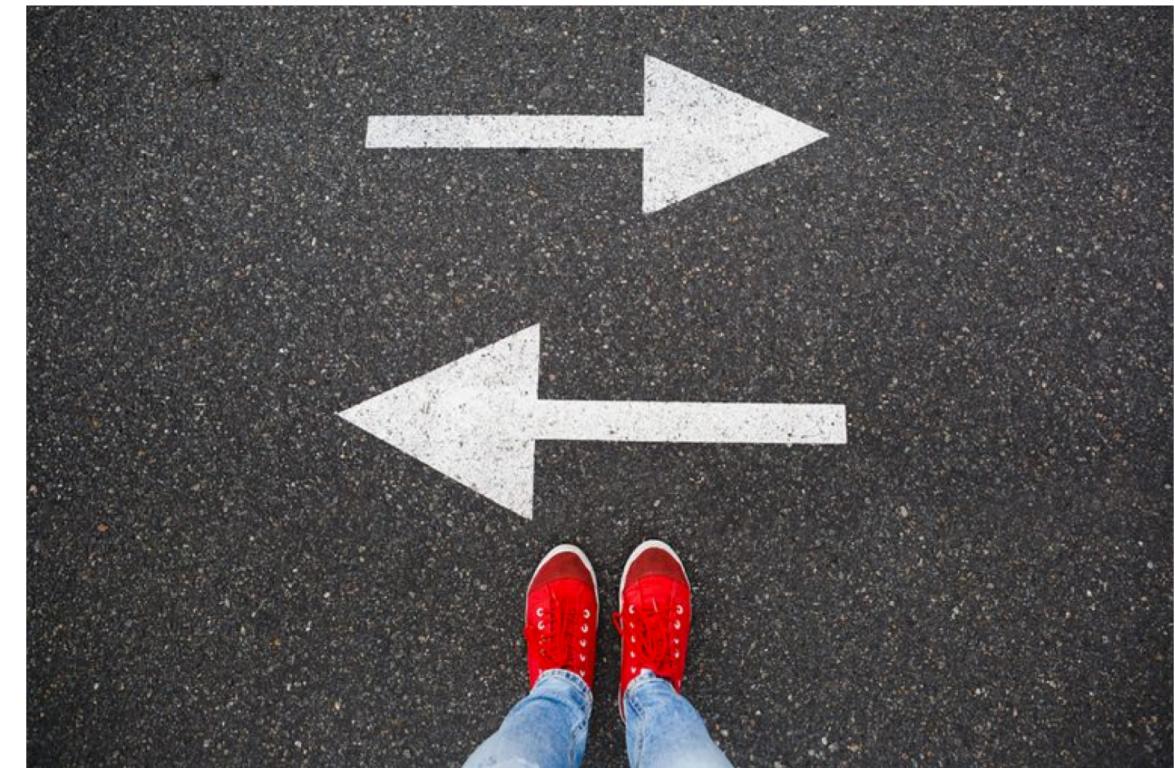
Model

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
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



The First Book

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

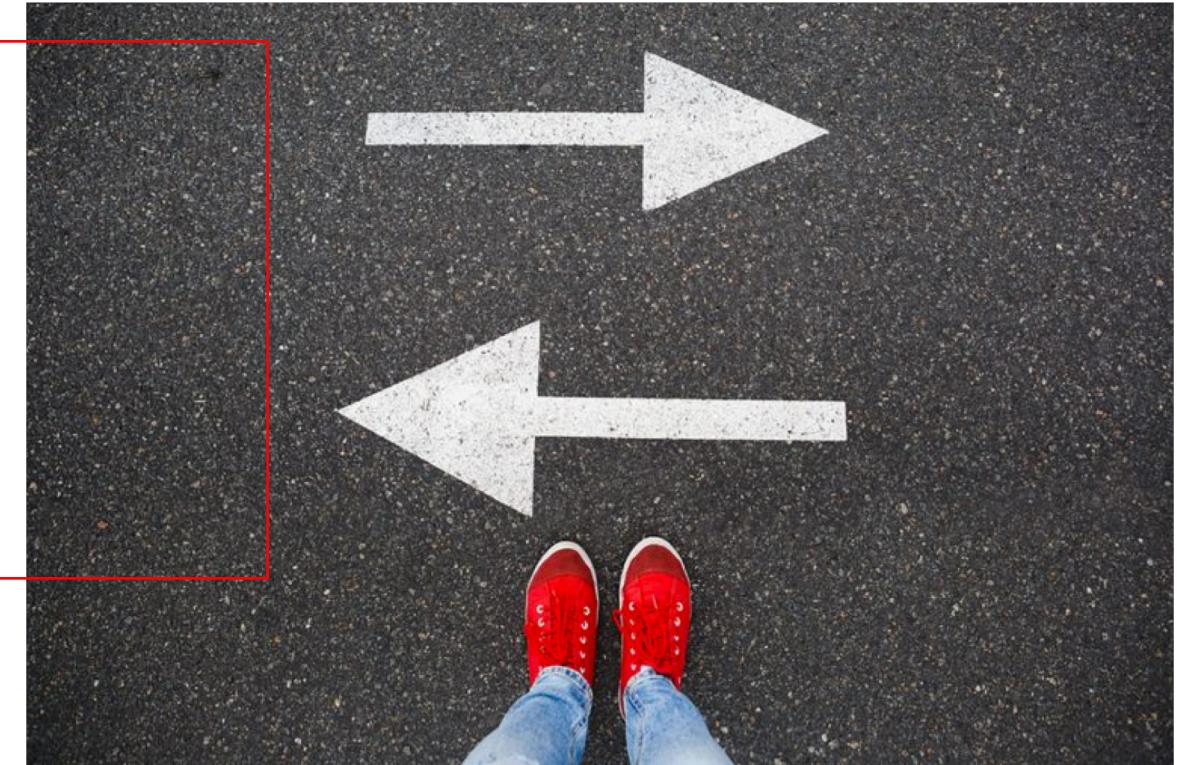
K-nearest neighbors

	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

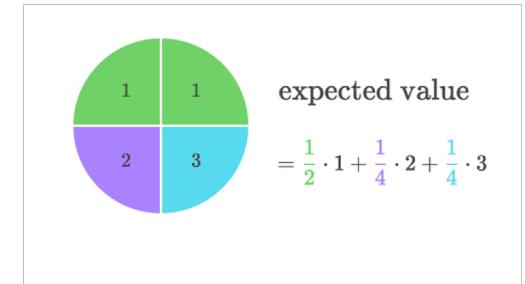
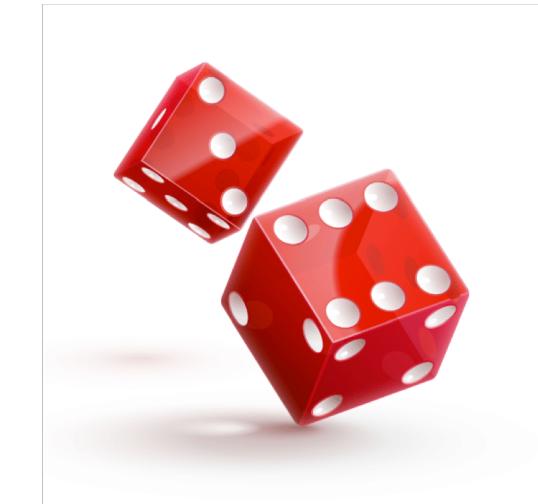
The second book



Decision Tree



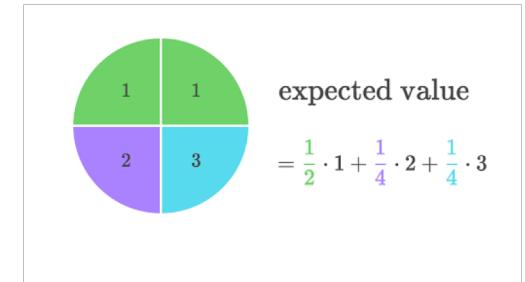
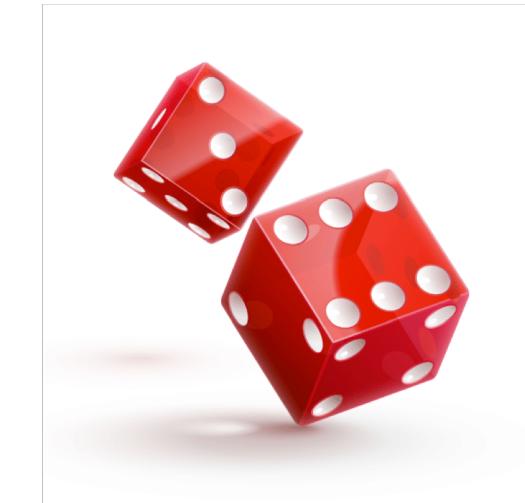
The third book



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

- $P(U|A_1, A_2, A_3)$
 - $\sim P(A_1|U) * P(A_2|U) * P(A_3|U)$
- Naïve Bayesian Classification



$$\text{Posterior} = \frac{\text{Likelihood} * \text{Prior}}{\text{Evidence}}$$
$$P(C_j | A_1, A_2, \dots, A_n) = \frac{\left(\prod_{i=1}^n P(A_i | C_j) \right) P(C_j)}{P(A_1, A_2, \dots, A_n)}$$

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	458	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	208	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

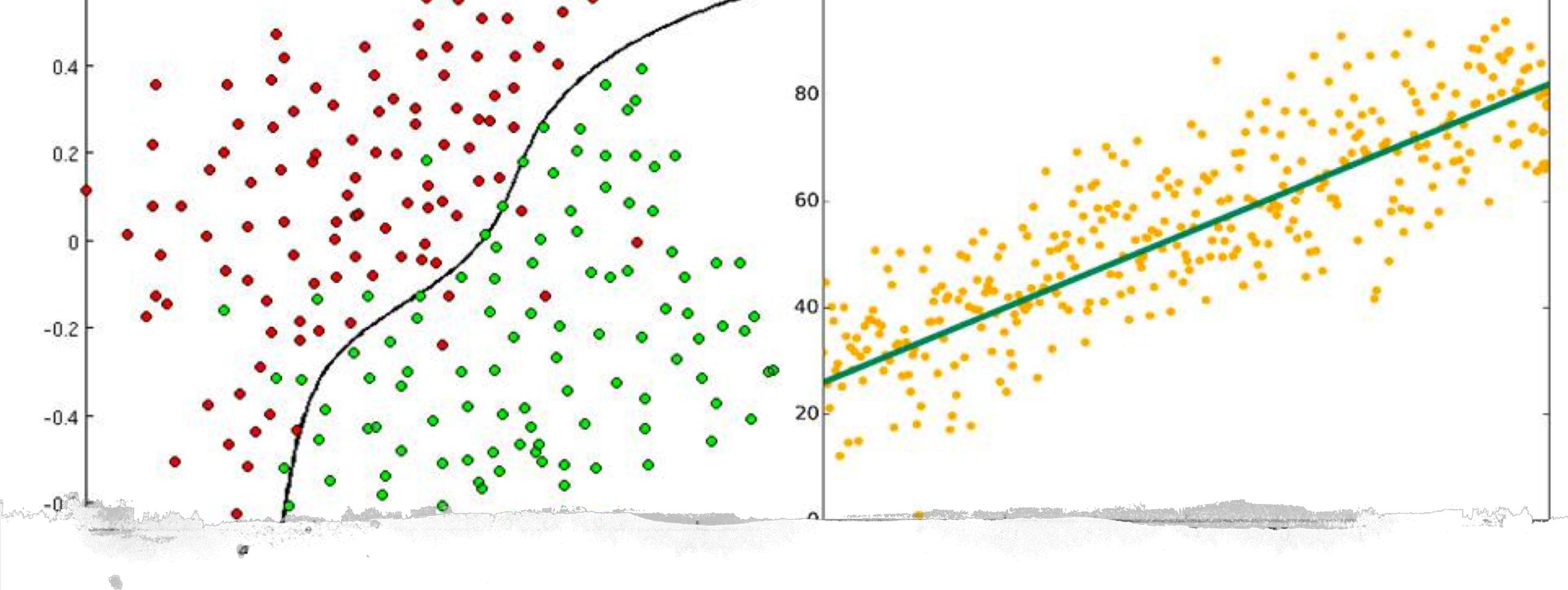
- 食堂打饭 不标价钱， 问， 每种菜多少钱?
-
- 第一次 茄子*2 + 西瓜*1 + 馒头 * 1 : 7.8元
- 第二次 茄子*1 + 西瓜*1 + 馒头 * 2 : 6.5元
- 第三次 玉米*1 + 菠菜*2 + 馒头 * 1: 5.6元
- ...
- 每种价格?
- Neural Network



Chat

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





Two Type

- 1. Classification
- 2. Regression

How to evaluate?

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

- Generation

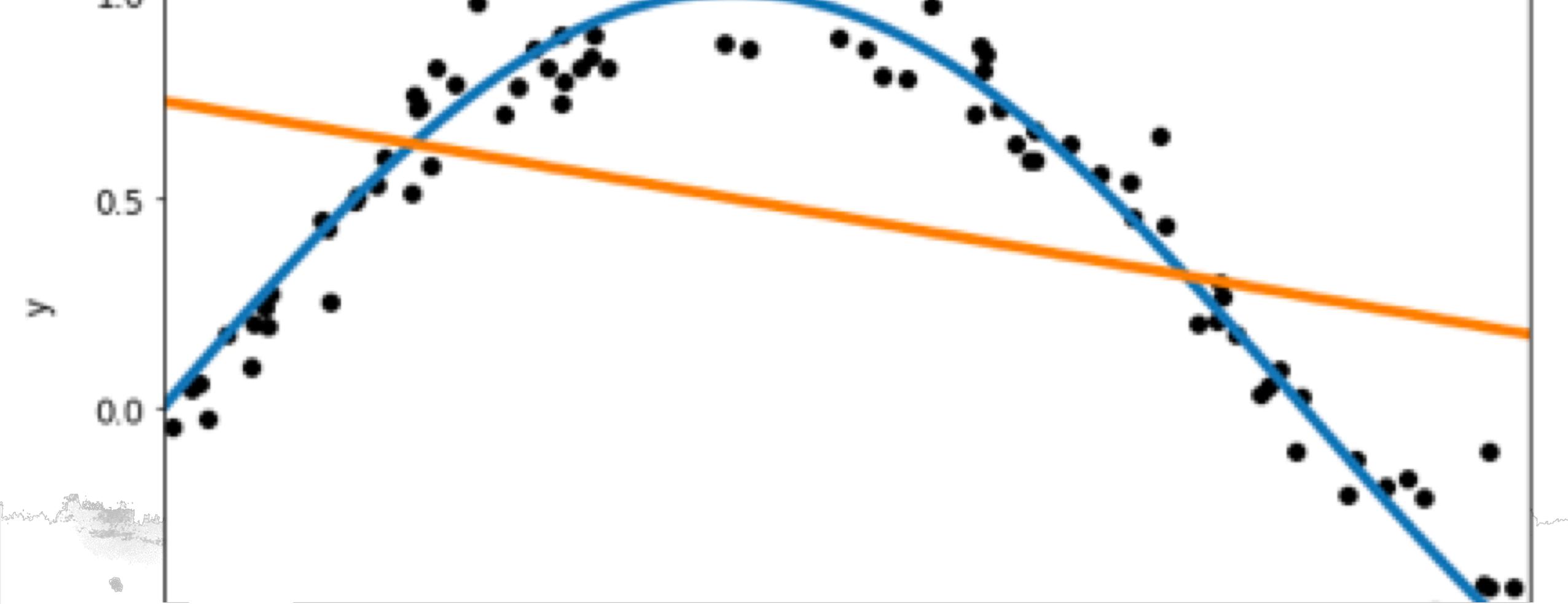
$$\text{Precision} = \frac{tp}{tp + fp}$$

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

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

$$F1 = 2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$$

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



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.

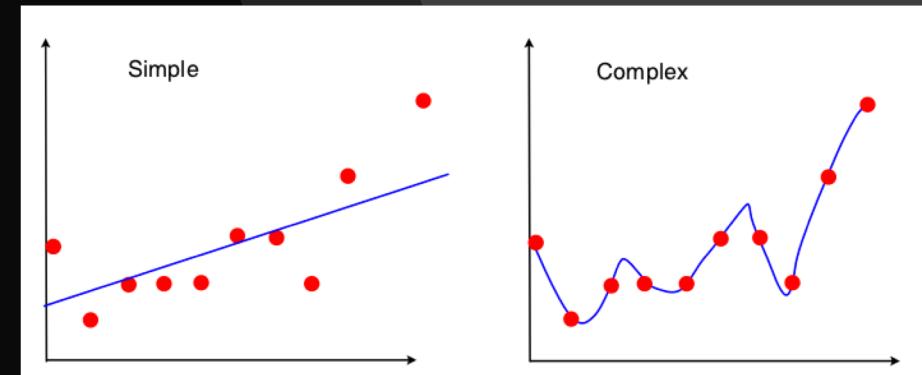
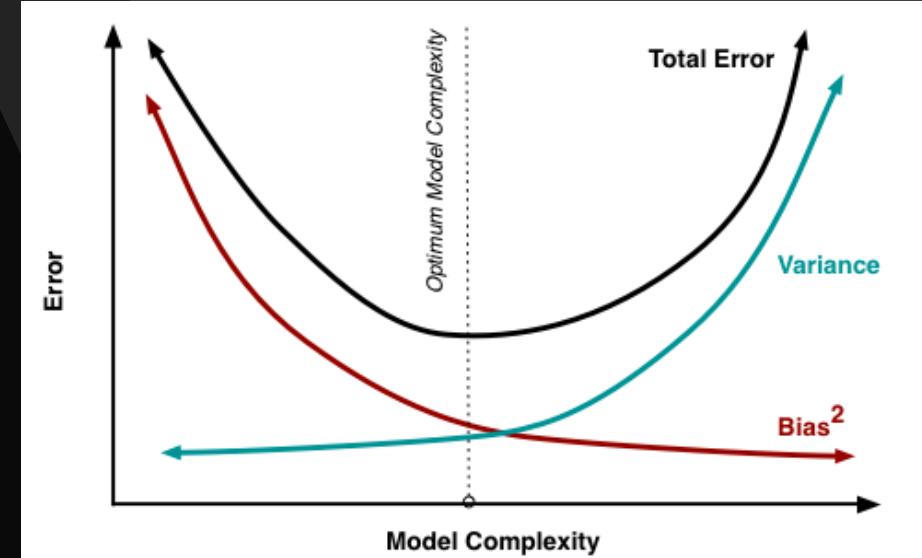
Outliner



- 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).



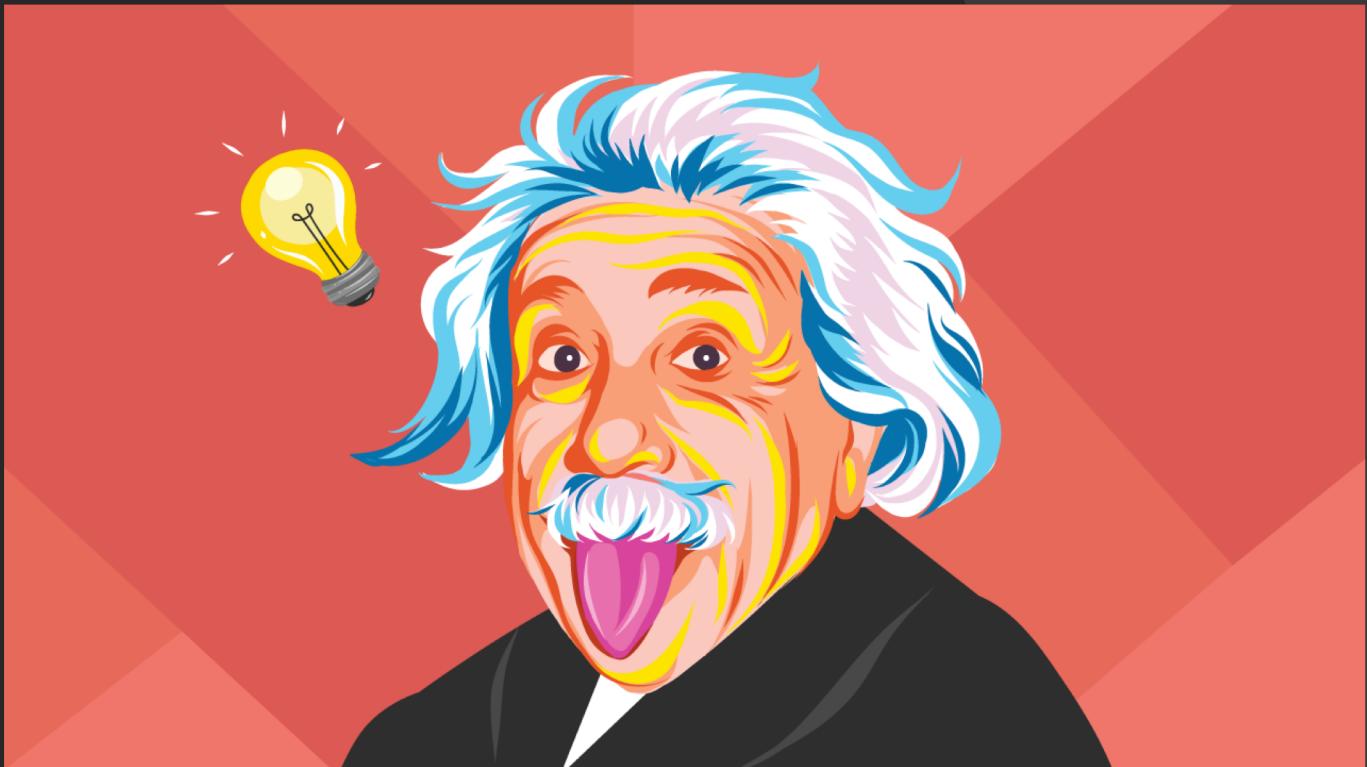
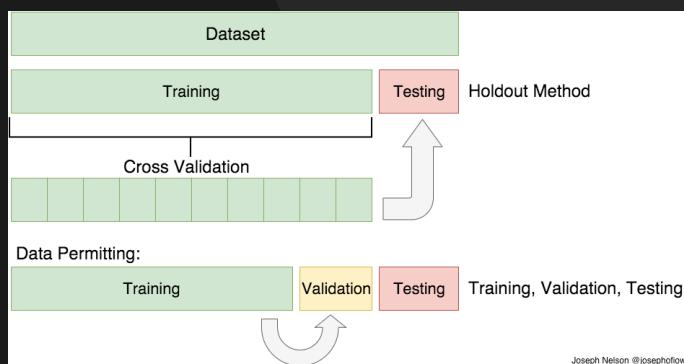
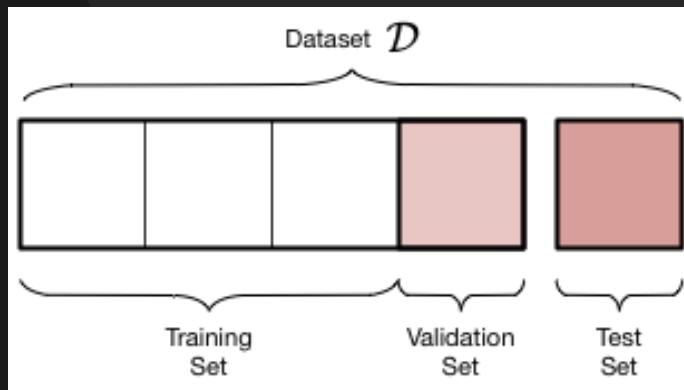
Train, validation, test

- 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.