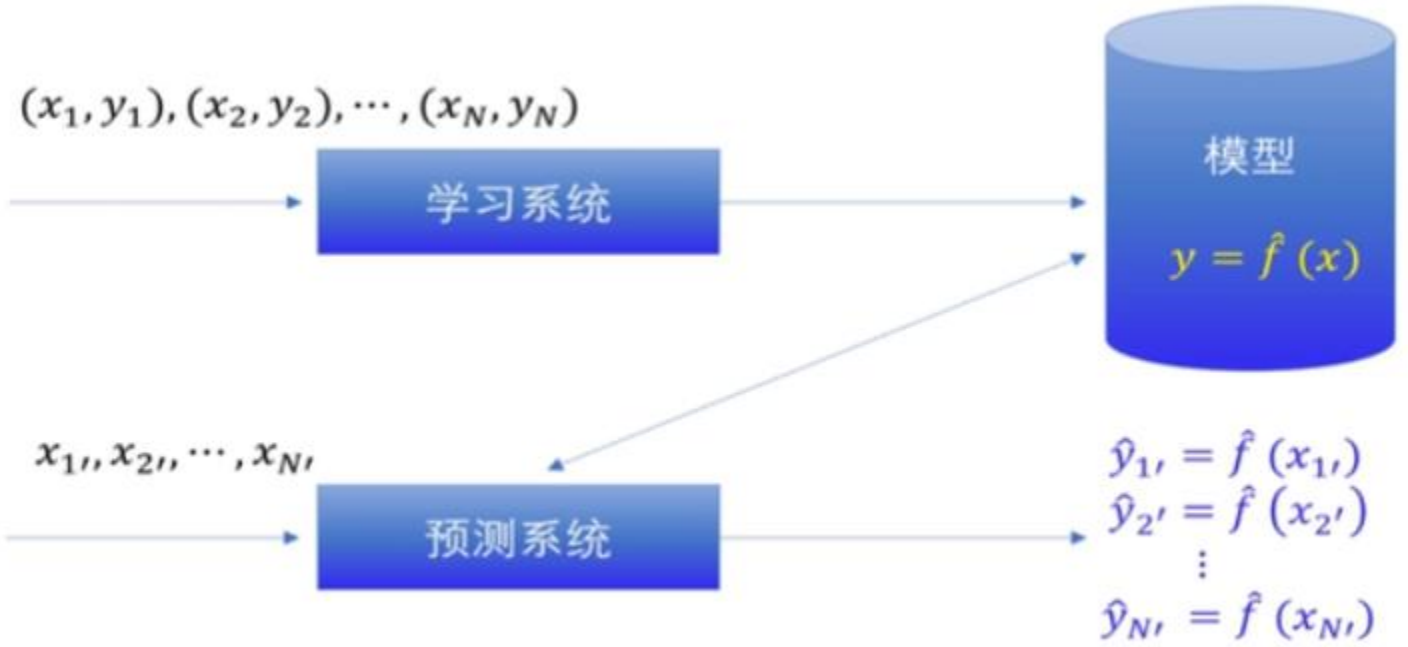
模型选择和评估

如何选择模型？

- 对一个给定的监督学习任务，应该选择哪个学习模型？
- 如何选择该模型的最优参数？
- 如何估计训练好的模型在学习样例之外的数据上可能的性能？

训练误差与测试误差


$(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)$
学习系统
模型
$y = \hat{f}(x)$
$x_{1'}, x_{2'}, \cdots, x_{N'}$
预测系统
$\hat{y}_{1'} = \hat{f}(x_{1'})$
$\hat{y}_{2'} = \hat{f}(x_{2'})$
$\hat{y}_{N'} = \hat{f}(x_{N'})$

训练集与测试集

- 训练集

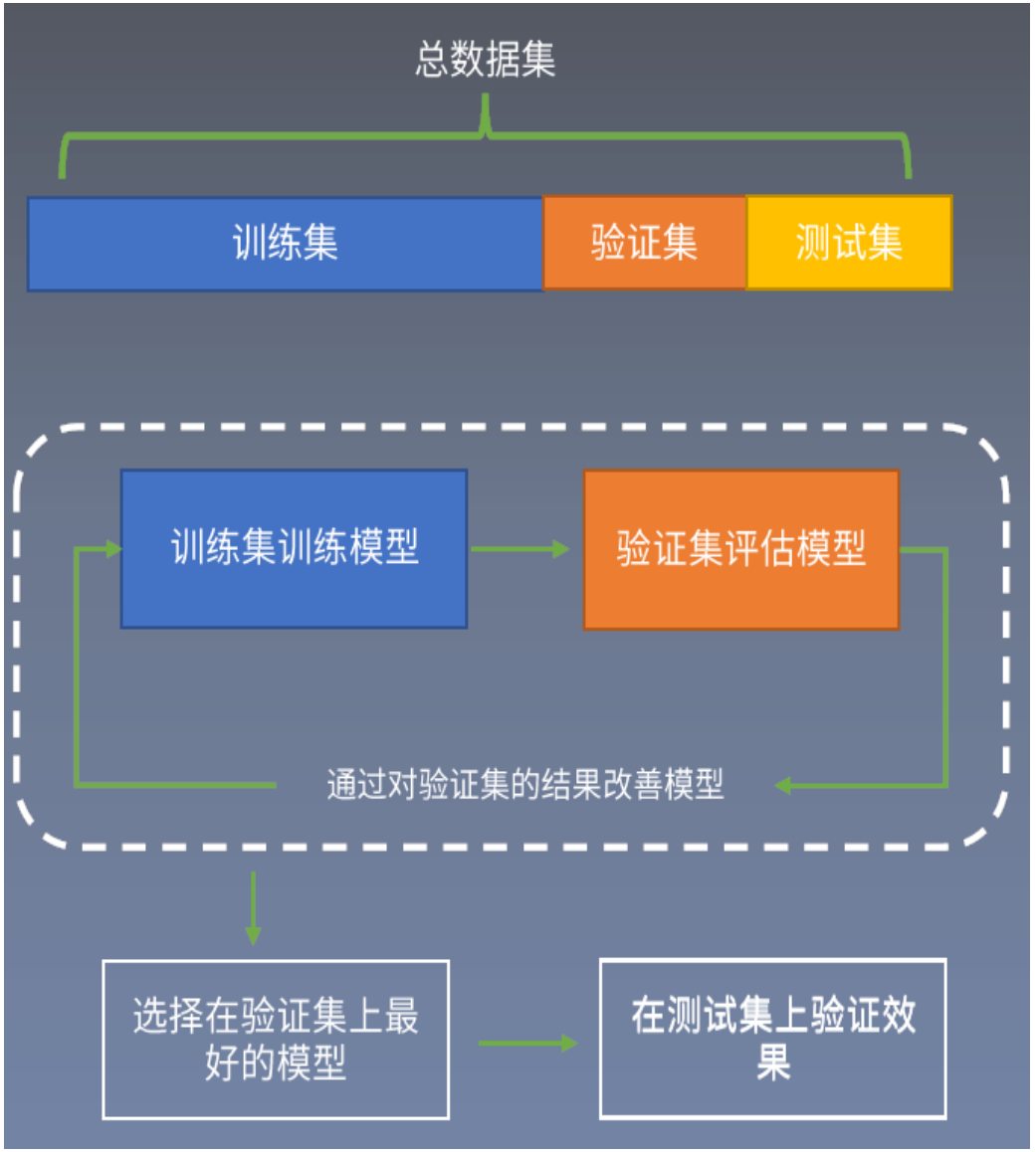
  用来训练模型，模型的迭代优化

- 测试集

  不参与训练流程，监测模型效果

- 验证集

  调整模型的超参数，优化模型



划分训练集和测试集

	Date	Title	Budget	DomesticTotalGross	Director	Rating	Runtime
0	2013-11-22	The Hunger Games: Catching Fire	130000000	424668047	Francis Lawrence	PG-13	146
1	2013-05-03	Iron Man 3	200000000	409013994	Shane Black	PG-13	129
2	2013-11-22	Frozen	150000000	400738009	Chris BuckJennifer Lee	PG	108
3	2013-07-03	Despicable Me 2	76000000	368061265	Pierre CoffinChris Renaud	PG	98
4	2013-06-14	Man of Steel	225000000	291045518	Zack Snyder	PG-13	143
5	2013-10-04	Gravity	100000000	274092705	Alfonso Cuaron	PG-13	91
6	2013-06-21	Monsters University	NaN	268492764	Dan Scanlon	G	107
7	2013-12-13	The Hobbit: The Desolation of Smaug	NaN	258366855	Peter Jackson	PG-13	161
8	2013-05-24	Fast & Furious 6	160000000	238679850	Justin Lin	PG-13	130
9	2013-03-08	Oz The Great and Powerful	215000000	234911825	Sam Raimi	PG	127
10	2013-05-16	Star Trek Into Darkness	190000000	228778661	J.J. Abrams	PG-13	123
11	2013-11-08	Thor: The Dark World	170000000	206362140	Alan Taylor	PG-13	120
12	2013-06-21	World War Z	190000000	202359711	Marc Forster	PG-13	116
13	2013-03-22	The Croods	135000000	187168425	Kirk De MiccoChris Sanders	PG	98
14	2013-06-28	The Heat	43000000	159582188	Paul Feig	R	117
15	2013-08-07	We're the Millers	37000000	150394119	Rawson Marshall Thurber	R	110
16	2013-12-13	American Hustle	40000000	150117807	David O. Russell	R	138
17	2013-05-10	The Great Gatsby	105000000	144840419	Baz Luhrmann	PG-13	143

划分训练集和测试集

	Date	Title	Budget	DomesticTotalGross	Director	Rating	Runtime
0	2013-11-22	The Hunger Games: Catching Fire	130000000	424668047	Francis Lawrence	PG-13	146
1	2013-05-03	Iron Man 3	200000000	409013994	Shane Black	PG-13	129
2	2013-11-22	Frozen	150000000	400738009	Chris BuckJennifer Lee	PG	108
3	2013-07-03	Despicable Me 2	76000000	368061265	Pierre CoffinChris Renaud	PG	98
4	2013-06-14	Man of Steel	225000000	291045518	Zack Snyder	PG-13	143
5	2013-10-04	Gravity	100000000	274092705	Alfonso Cuaron	PG-13	91
6	2013-06-21	Monsters University	NaN	268492764	Dan Scanlon	G	107
7	2013-12-13	The Hobbit: The Desolation of Smaug	NaN	258366855	Peter Jackson	PG-13	161
8	2013-05-24	Fast & Furious 6	160000000	238679850	Justin Lin	PG-13	130
9	2013-03-08	Oz The Great and Powerful	215000000	234911825	Sam Raimi	PG	127
10	2013-05-16	Star Trek Into Darkness	190000000	228778661	J.J. Abrams	PG-13	123
11	2013-11-08	Thor: The Dark World	170000000	206362140	Alan Taylor	PG-13	120
12	2013-06-21	World War Z	190000000	202359711	Marc Forster	PG-13	116
13	2013-03-22	The Croods	135000000	187168425	Kirk De MiccoChris Sanders	PG	98
14	2013-06-28	The Heat	43000000	159582188	Paul Feig	R	117
15	2013-08-07	We're the Millers	37000000	150394119	Rawson Marshall Thurber	R	110
16	2013-12-13	American Hustle	40000000	150117807	David O. Russell	R	138
17	2013-05-10	The Great Gatsby	105000000	144840419	Baz Luhrmann	PG-13	143

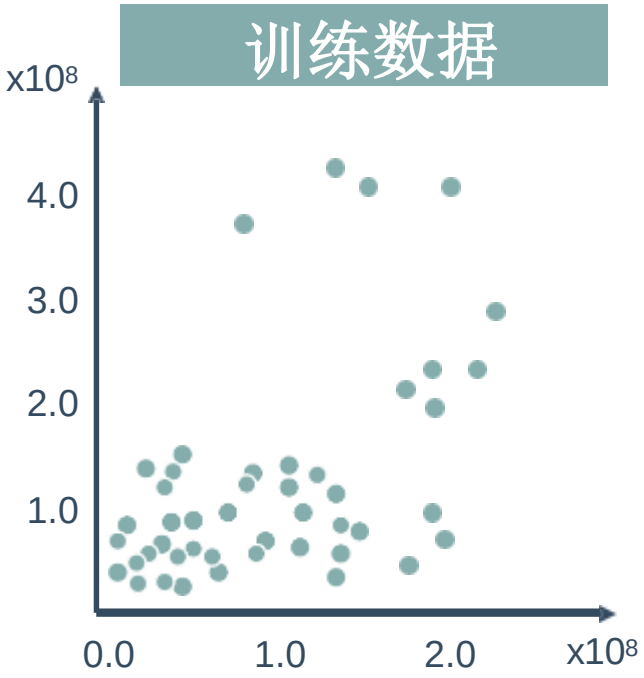
训练数据

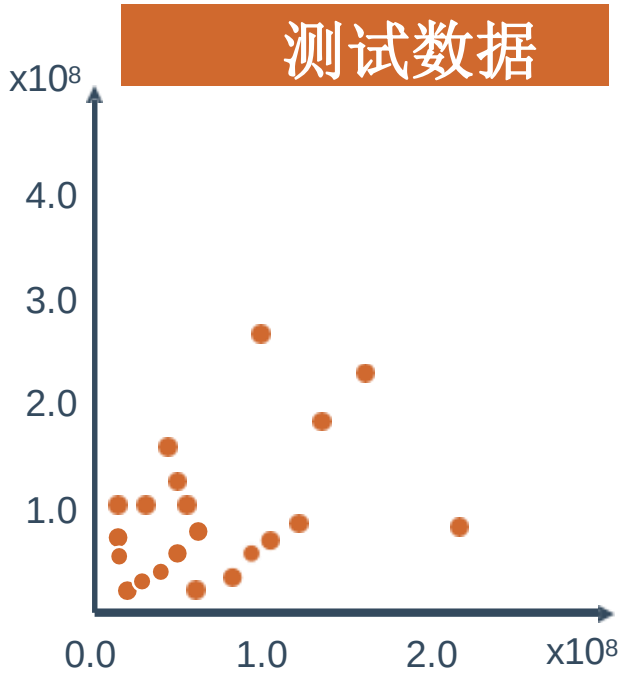
测试数据

使用训练集和测试集

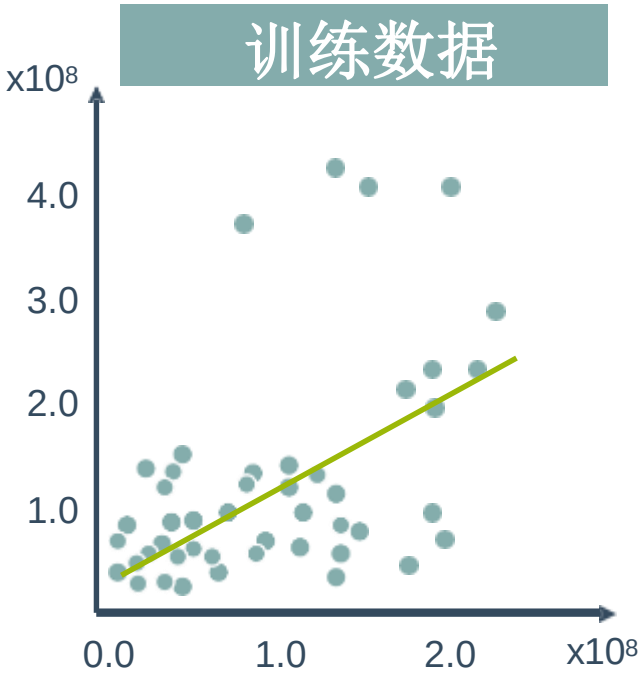
训练数据	训练模型
测试数据	评价模型 - 用模型预测类别标签 - 和真实值比较 - 计算误差

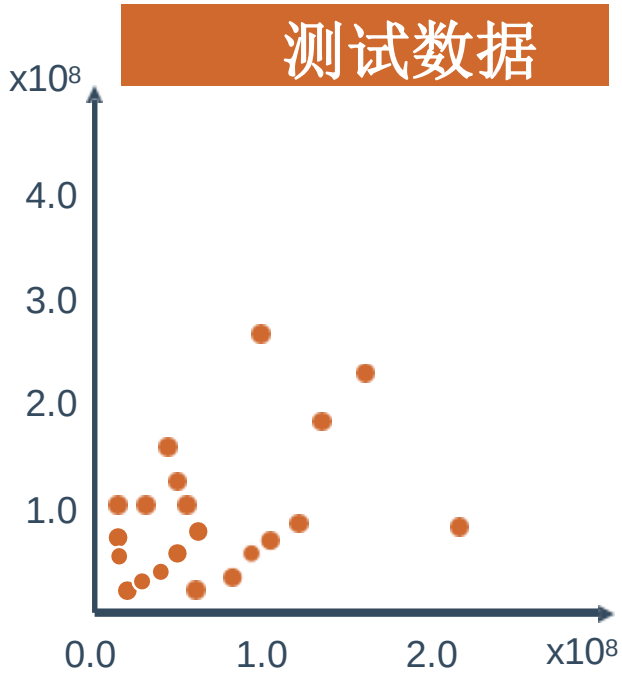
使用训练集和测试集


训练数据
x10⁸
4.0
3.0
2.0
1.0
0.0
1.0
2.0
x10⁸


测试数据
x10⁸
4.0
3.0
2.0
1.0
0.0
1.0
2.0
x10⁸

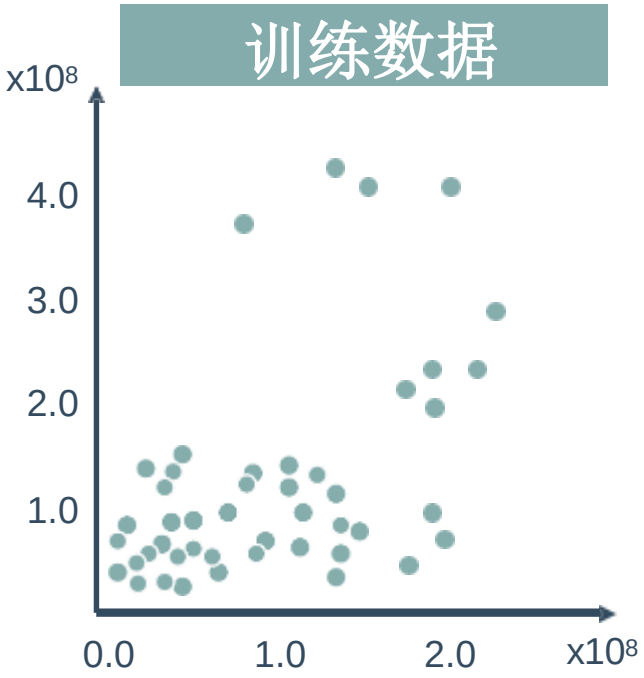
使用训练集和测试集

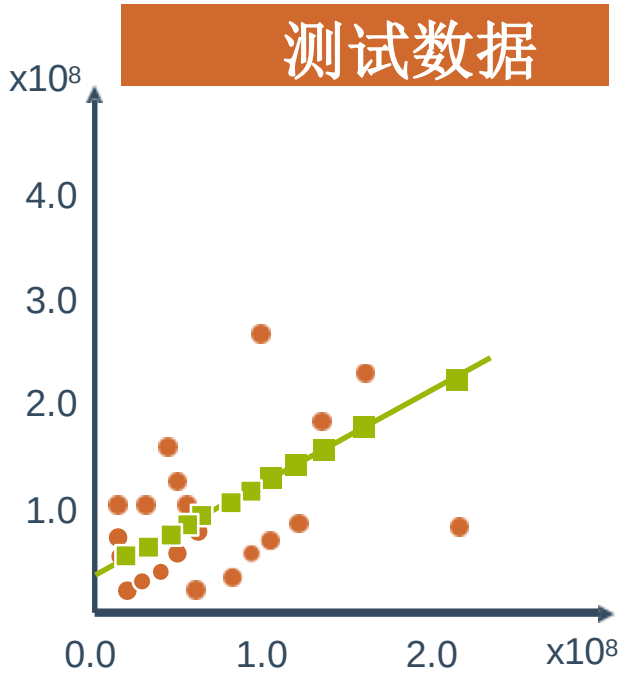




训练模型

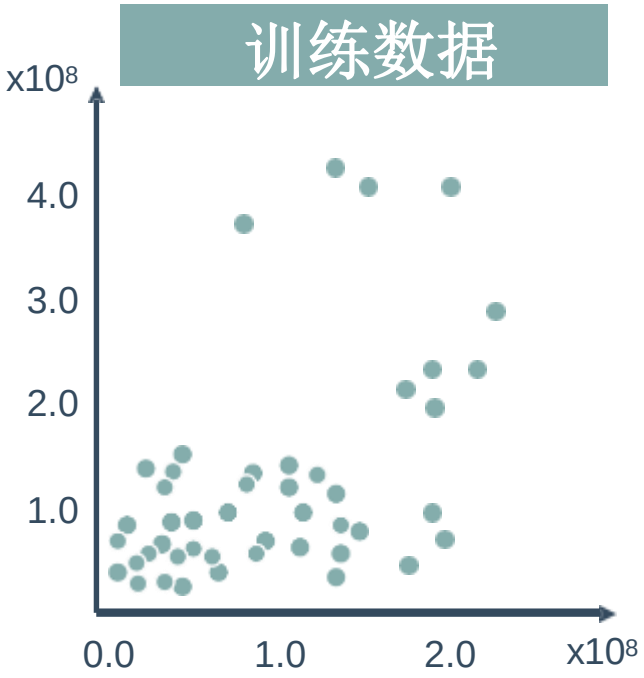
使用训练集和测试集

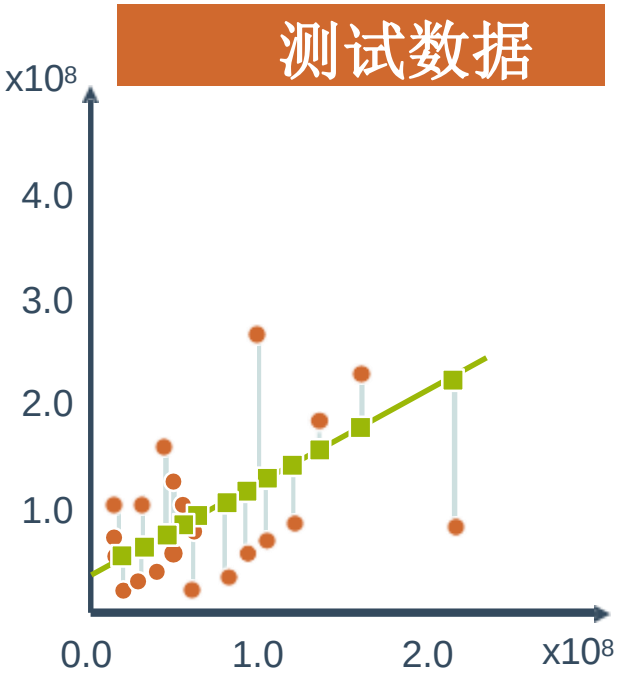




预测

使用训练集和测试集





计算误差
（或精度）

- 训练误差（training error）

  也称经验误差（empirical error） 、近似误差（approximation error），是对现有训练集的训练误差，对应训练集数据。

  - 学习到的模型

$$Y = \hat{f}(X)$$

  - 训练集

$$T = \{(x_1, y_1), (x_2, y_2) \cdots, (x_N, y_N)\}$$

  - 训练误差

$$R_{emp}(\hat{f}) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, \hat{f}(x_i))$$

- 测试误差（test error）

  也称泛化误差（generalization error）、估计误差（estimation error），对测试集的测试误差，在未知样本上的误差，对应测试集数据。

  - 学习到的模型

$$Y = \hat{f}(X)$$

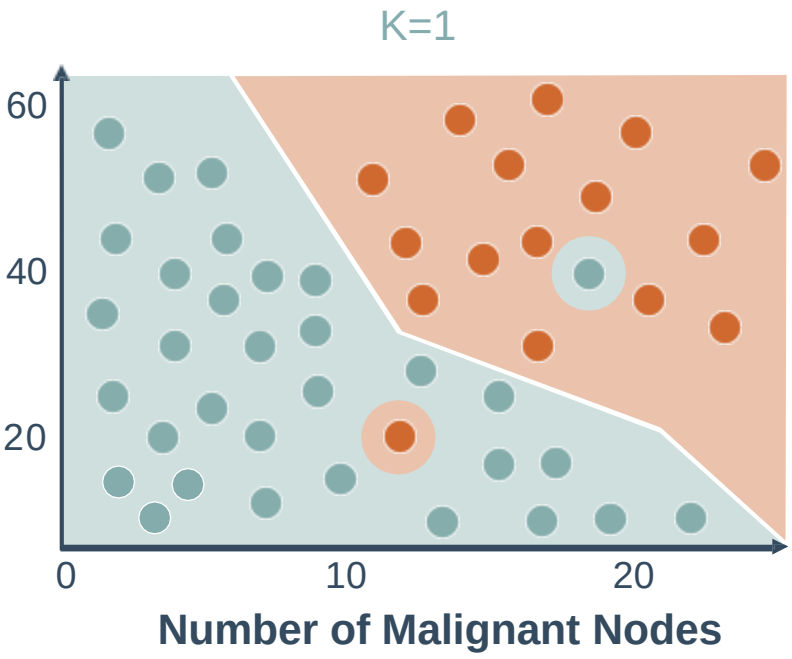
  - 测试集

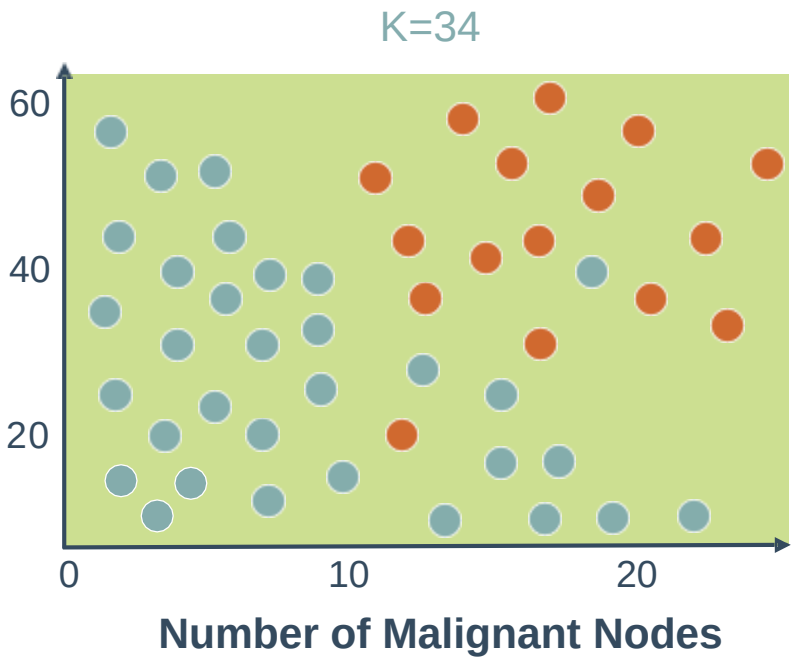
$$T' = \{(x_{1'}, y_{1'}), (x_{2'}, y_{2'}) \cdots, (x_{N'}, y_{N'})\}$$

  - 测试误差

$$e_{test} = \frac{1}{N'} \sum_{i'=1}^{N'} L(y_{i'}, \hat{f}(x_{i'}))$$

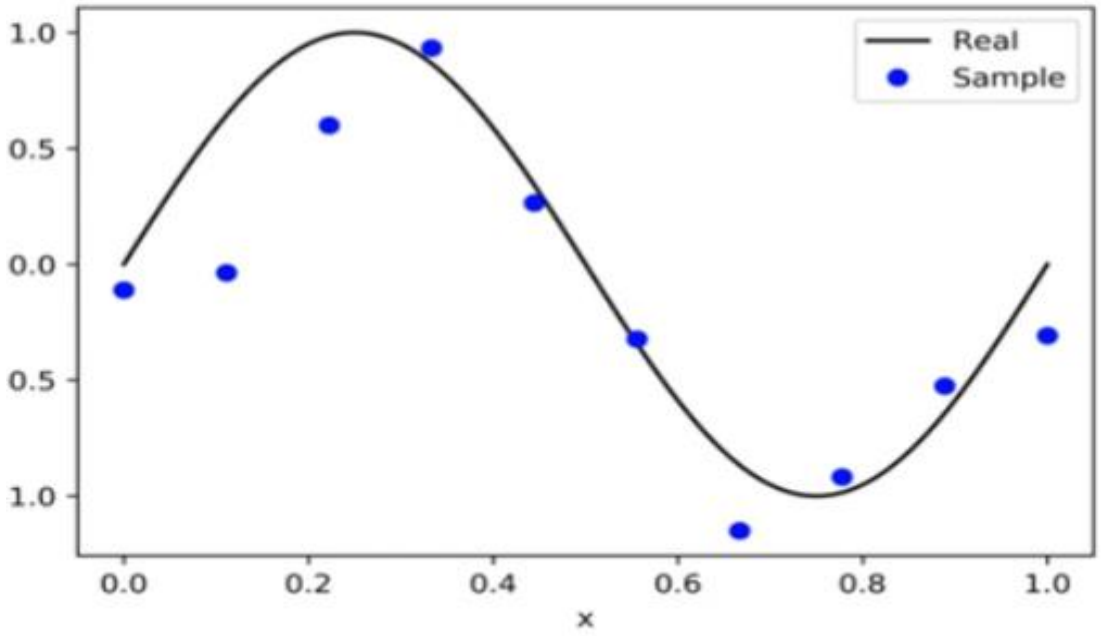
K值会影响判定边界


K=1
60
40
20
0
10
20
Number of Malignant Nodes


K=34
60
40
20
0
10
20
Number of Malignant Nodes

例: 函数为 $y = \sin(2\pi x)$，样本为 $y_i = \sin(2\pi x_i) + \varepsilon_i$，训练集为

$$T = \{(x_1, y_1), (x_2, y_2) \cdots, (x_{10}, y_{10})\}$$



M 次多项式：

$$f_M(x, w) = w_0 + w_1 x + w_2 x^2 + \cdots + w_M x^M = \sum_{j=0}^{M} w_j x^j$$

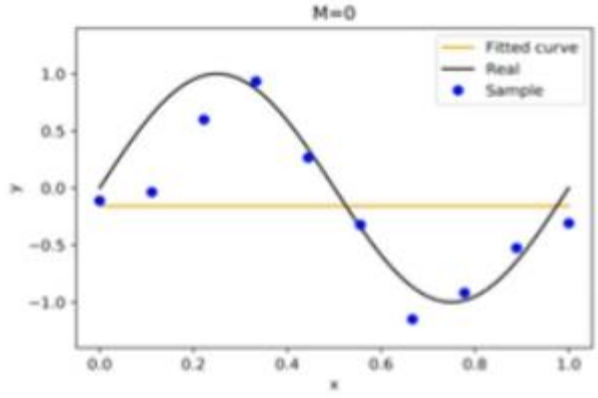
经验风险：

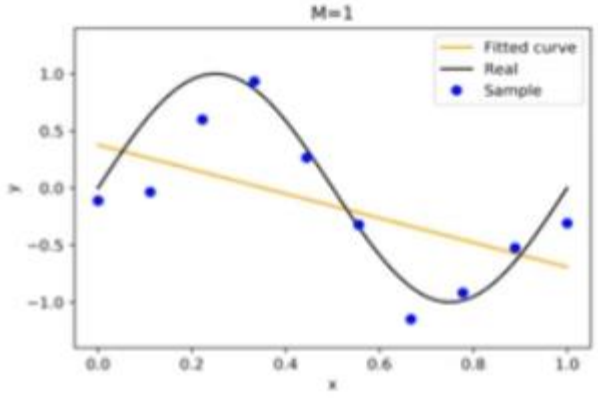
$$L(w) = \frac{1}{2} \sum_{i=1}^{N} (f_M(x_i, w) - y_i)^2$$

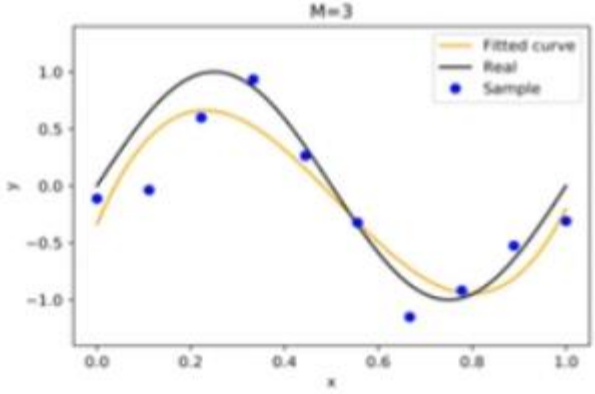
代入多项式：

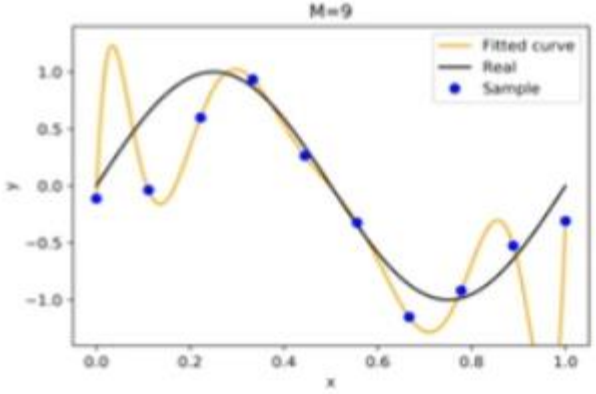
$$L(w) = \frac{1}{2} \sum_{i=1}^{N} (\sum_{j=0}^{M} w_j x_i^j - y_i)^2$$

通过最小二乘法求解参数

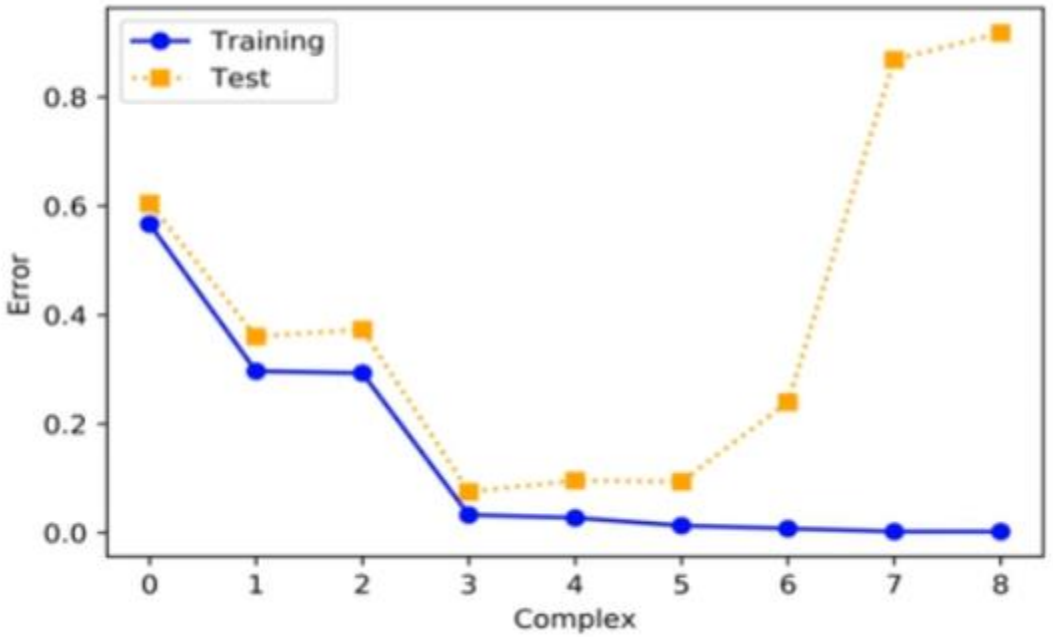

M=0
Fitted curve
Real
Sample
1.0
0.5
0.0
−0.5
−1.0
y
0.0
0.2
0.4
0.6
0.8
1.0
x


M=1
Fitted curve
Real
Sample
1.0
0.5
0.0
−0.5
−1.0
y
0.0
0.2
0.4
0.6
0.8
1.0
x

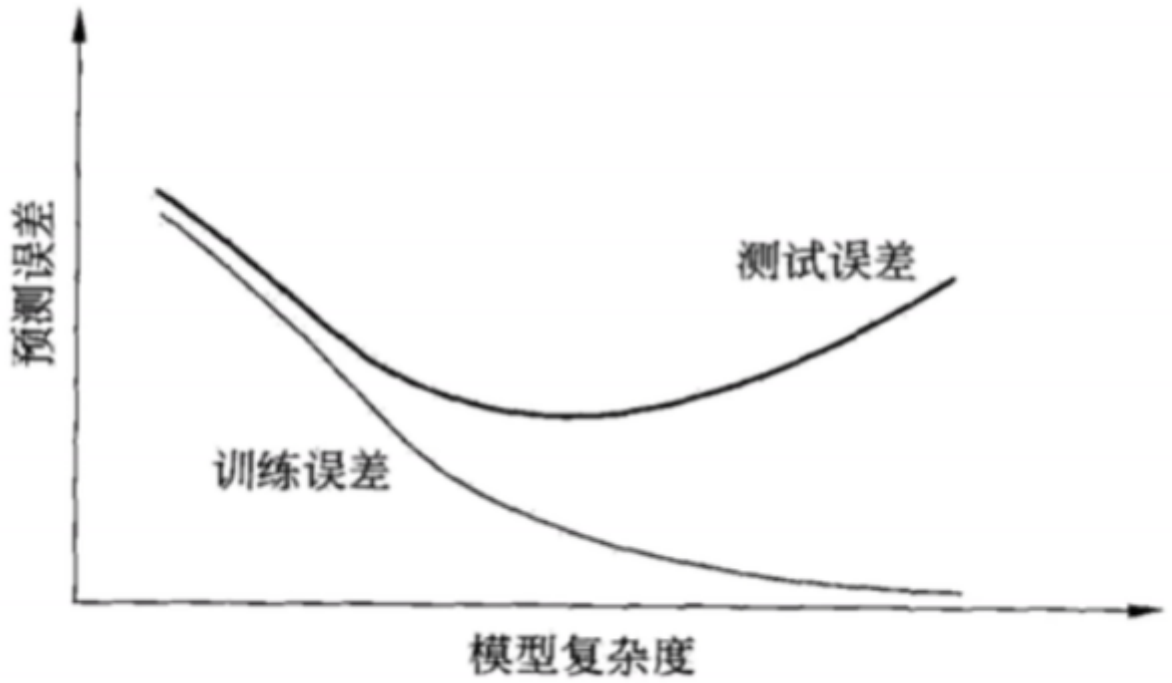

M=3
Fitted curve
Real
Sample
1.0
0.5
0.0
−0.5
−1.0
y
0.0
0.2
0.4
0.6
0.8
1.0
x


M=9
Fitted curve
Real
Sample
1.0
0.5
0.0
−0.5
−1.0
y
0.0
0.2
0.4
0.6
0.8
1.0
x

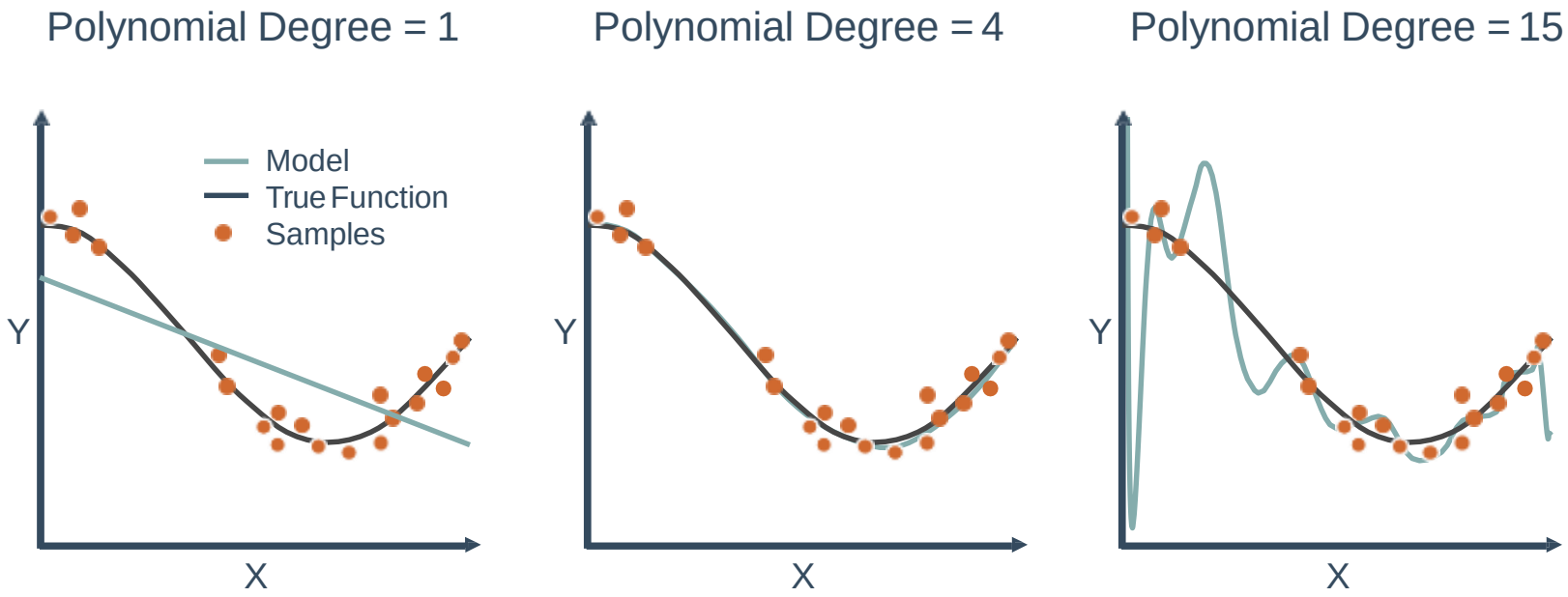
不同复杂度的模型



不同复杂度的模型

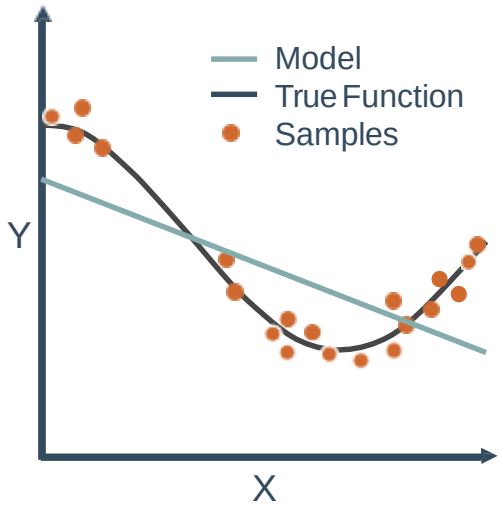


不同复杂度的模型


Polynomial Degree = 1
Polynomial Degree = 4
Polynomial Degree = 15
Model
True Function
Samples
Y
X
Y
X
Y
X

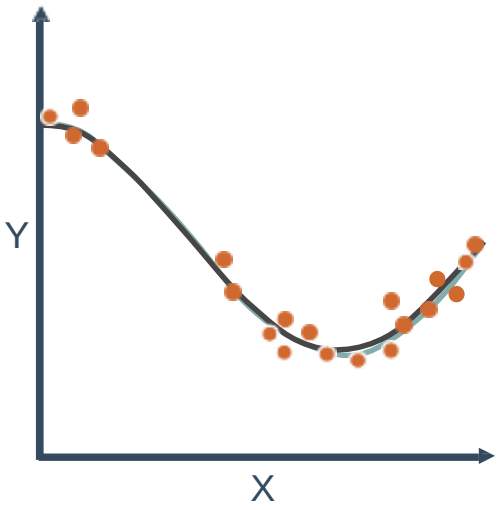
不同模型的泛化能力

Polynomial Degree = 1



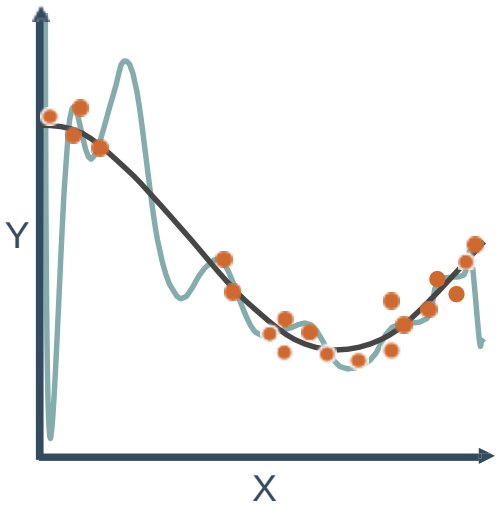
Poor at Training
Poor at Predicting

Polynomial Degree = 4



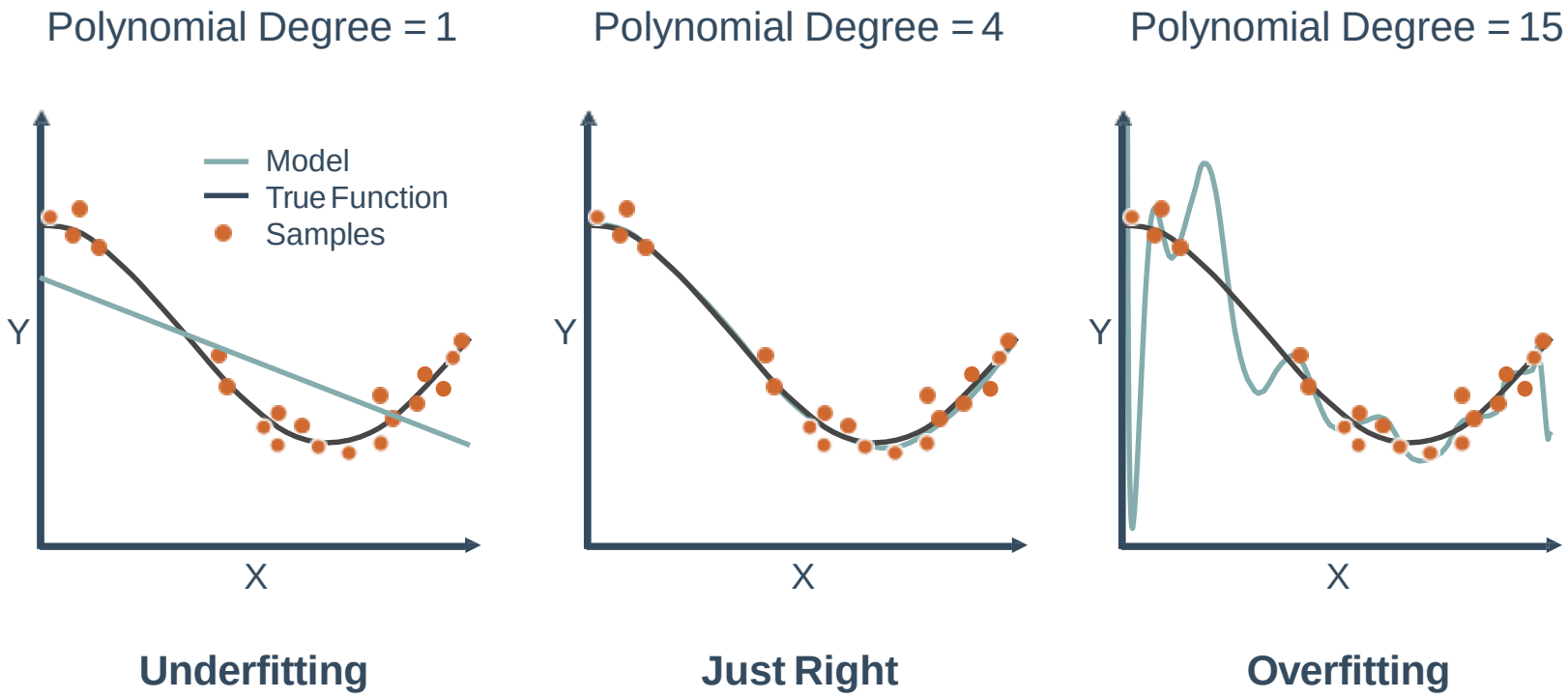
Just Right

Polynomial Degree = 15



Good at Training
Poor at Predicting

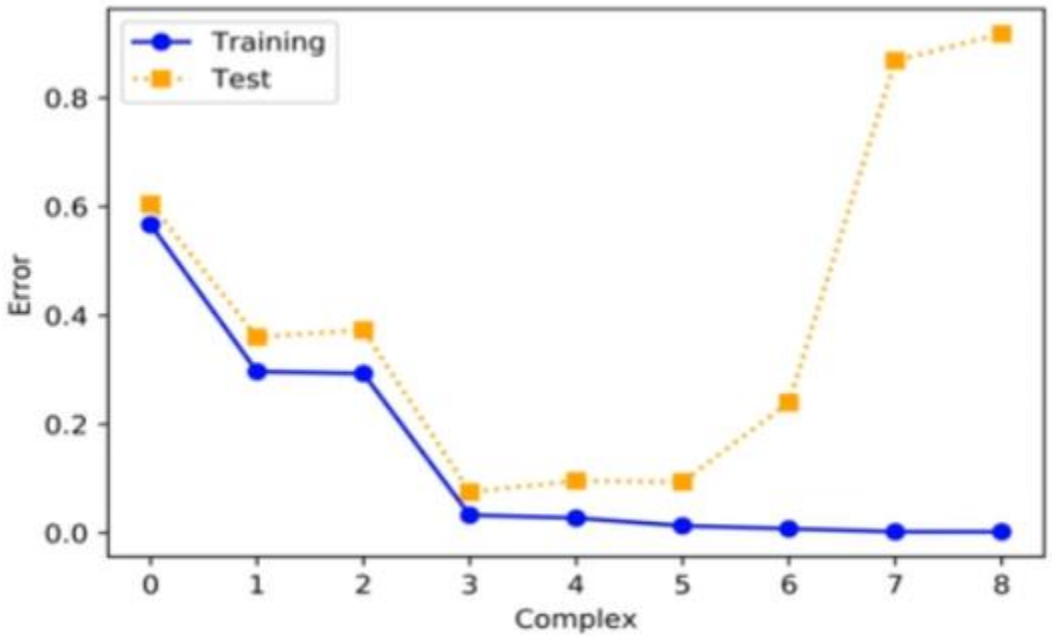
欠拟合与过拟合



欠拟合和过拟合都会导致较大的泛化误差。

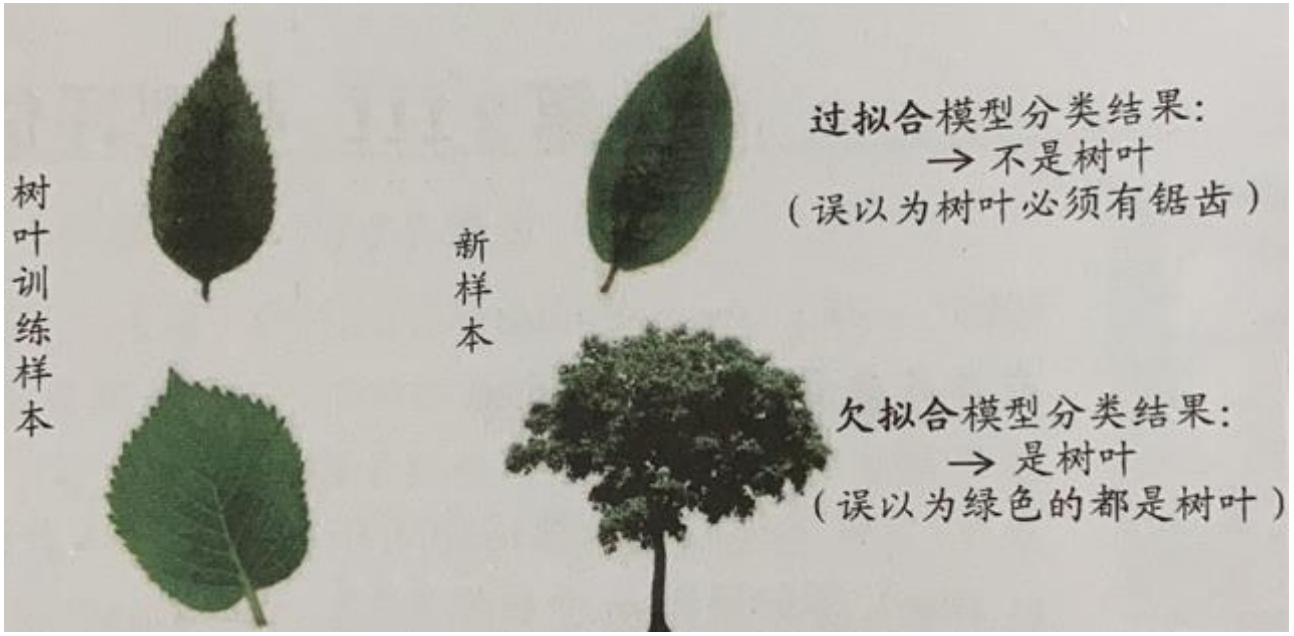
- 过拟合（overfitting）

  学习所得模型包含参数过多，出现对已知数据预测很好，但对未知数据预测很差的现象。



- 欠拟合（underfitting）

对训练样本的一般性质尚未学习好。



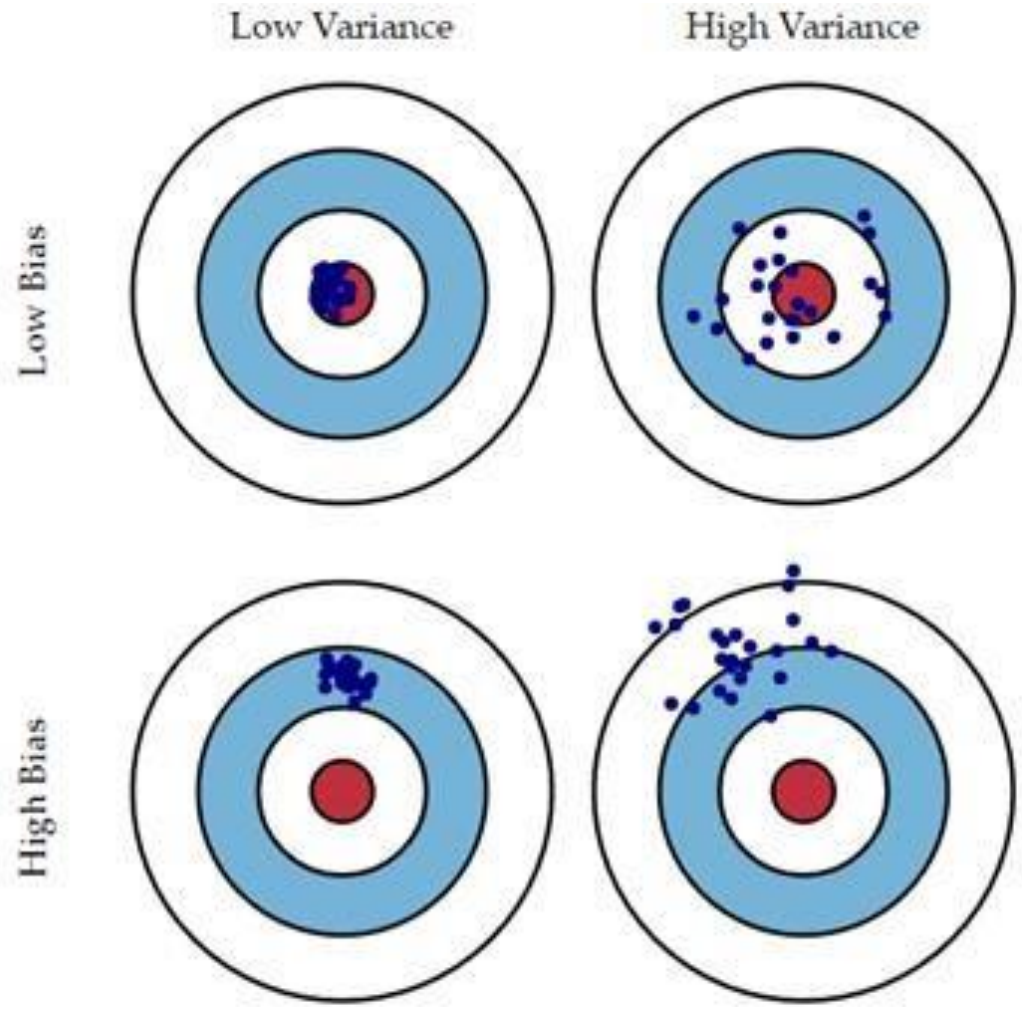
过拟合和欠拟合

监督学习中的误差来源

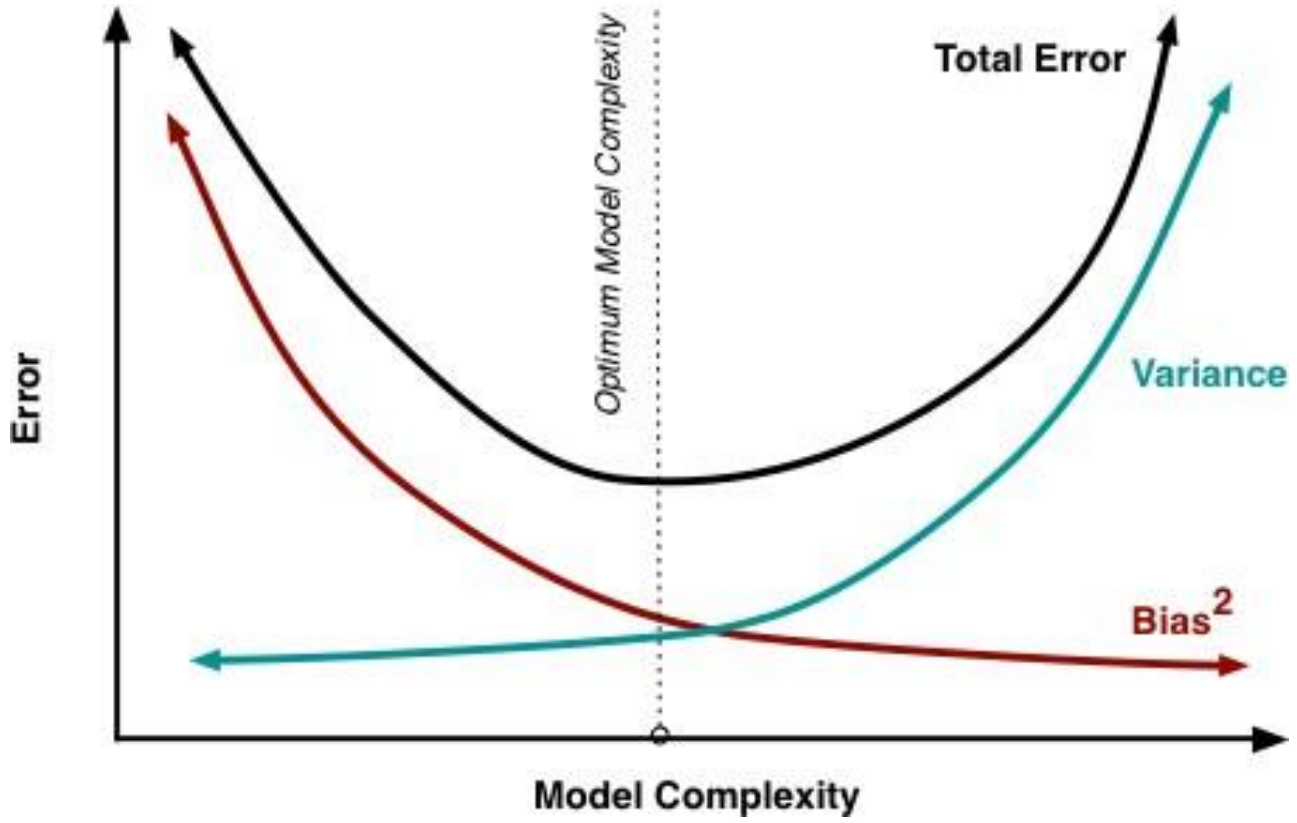
$$Error = Bias^2 + Variance + Noise$$

- **偏差**（Bias）：模型的期望输出值（即用不同数据集训练出的所有模型输出的平均值）与真实值之间的差异。即学习算法的期望预测与真实结果的偏离程度，刻画了学习算法本身的拟合能力。
- **方差**（Variance）：用不同数据集训练出的模型的输出值之间的差异。即数据的变动所导致的学习性能的变化，刻画了学习算法的稳定性。

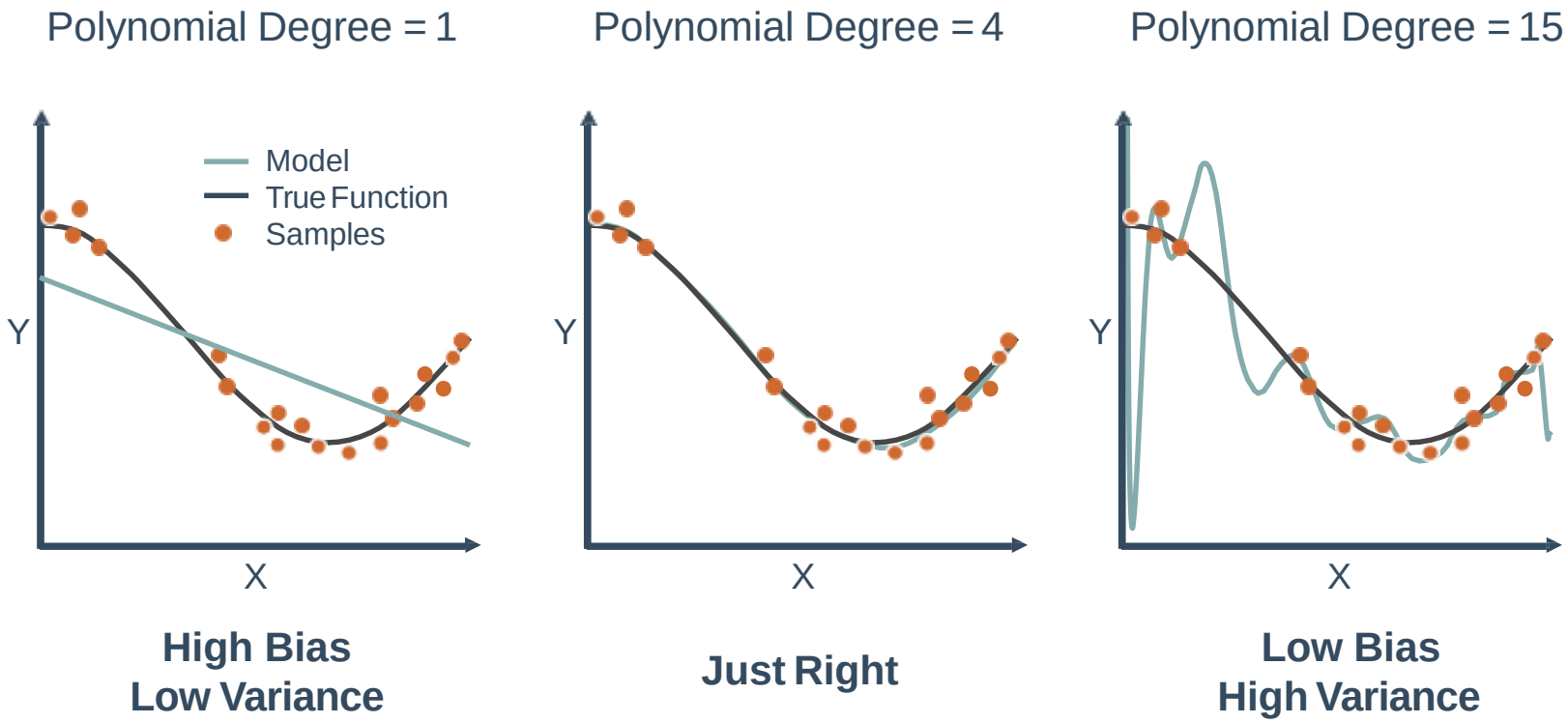
偏差与方差



偏差-方差权衡


Error
Optimum Model Complexity
Total Error
Variance
Bias^2
Model Complexity

偏差-方差权衡


Polynomial Degree = 1
Polynomial Degree = 4
Polynomial Degree = 15
Model
True Function
Samples
Y
X
Y
X
Y
X
High Bias
Low Variance
Just Right
Low Bias
High Variance

常用的模型评估与选择方法

- 正则化
- 留出法
- 交叉验证法
- 自助法

- 正则化（regularization）

  实现结构风险最小化策略。

  - 一般形式：

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f)$$

  - 经验风险：

$$\frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i))$$

  - 正则化项：

$$\lambda J(f)$$

    其中，λ 权衡经验风险和模型复杂度

- 正则化项
  - L_1 范数：

$$L(w) = \frac{1}{N}\sum_{i=1}^{N}(f(x_i; w) - y_i)^2 + \lambda||w||_1$$

其中，$||w||_1 = \sum_j |w_j|$

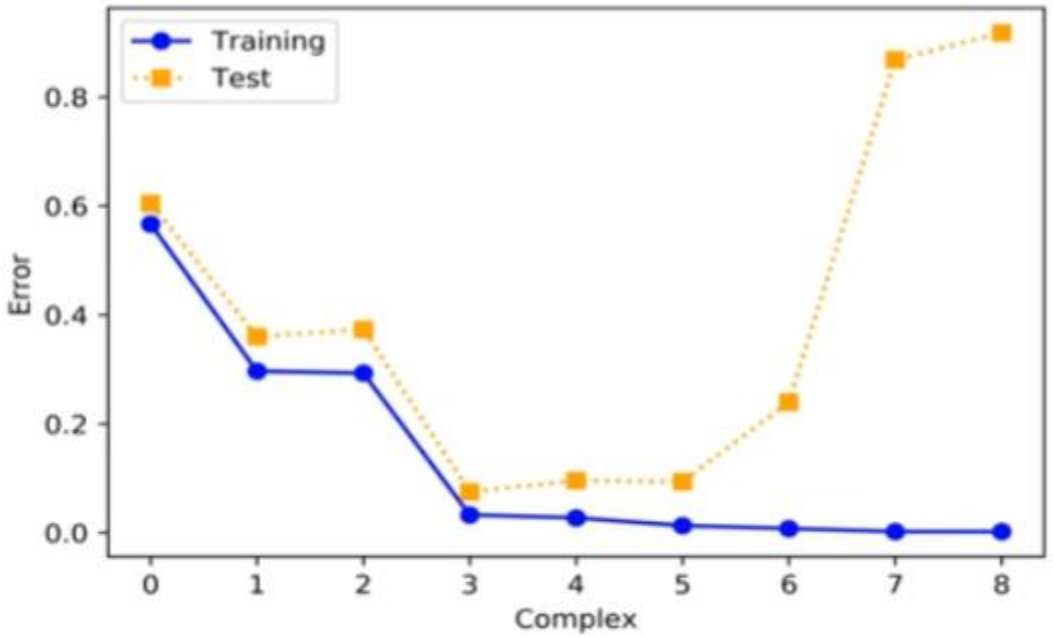
  - L_2 范数：

$$L(w) = \frac{1}{N}\sum_{i=1}^{N}(f(x_i; w) - y_i)^2 + \frac{\lambda}{2}||w||_2^2$$

其中，$||w||_2 = \sqrt{\sum_j w_j^2}$，$||w||_2^2 = \sum_j w_j^2$

- 奥卡姆剃刀原理

  在模型选择时，选择所有可能模型中，能很好解释已知数据并且十分简单的模型。



- 留出法（hold-out）

  直接将数据集D划分为两个互斥的部分，其中一部分作为训练集S，另一部分用作测试集T 。

  通常训练集和测试集的比例为0.7：0.3。划分时注意：

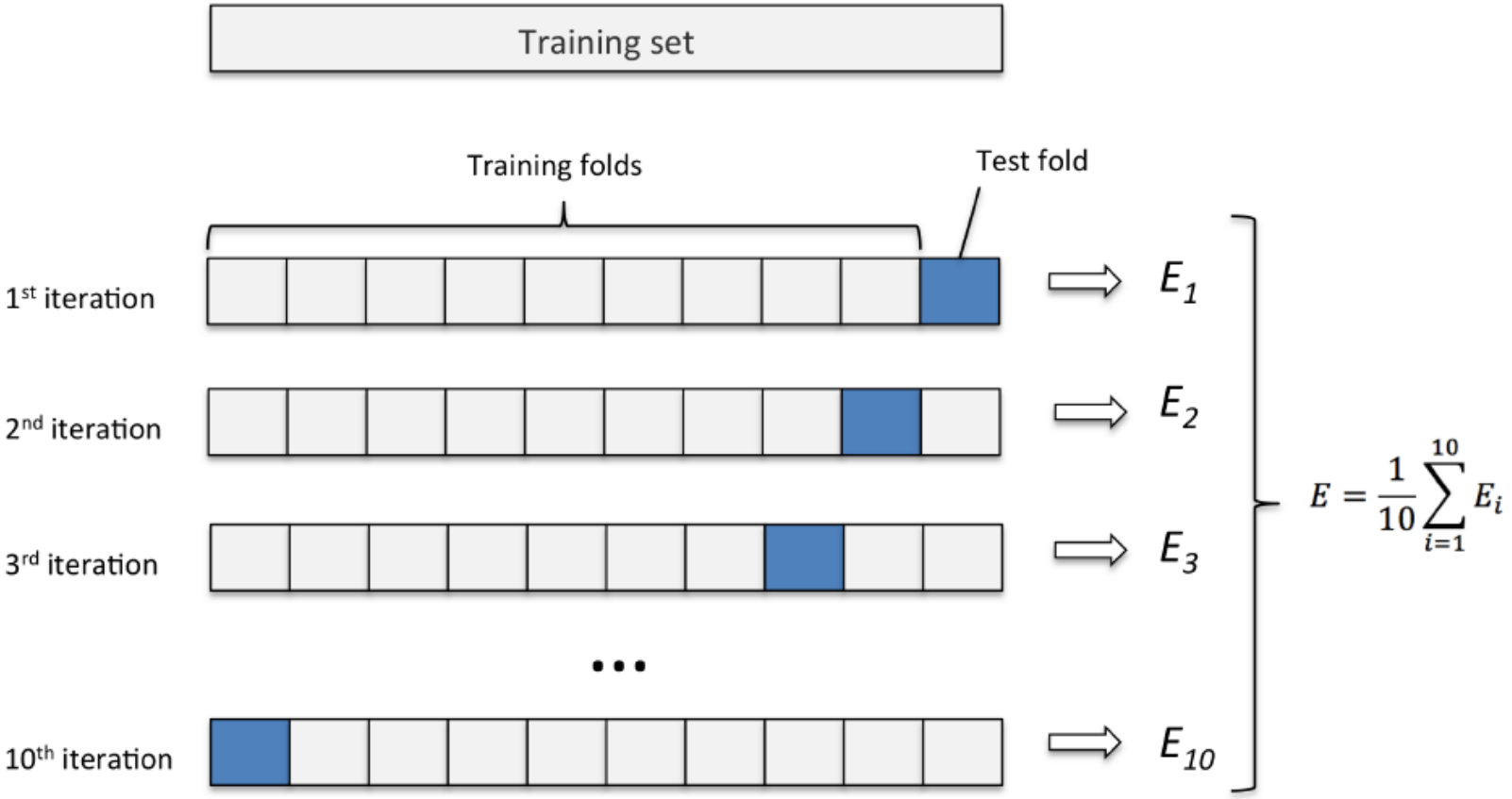
  - 尽可能保持数据分布的一致性。避免因数据划分过程引入的额外偏差而对最终结果产生影响。
  - 采用若干次随机划分避免单次使用留出法的不稳定性。

- 交叉验证法（ cross validation ）

先将数据集划分为k个大小相似的互斥子集，每次采用k-1个子集的并集作为训练集，剩下的那个子集作为测试集。进行k次训练和测试，最终返回k个测试结果的均值。又称为“k折交叉验证”（k-fold cross validation）。

为减少因样本划分带来的偏差，通常重复p次不同的划分，最终结果是p次k折交叉验证结果的均值。

超越单个测试集：交叉验证



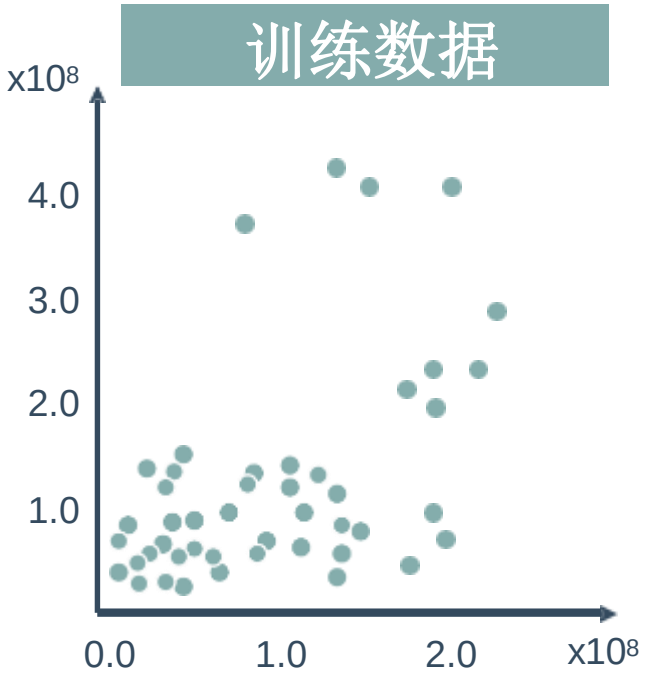
超越单个测试集：交叉验证

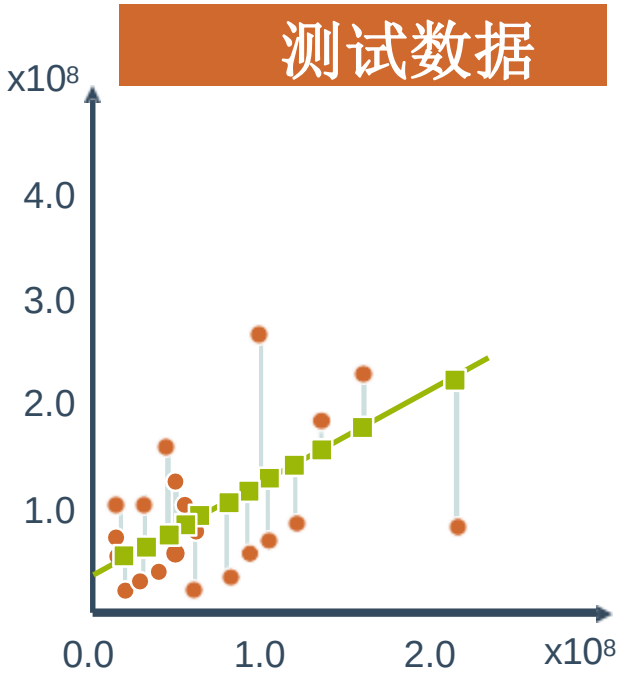
	Date	Title	Budget	DomesticTotalGross	Director	Rating	Runtime
0	2013-11-22	The Hunger Games: Catching Fire	130000000	424668047	Francis Lawrence	PG-13	146
1	2013-05-03	Iron Man 3	200000000	409013994	Shane Black	PG-13	129
2	2013-11-22	Frozen	150000000	400738009	Chris BuckJennifer Lee	PG	108
3	2013-07-03	Despicable Me 2	76000000	368061265	Pierre CoffinChris Renaud	PG	98
4	2013-06-14	Man of Steel	225000000	291045518	Zack Snyder	PG-13	143
5	2013-10-04	Gravity	100000000	274092705	Alfonso Cuaron	PG-13	91
6	2013-06-21	Monsters University	NaN	268492764	Dan Scanlon	G	107
7	2013-12-13	The Hobbit: The Desolation of Smaug	NaN	258366855	Peter Jackson	PG-13	161
8	2013-05-24	Fast & Furious 6	160000000	238679850	Justin Lin	PG-13	130
9	2013-03-08	Oz The Great and Powerful	215000000	234911825	Sam Raimi	PG	127
10	2013-05-16	Star Trek Into Darkness	190000000	228778661	J.J. Abrams	PG-13	123
11	2013-11-08	Thor: The Dark World	170000000	206362140	Alan Taylor	PG-13	120
12	2013-06-21	World War Z	190000000	202359711	Marc Forster	PG-13	116
13	2013-03-22	The Croods	135000000	187168425	Kirk De MiccoChris Sanders	PG	98
14	2013-06-28	The Heat	43000000	159582188	Paul Feig	R	117
15	2013-08-07	We're the Millers	37000000	150394119	Rawson Marshall Thurber	R	110
16	2013-12-13	American Hustle	40000000	150117807	David O. Russell	R	138
17	2013-05-10	The Great Gatsby	105000000	144840419	Baz Luhrmann	PG-13	143

训练数据

验证数据

超越单个测试集：交叉验证





对这个测试集的最优模型

超越单个测试集：交叉验证

	Date	Title	Budget	DomesticTotalGross	Director	Rating	Runtime
0	2013-11-22	The Hunger Games: Catching Fire	130000000	424668047	Francis Lawrence	PG-13	146
1	2013-05-03	Iron Man 3	200000000	409013994	Shane Black	PG-13	129
2	2013-11-22	Frozen	150000000	400738009	Chris BuckJennifer Lee	PG	108
3	2013-07-03	Despicable Me 2	76000000	368061265	Pierre CoffinChris Renaud	PG	98
4	2013-06-14	Man of Steel	225000000	291045518	Zack Snyder	PG-13	143
5	2013-10-04	Gravity	100000000	274092705	Alfonso Cuaron	PG-13	91
6	2013-06-21	Monsters University	NaN	268492764	Dan Scanlon	G	107
7	2013-12-13	The Hobbit: The Desolation of Smaug	NaN	258366855	Peter Jackson	PG-13	161
8	2013-05-24	Fast & Furious 6	160000000	238679850	Justin Lin	PG-13	130
9	2013-03-08	Oz The Great and Powerful	215000000	234911825	Sam Raimi	PG	127
10	2013-05-16	Star Trek Into Darkness	190000000	228778661	J.J. Abrams	PG-13	123
11	2013-11-08	Thor: The Dark World	170000000	206362140	Alan Taylor	PG-13	120
12	2013-06-21	World War Z	190000000	202359711	Marc Forster	PG-13	116
13	2013-03-22	The Croods	135000000	187168425	Kirk De MiccoChris Sanders	PG	98
14	2013-06-28	The Heat	43000000	159582188	Paul Feig	R	117
15	2013-08-07	We're the Millers	37000000	150394119	Rawson Marshall Thurber	R	110
16	2013-12-13	American Hustle	40000000	150117807	David O. Russell	R	138
17	2013-05-10	The Great Gatsby	105000000	144840419	Baz Luhrmann	PG-13	143

训练数据1

验证数据1

超越单个测试集：交叉验证

	Date	Title	Budget	DomesticTotalGross	Director	Rating	Runtime
0	2013-11-22	The Hunger Games: Catching Fire	130000000	424668047	Francis Lawrence	PG-13	146
1	2013-05-03	Iron Man 3	200000000	409013994	Shane Black	PG-13	129
2	2013-11-22	Frozen	150000000	400738009	Chris BuckJennifer Lee	PG	108
3	2013-07-03	Despicable Me 2	76000000	368061265	Pierre CoffinChris Renaud	PG	98
4	2013-06-14	Man of Steel	225000000	291045518	Zack Snyder	PG-13	143
5	2013-10-04	Gravity	100000000	274092705	Alfonso Cuaron	PG-13	91
6	2013-06-21	Monsters University	NaN	268492764	Dan Scanlon	G	107
7	2013-12-13	The Hobbit: The Desolation of Smaug	NaN	258366855	Peter Jackson	PG-13	161
8	2013-05-24	Fast & Furious 6	160000000	238679850	Justin Lin	PG-13	130
9	2013-03-08	Oz The Great and Powerful	215000000	234911825	Sam Raimi	PG	127
10	2013-05-16	Star Trek Into Darkness	190000000	228778661	J.J. Abrams	PG-13	123
11	2013-11-08	Thor: The Dark World	170000000	206362140	Alan Taylor	PG-13	120
12	2013-06-21	World War Z	190000000	202359711	Marc Forster	PG-13	116
13	2013-03-22	The Croods	135000000	187168425	Kirk De MiccoChris Sanders	PG	98
14	2013-06-28	The Heat	43000000	159582188	Paul Feig	R	117
15	2013-08-07	We're the Millers	37000000	150394119	Rawson Marshall Thurber	R	110
16	2013-12-13	American Hustle	40000000	150117807	David O. Russell	R	138
17	2013-05-10	The Great Gatsby	105000000	144840419	Baz Luhrmann	PG-13	143

训练数据2

验证数据2

超越单个测试集：交叉验证

	Date	Title	Budget	DomesticTotalGross	Director	Rating	Runtime
0	2013-11-22	The Hunger Games: Catching Fire	130000000	424668047	Francis Lawrence	PG-13	146
1	2013-05-03	Iron Man 3	200000000	409013994	Shane Black	PG-13	129
2	2013-11-22	Frozen	150000000	400738009	Chris BuckJennifer Lee	PG	108
3	2013-07-03	Despicable Me 2	76000000	368061265	Pierre CoffinChris Renaud	PG	98
4	2013-06-14	Man of Steel	225000000	291045518	Zack Snyder	PG-13	143
5	2013-10-04	Gravity	100000000	274092705	Alfonso Cuaron	PG-13	91
6	2013-06-21	Monsters University	NaN	268492764	Dan Scanlon	G	107
7	2013-12-13	The Hobbit: The Desolation of Smaug	NaN	258366855	Peter Jackson	PG-13	161
8	2013-05-24	Fast & Furious 6	160000000	238679850	Justin Lin	PG-13	130
9	2013-03-08	Oz The Great and Powerful	215000000	234911825	Sam Raimi	PG	127
10	2013-05-16	Star Trek Into Darkness	190000000	228778661	J.J. Abrams	PG-13	123
11	2013-11-08	Thor: The Dark World	170000000	206362140	Alan Taylor	PG-13	120
12	2013-06-21	World War Z	190000000	202359711	Marc Forster	PG-13	116
13	2013-03-22	The Croods	135000000	187168425	Kirk De MiccoChris Sanders	PG	98
14	2013-06-28	The Heat	43000000	159582188	Paul Feig	R	117
15	2013-08-07	We're the Millers	37000000	150394119	Rawson Marshall Thurber	R	110
16	2013-12-13	American Hustle	40000000	150117807	David O. Russell	R	138
17	2013-05-10	The Great Gatsby	105000000	144840419	Baz Luhrmann	PG-13	143

验证数据3

训练数据3

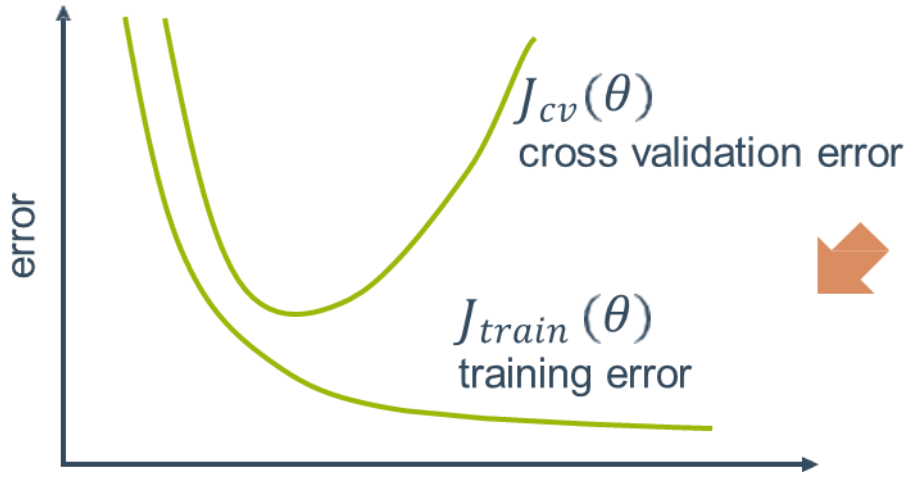
超越单个测试集：交叉验证

	Date	Title	Budget	DomesticTotalGross	Director	Rating	Runtime
0	2013-11-22	The Hunger Games: Catching Fire	130000000	424668047	Francis Lawrence	PG-13	146
1	2013-05-03	Iron Man 3	200000000	409013994	Shane Black	PG-13	129
2	2013-11-22	Frozen	150000000	400738009	Chris BuckJennifer Lee	PG	108
3	2013-07-03	Despicable Me 2	76000000	368061265	Pierre CoffinChris Renaud	PG	98
4	2013-06-14	Man of Steel	225000000	291045518	Zack Snyder	PG-13	143
5	2013-10-04	Gravity	100000000	274092705	Alfonso Cuaron	PG-13	91
6	2013-06-21	Monsters University	NaN	268492764	Dan Scanlon	G	107
7	2013-12-13	The Hobbit: The Desolation of Smaug	NaN	258366855	Peter Jackson	PG-13	161
8	2013-05-24	Fast & Furious 6	160000000	238679850	Justin Lin	PG-13	130
9	2013-03-08	Oz The Great and Powerful	215000000	234911825	Sam Raimi	PG	127
10	2013-05-16	Star Trek Into Darkness	190000000	228778661	J.J. Abrams	PG-13	123
11	2013-11-08	Thor: The Dark World	170000000	206362140	Alan Taylor	PG-13	120
12	2013-06-21	World War Z	190000000	202359711	Marc Forster	PG-13	116
13	2013-03-22	The Croods	135000000	187168425	Kirk De MiccoChris Sanders	PG	98
14	2013-06-28	The Heat	43000000	159582188	Paul Feig	R	117
15	2013-08-07	We're the Millers	37000000	150394119	Rawson Marshall Thurber	R	110
16	2013-12-13	American Hustle	40000000	150117807	David O. Russell	R	138
17	2013-05-10	The Great Gatsby	105000000	144840419	Baz Luhrmann	PG-13	143

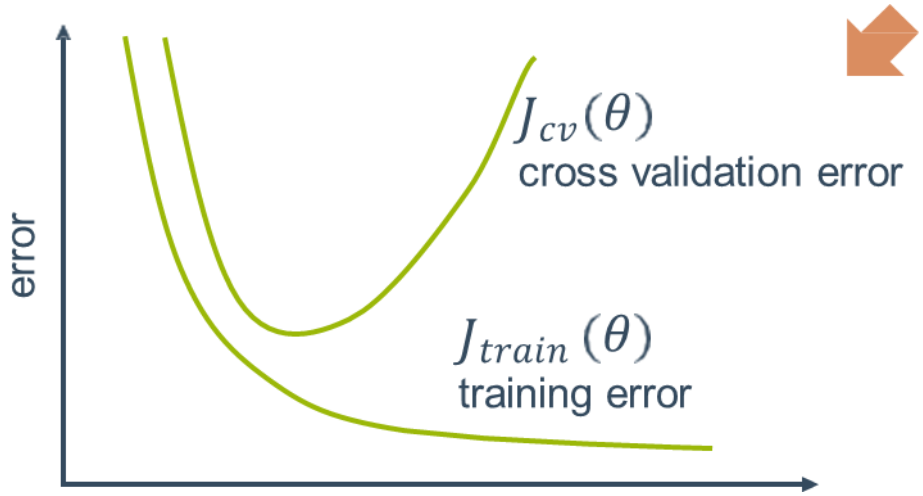
验证数据4

训练数据4

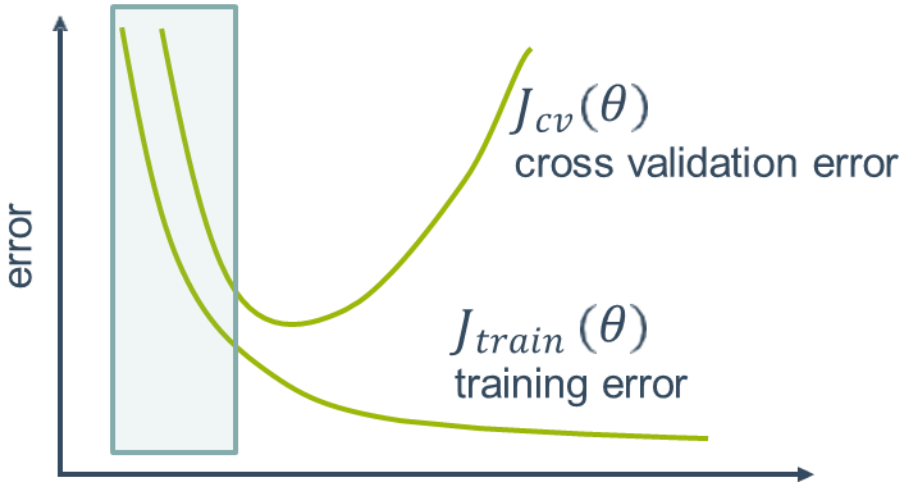
模型复杂度与误差

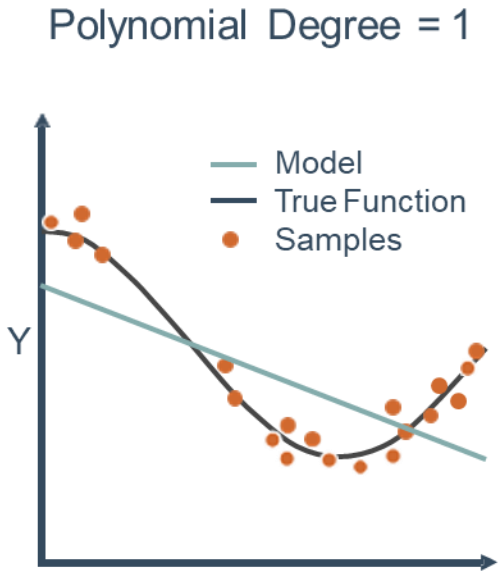

error
$J_{cv}(\theta)$
cross validation error
$J_{train}(\theta)$
training error

模型复杂度与误差


error
$J_{cv}(\theta)$
cross validation error
$J_{train}(\theta)$
training error

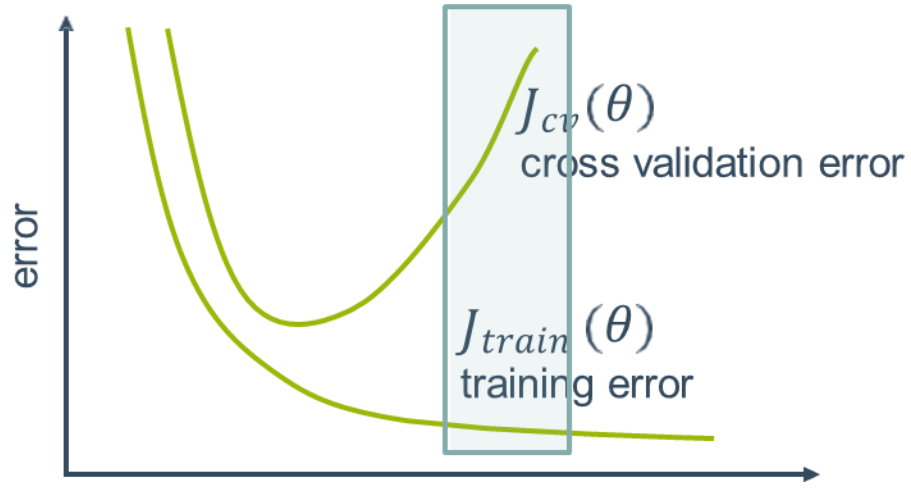
模型复杂度与误差

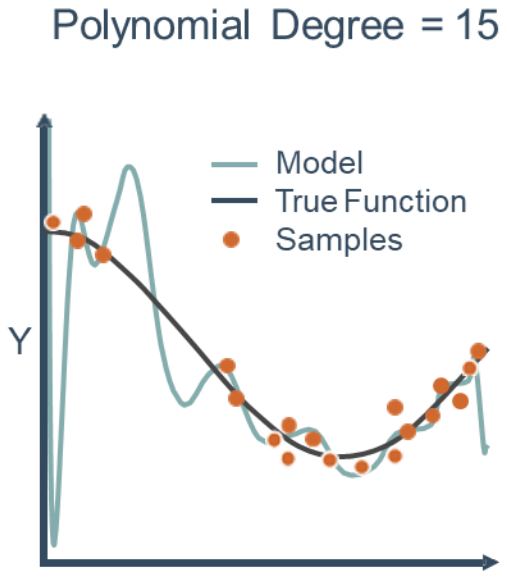




欠拟合：训练误差和交叉验证误差都很高

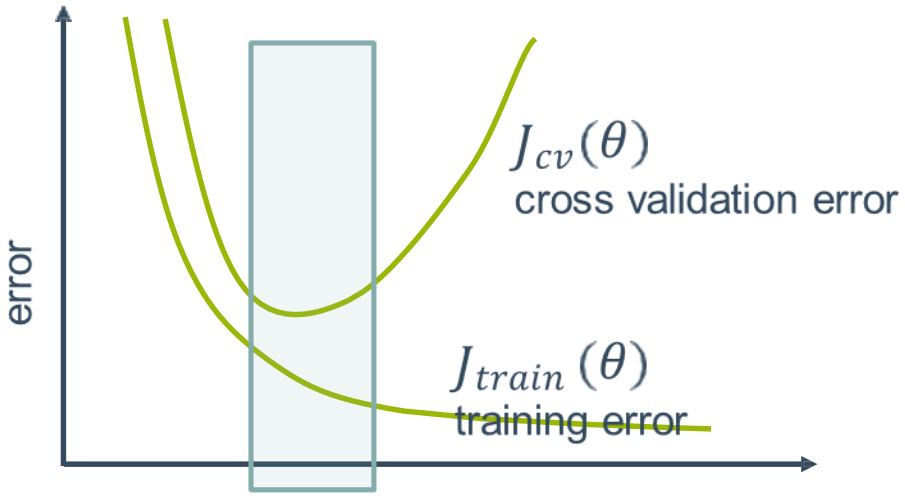
模型复杂度与误差

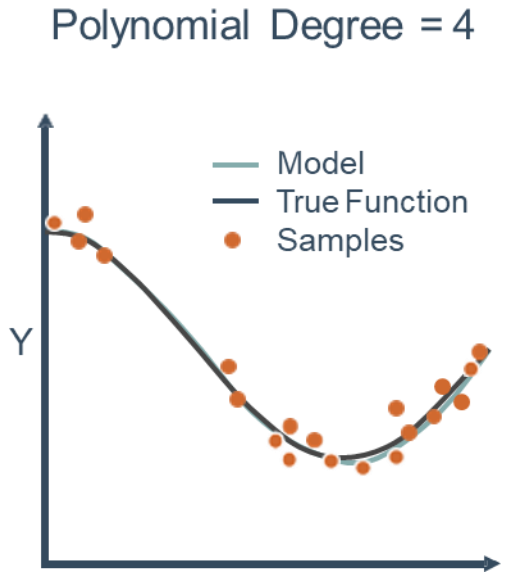




过拟合：训练误差低，交叉验证误差高

模型复杂度与误差


error
$J_{cv}(\theta)$
cross validation error
$J_{train}(\theta)$
training error


Polynomial Degree = 4
Model
True Function
Samples
Y

- 自助法（booststrapping）

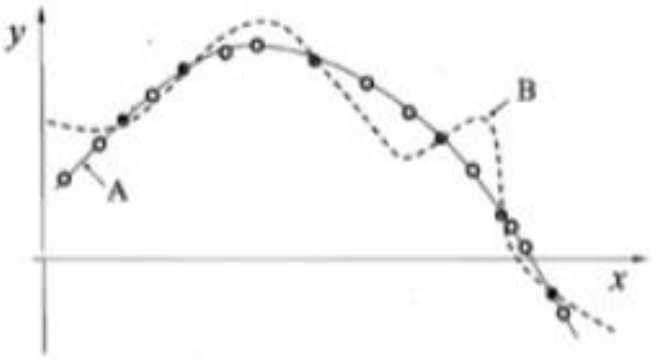
以自助采样为基础（有放回采样）。每次随机从D中挑选一个样本，放入D′中，然后将样本放回D中，重复m次之后，得到了包含m个样本的数据集。

样本在m次采样中始终不被采到的概率是$(1-1/m)^m$，取极限得到$\lim_{m\to\infty}(1-1/m)^m = 1/e = 0.368$。即D中约有36.8%的样本未出现在D′中。于是将D′用作训练集，D 用作测试集。这样，仍然使用m个训练样本，但约有1/3未出现在训练集中的样本被用作测试集。

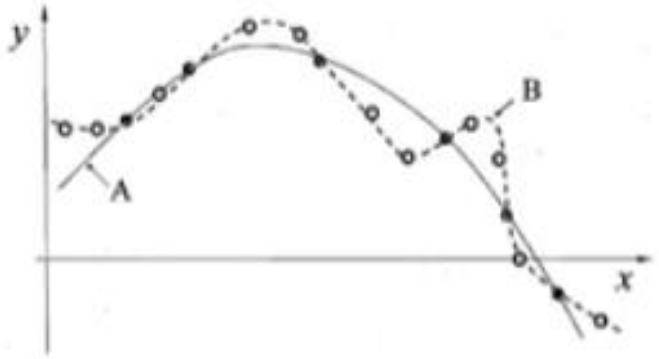
优点：自助法在数据集较小、难以有效划分训练/测试集时很有用。

缺点：然而自助法改变了初始数据集的分布，这会引入估计偏差。

模型性能度量



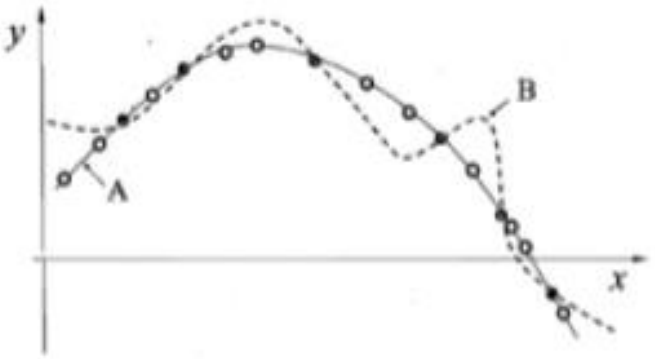
A优于B



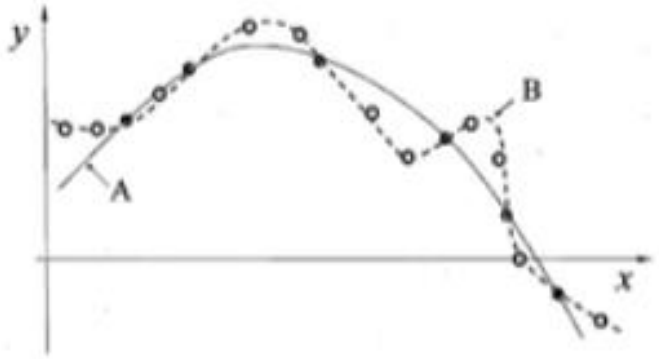
B优于A

- 模型没有高低，只有是否适合。

模型性能度量



A优于B



B优于A

- 那么如何量化模型对于问题的适应性?
  - 模型预测是否足够准确
  - 模型预测是否少犯错
  - 模型预测能力是否稳定

精度指标的局限性

- 要求你为白血病的诊断构建一个分类器
- **训练数据：** 1% 的样例患有白血病，99% 是健康的
- **评价指标是预测精度：** 即预测正确的百分比
- 那么构建一个最简单的分类器，对所有输入都回答“健康”
- 仍然可以达到99%的精度。。。

现实中样本在不同类别的分布不平衡，导致精度不能很好地反应分类器的性能

回归任务性能度量

在预测任务中，给定数据集$D=\{(x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)\}$，其中 y_i 是示例 x_i 的真实标记。要评估学习器 f 的性能，就要把学习器预测结果 $f(x)$ 与真实标记 y 进行比较。

回归任务最常用的性能度量是“均方误差”（mean squared error）：

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}(f(x_i)-y_i)^2$$

对于数据分布 $\mathcal{D}$ 和概率密度函数 $p(\cdot)$，均方误差为：

$$E(f;\mathcal{D})=\int_{x\sim\mathcal{D}}(f(x)-y)^2p(x)dx$$

分类任务性能度量

- 错误率与精度
- 查准率、查全率与$F1$分数
- ROC与AUC
- 代价敏感错误率与代价曲线

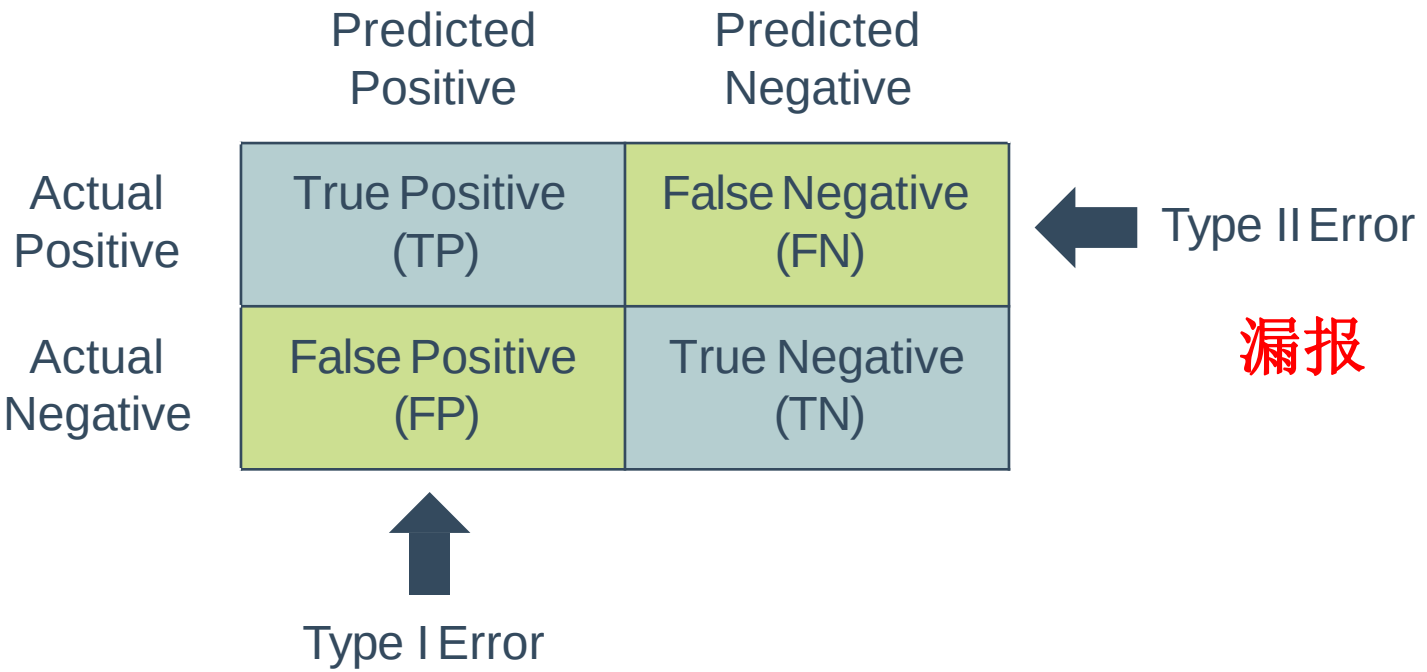
混淆矩阵

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

样本总数
正样本数
负样本数
判断为正样本数
判断为负样本数
正确分类样本数
错误分类样本数

混淆矩阵（confusion matrix）可以展示**各种类型的错误**，能更好地描述模型的性能；从混淆矩阵中可计算出多种指标。

混淆矩阵



真正率(true positive rate,TPR):
模型正确预测的正样本的比例。
TPR = TP \(TP+FN)

真负率(true negative rate,TNR)：
模型正确预测的负样本的比例 。
TNR = TN \ (TN+FP)

假正率(false positive rate, FPR)：
FPR = FP \(TN+FP)

假负率(flase negative rate, FNR)：
FNR = FN \(TP+FN)

- 错误率与精度

  分类任务中常用的性能度量，既适用于二分类，也适用于多分类。

  - 错误率（error rate）

    分类错误的样本数占样本总数的比例，对数据集 D，定义为：

$$E(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathrm{II}\left(f(x_i)\neq y_i\right)$$

  - 精度（accuracy）

    分类正确的样本数占样本总数的比例，定义为：

$$\mathrm{acc}(f;D)=\frac{1}{m}\sum_{i=1}^{m}\mathrm{II}\left(f(x_i)=y_i\right)$$

$$=1\text{-}E(f;D)$$

$\mathrm{II}(\cdot)$为指示函数，$\cdot$为真假时分别取1和0

- 错误率与精度

  分类任务中常用的性能度量，既适用于二分类，也适用于多分类。

对于数据分布 $\mathcal{D}$ 和概率密度函数 $p(\cdot)$，错误率和精度为：

$$E(f;\mathcal{D}) = \int_{x\sim\mathcal{D}} \mathrm{II}\,(f(x) \neq y)p(x)dx$$

$$\mathrm{acc}(f;\mathcal{D}) = \int_{x\sim\mathcal{D}} \mathrm{II}\,(f(x) = y)p(x)dx$$

$$=1\text{-}E(f;\mathcal{D})$$

精度：预测正确的比例

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

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

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

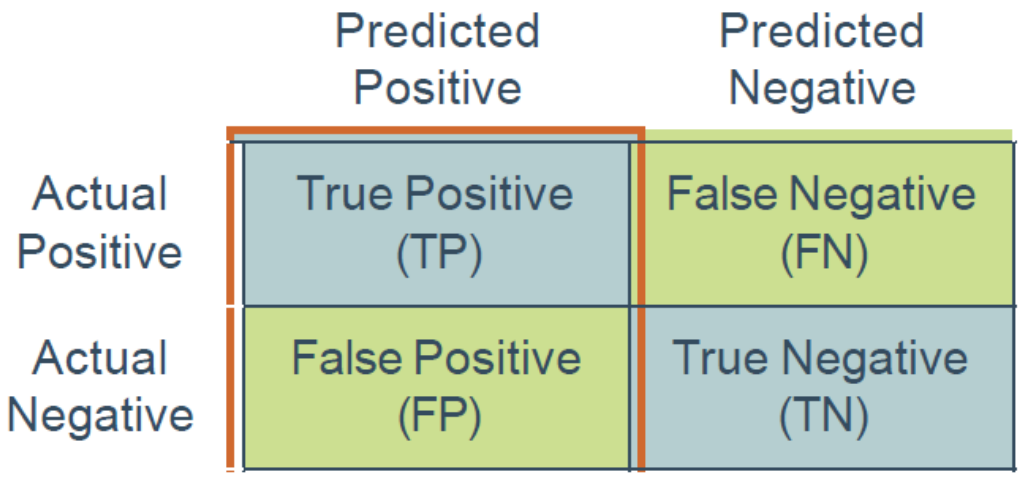
$$= 1 - \text{Accuracy}$$

- 查准率、查全率与$F1$分数

  - 查准率（precision，P）

    也叫准确率、精确率，表示预测为正的样例中有多少是真正的正样例，针对的是预测结果。

查准率：识别出的都是正例



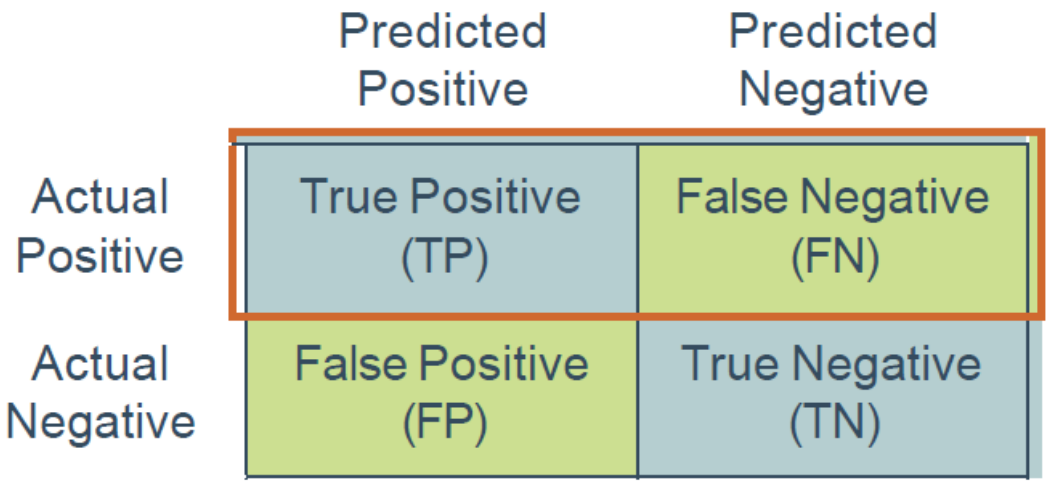
$$\text{Precision} = \frac{TP}{TP + FP}$$

- 查准率、查全率与$F1$分数

  - 查全率（recall，R）

    也叫召回率、敏感度、真正例，表示样例中的正例有多少被预测正确，针对的是原来的样本。

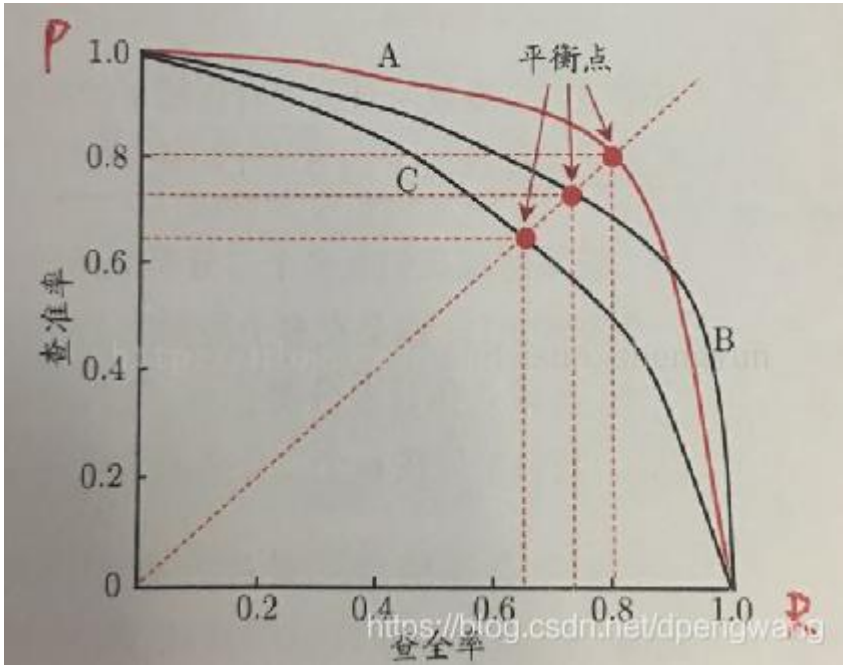
查全率或敏感度：识别出所有正例



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

- 查准率、查全率与$F1$分数

  - P-R曲线



平衡点（break-even point，BEP）：查准率=查全率

- 查准率、查全率与$F1$分数

  - $F1$分数

查准率与查全率的调和平均数，定义为：

$$\frac{1}{\mathrm{F_1}}=\frac{1}{2}\left(\frac{1}{P}+\frac{1}{R}\right)$$

$$\mathrm{F_1}=\frac{2\times \mathrm{P}\times \mathrm{R}}{\mathrm{P}+\mathrm{R}}=\frac{2\times \mathrm{TP}}{\text{样本总数}+\mathrm{TP}-\mathrm{TN}}$$

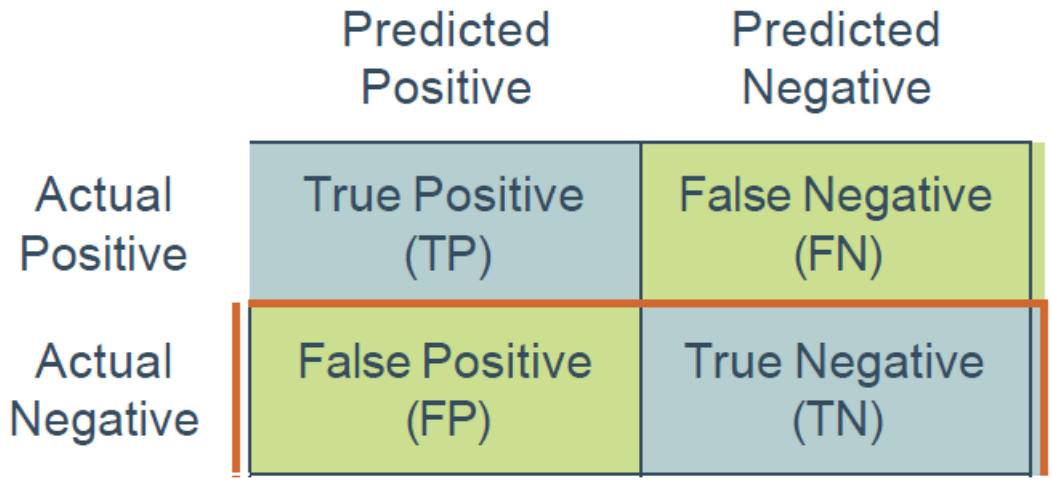
$F1$的一般形式：

$$\mathrm{F_\beta}=\frac{(1+\beta^2)\times \mathrm{P}\times \mathrm{R}}{\beta^2\times \mathrm{P}+\mathrm{R}}$$

$\beta>0$，度量了查全率对查准率的相对重要性：

$0<\beta<1$，查准率影响大
$B=1$，$F1$
$\beta>1$，查全率影响大

特异度：避免误报



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

错误评价指标

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

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

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

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

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

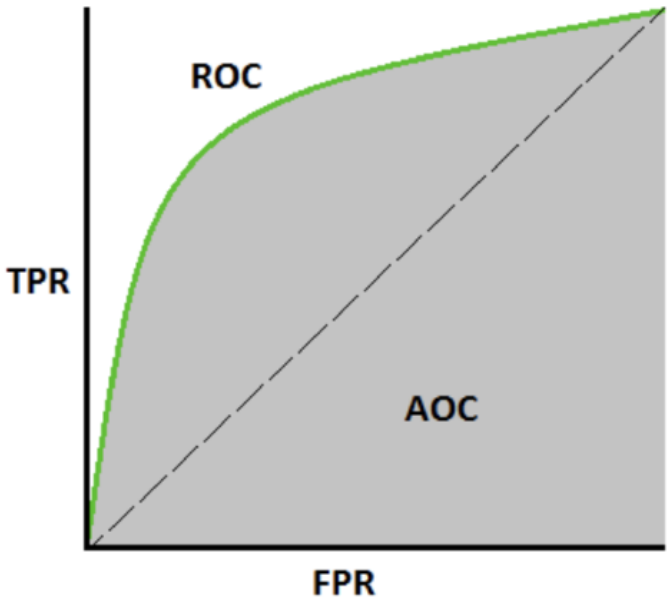
$$F1 = 2 \times \frac{P \times R}{P + R}$$

$$F_\beta = \frac{(1+\beta^2) \times P \times R}{\beta^2 \times P + R}$$

• ROC曲线与AUC

ROC曲线最早是运用在军事上的,后来逐渐运用到医学领域,并于20世纪80年代后期被引入机器学习领域。相传在第二次世界大战期间,雷达兵的任务之一就是死死地盯住雷达显示器,观察是否有敌机来袭。理论上讲,只要有敌机来袭,雷达屏幕上就会出现相应的信号。但是实际上,如果飞鸟出现在雷达扫描区域时,雷达屏幕上有时也会出现信号。这种情况令雷达兵烦恼不已,如果过于谨慎,凡是有信号就确定为敌机来袭,显然会增加误报风险;如果过于大胆,凡是信号都认为是飞鸟,又会增加漏报的风险。每个雷达兵都竭尽所能地研究飞鸟信号和飞机信号之间的区别,以便增加预报的准确性。但问题在于,每个雷达兵都有自己的判别标准,有的雷达兵比较谨慎,容易出现误报;有的雷达兵则比较胆大,容易出现漏报。

为了研究每个雷达兵预报的准确性,雷达兵的管理者汇总了所有雷达兵的预报特点,特别是他们漏报和误报的概率,并将这些概率画到一个二维坐标系里。这个二维坐标的纵坐标为敏感性(真阳性率),即在所有敌机来袭的事件中,每个雷达兵准确预报的概率。而横坐标则为1-特异性(假阳性率),表示在所有非敌机来袭信号中,雷达兵预报错误的概率。由于每个雷达兵的预报标准不同,且得到的敏感性和特异性的组合也不同。将这些雷达兵的预报性能进行汇总后,雷达兵管理员发现他们刚好在一条曲线上,这条曲线就是后来被广泛应用在医疗和机器学习领域的ROC曲线。



摘自《百面机器学习》

受试者工作特征（Receiver Operating Characteristic，ROC）曲线

真正例率（TPR）：在所有实际为阳性的样本中，被正确地判断为阳性之比率。
假正例率（FPR）：在所有实际为阴性的样本中，被错误地判断为阳性之比率。

取所有可能的阈值，计算（FPR, TPR)

- ROC曲线与AUC

➢ 士兵a-大胆型

	预测正	预测负
真实正	0	10
真实负	0	10

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

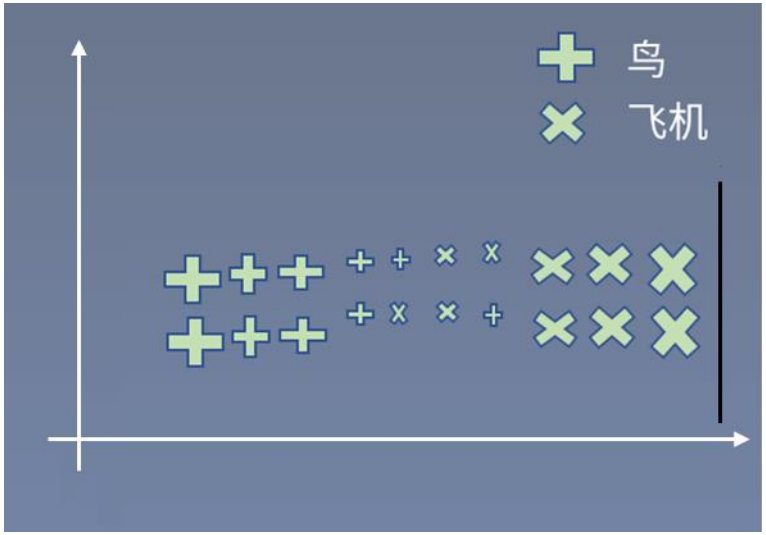
Y轴：TPR

X轴：FPR

真正率TPR：0

假正率FPR：0

画一个点：（0，0）



- ROC曲线与AUC

➢ 士兵b-一般型

	预测正	预测负
真实正	7	3
真实负	1	9

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

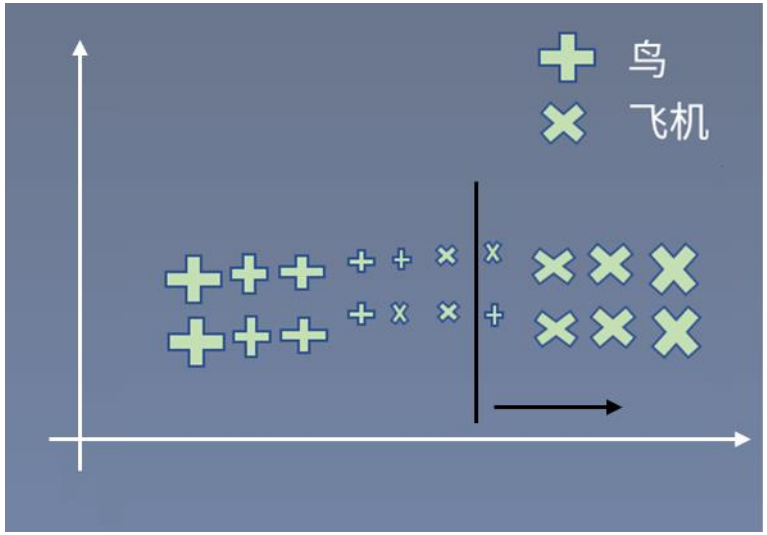
Y轴：TPR

X轴：FPR

真正率TPR：0.7

假正率FPR：0.1

画一个点：（0.1，0.7）



- ROC曲线与AUC

➢ 士兵c-谨慎型

	预测正	预测负
真实正	10	0
真实负	10	0

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

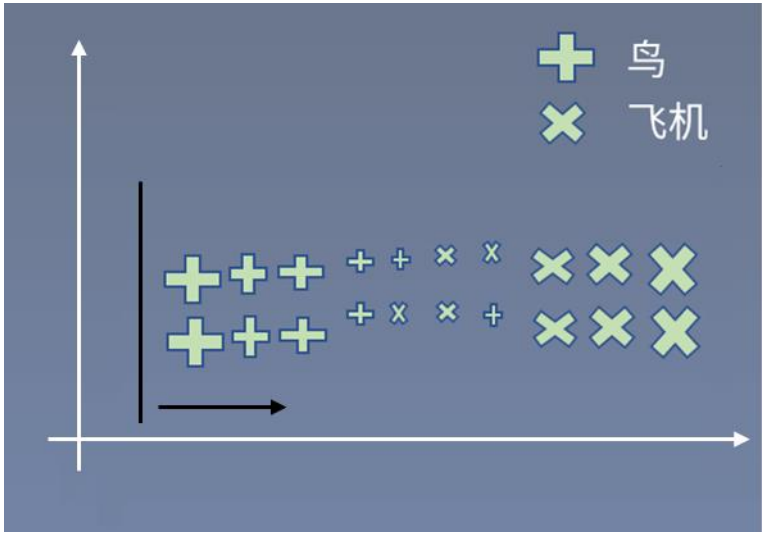
Y轴：TPR

X轴：FPR

真正率TPR：1

假正率FPR：1

画一个点：（1，1）



- ROC曲线与AUC

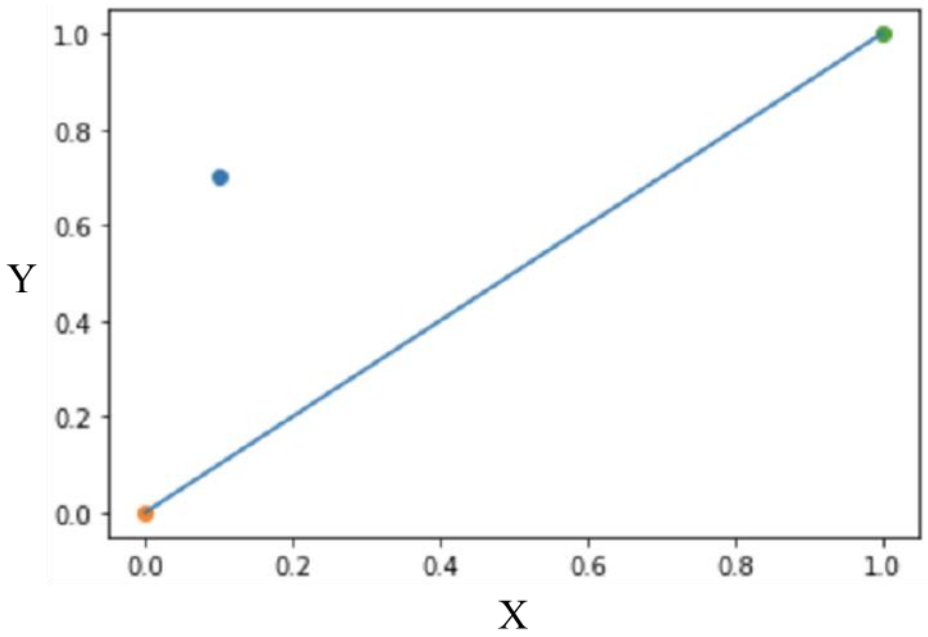
3个士兵3个点：

(0，0)

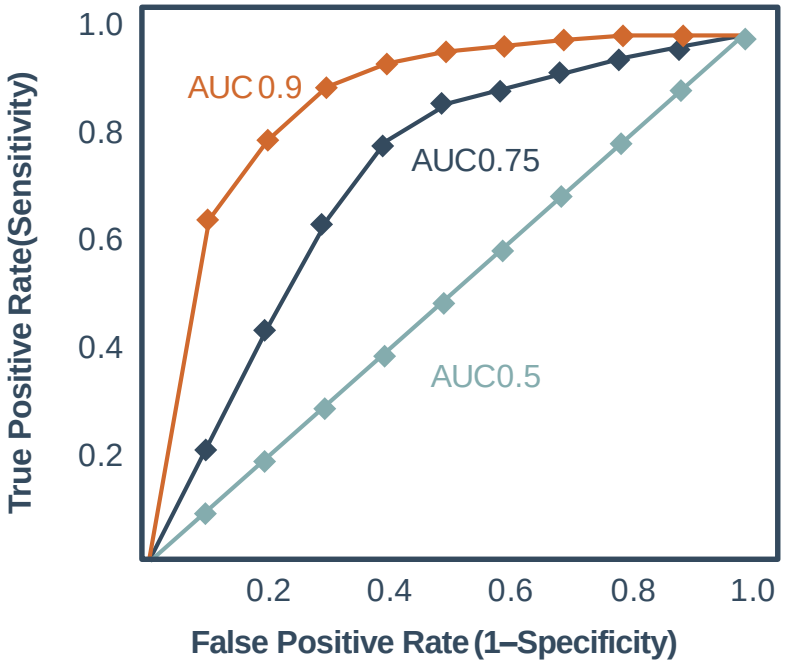
(0.1，0.7)

(1，1)

……



- ROC曲线与AUC



衡量ROC曲线下的面积（AUC）

- ROC曲线与AUC

  - ROC曲线的作用与优点

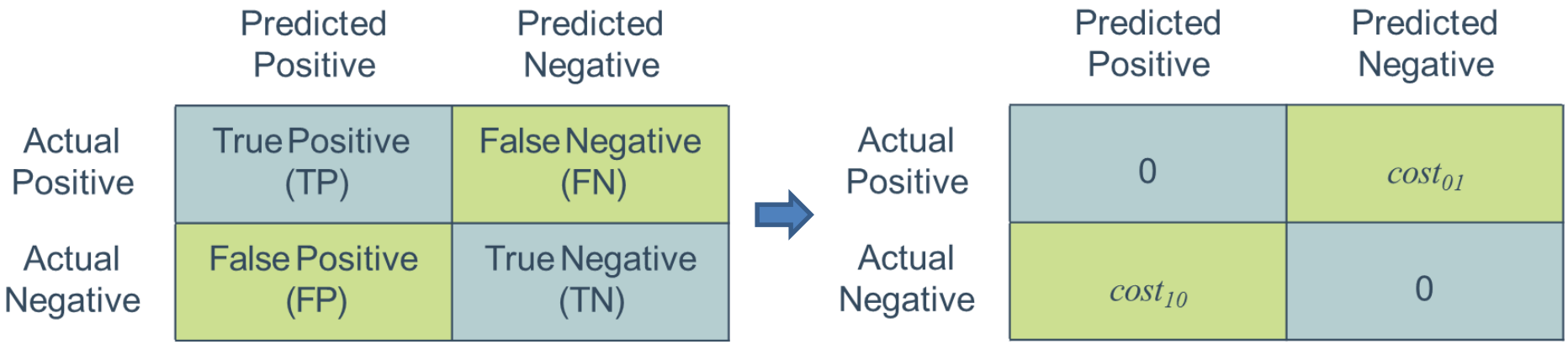
    - ✓ 能查出任意阈值对学习器泛化性能的影响
    - ✓ 有助于选择最佳的阈值
    - ✓ 可以比较不同学习器的性能

- 代价敏感错误率与代价曲线

  - 非均等代价（unequal cost）

    用来衡量不同类型错误所造成的不同损失。

代价矩阵



$cost_{ij}$表示把第i类样本预测为第j类样本的代价。

- 代价敏感错误率与代价曲线

➢ 代价敏感错误率（cost-sensitive error rate）

非均等代价下，不再是简单地最小化错误次数，而是希望最小化总体代价（total cost）。令D^+和D^-分别代表样例集D的正例子集和负例子集，则代价敏感错误率为：

$$E(f;D;\ \text{cost}) = \frac{1}{m}\left(\sum_{x_i \in D^+} \text{II}\left(f(x_i) \neq y_i\right) \times cost_{01} + \sum_{x_i \in D^-} \text{II}\left(f(x_i) \neq y_i\right) \times cost_{10}\right)$$

- 代价敏感错误率与代价曲线

  - 代价曲线（cost curve）

    在非均等代价下，ROC曲线不能直接反映出学习器的期望总体代价，而“代价曲线”则可达到该目的。

- 代价敏感错误率与代价曲线

  - 代价曲线（cost curve）

    横轴X：取值为[0, 1]的正例概率代价

$$P(+)\cos t = \frac{p * \cos t_{01}}{p * \cos t_{01} + (1-p)\cos t_{10}}$$

    P为样例为正例的概率。

- 代价敏感错误率与代价曲线

  - 代价曲线（cost curve）

    纵轴Y：取值为[0, 1]的归一化代价

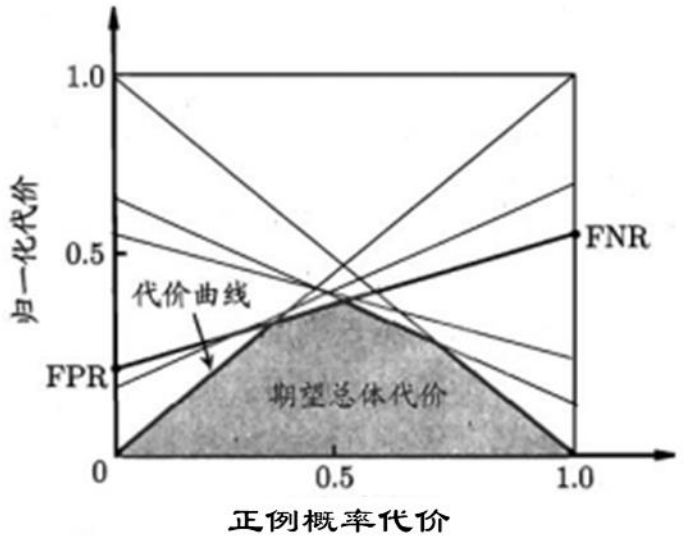
$$\text{cost}_{norm} = \frac{FNR * P * \cos t_{01} + FPR * (1 - P) * \cos t_{10}}{p * \cos t_{01} + (1 - p) * \cos t_{10}}$$

    P为样例为正例的概率。

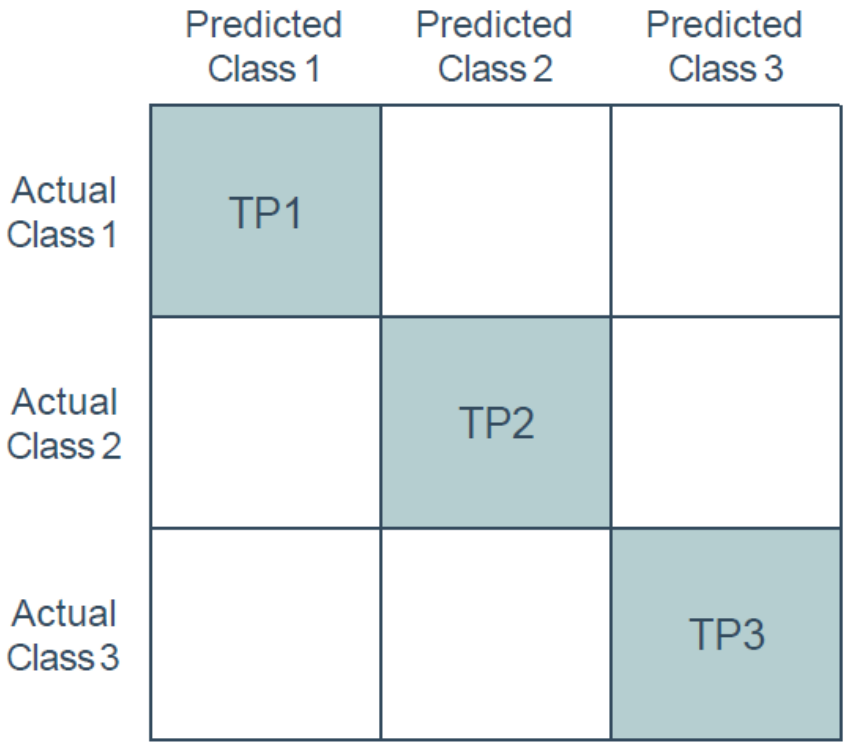
- 代价敏感错误率与代价曲线

  - 代价曲线（cost curve）

$$Y = (FNR-FPR)*X+FPR$$



多分类错误评价指标



$$\text{Accuracy} = \frac{TP1 + TP2 + TP3}{\text{Total}}$$

大部分多分类错误评价指标和二分类的类似—只是扩展为求和取平均。

宏平均和微平均

- 宏平均（**Macro-averaging**）

  先对每个类统计指标值，然后再对所有类求算术平均值。

$$\text{Macro\_P} = \frac{1}{n}\sum_{i=1}^{n} P_i$$

$$\text{Macro\_R} = \frac{1}{n}\sum_{i=1}^{n} R_i$$

$$\text{Macro\_F1} = \frac{1}{n}\sum_{i=1}^{n} F1_i$$

$$\text{Macro\_F1} = \frac{2\times Macro\_P\times Macro\_R}{Macro\_P+Macro\_R}$$

宏平均和微平均

- 微平均（**Micro-averaging**）

  对数据集中每个实例不分类别进行统计，建立全局混淆矩阵，然后再计算相应指标。

$$\text{Micro\_P} = \frac{\overline{\text{TP}}}{\overline{\text{TP}} + \overline{\text{FP}}} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FP_i}$$

$$\text{Micro\_R} = \frac{\overline{\text{TP}}}{\overline{\text{TP}} + \overline{\text{FN}}} = \frac{\sum_{i=1}^{n} TP_i}{\sum_{i=1}^{n} TP_i + \sum_{i=1}^{n} FN_i}$$

$$\text{Micro\_F1} = \frac{2 \times Micro\_P \times Micro\_R}{Micro\_P + Micro\_R}$$

练习

- 有10个样本，属于A、B、C三个类别。假设这10个样本的真实类别和预测的类别 分别是：
  - 真实：A A A C B C A B B C
  - 预测：A A C B A C A C B C

1. 求出每个类别的精度和错误率，P, R, 和F1
2. 求出宏平均P, R, 和F1
3. 求出微平均P, R, 和F1

练习

- 数据集包含1000个样本，其中500个正例，500个反例，将其划分为包含70%样本的训练集和30%个样本的测试集用于留出法评估，试估算共有多少种划分方法。