统计学习方法概论

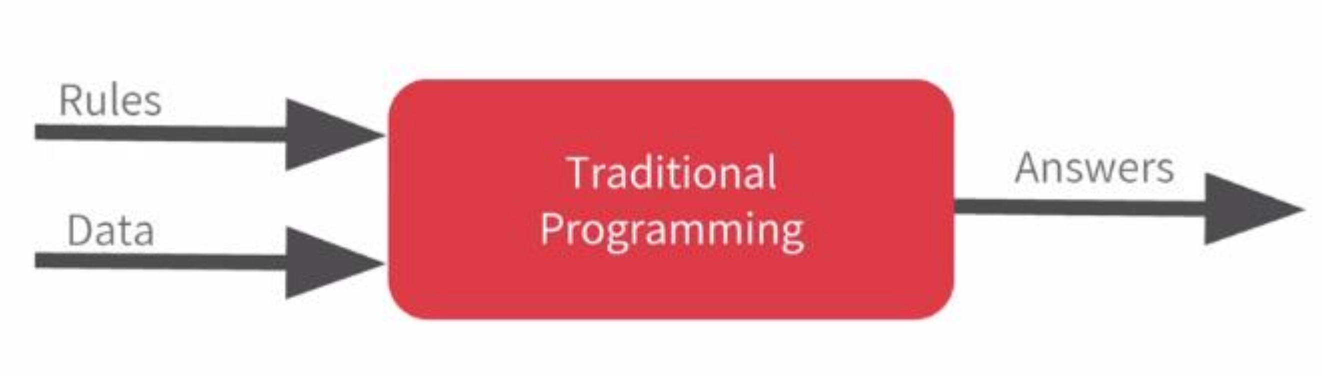
2019/10/10

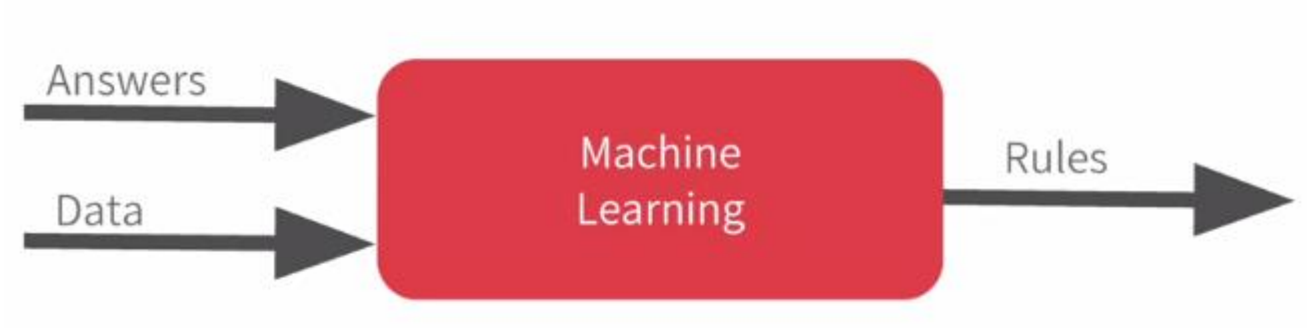
什么是机器学习?

A program can be said to learn from experience ***E*** *which respect to some class of tasks* ***T*** *and performance measure* ***P****, if its performance at tasks in* ***T****, as measured by* ***P****, improves which experience* ***E****.*

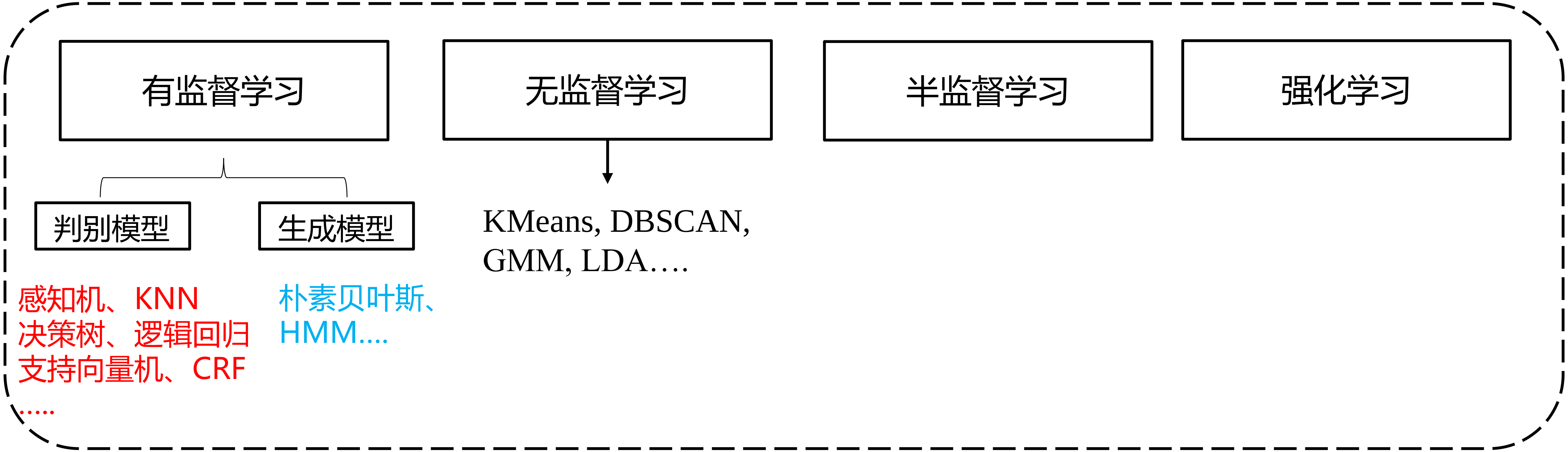
—— *Tom Mitchell*

机器学习 vs 传统程序


Rules
Data
Traditional Programming
Answers


Answers
Data
Machine Learning
Rules

机器学习算法分类



有监督学习

判别模型

感知机、KNN
决策树、逻辑回归
支持向量机、CRF
…..

生成模型

朴素贝叶斯、
HMM….

无监督学习

KMeans, DBSCAN,
GMM, LDA….

半监督学习

强化学习

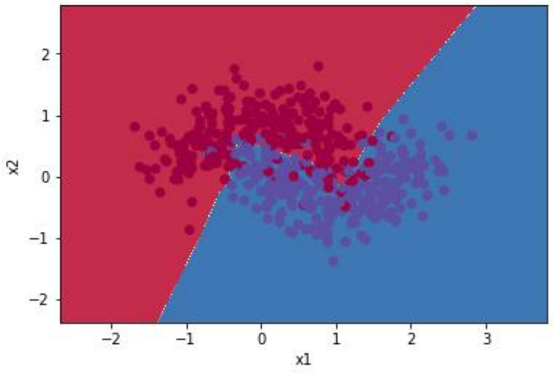
统计学习 三要素

模型 + 策略 + 算法

模型

决策函数

$$y = ax + b$$



策略

损失函数

- 0-1损失函数
- 均方损失函数
- 绝对损失函数
- 对数损失函数
- softmax损失

损失函数值越小，说明模型对数据集拟合的越好

风险函数

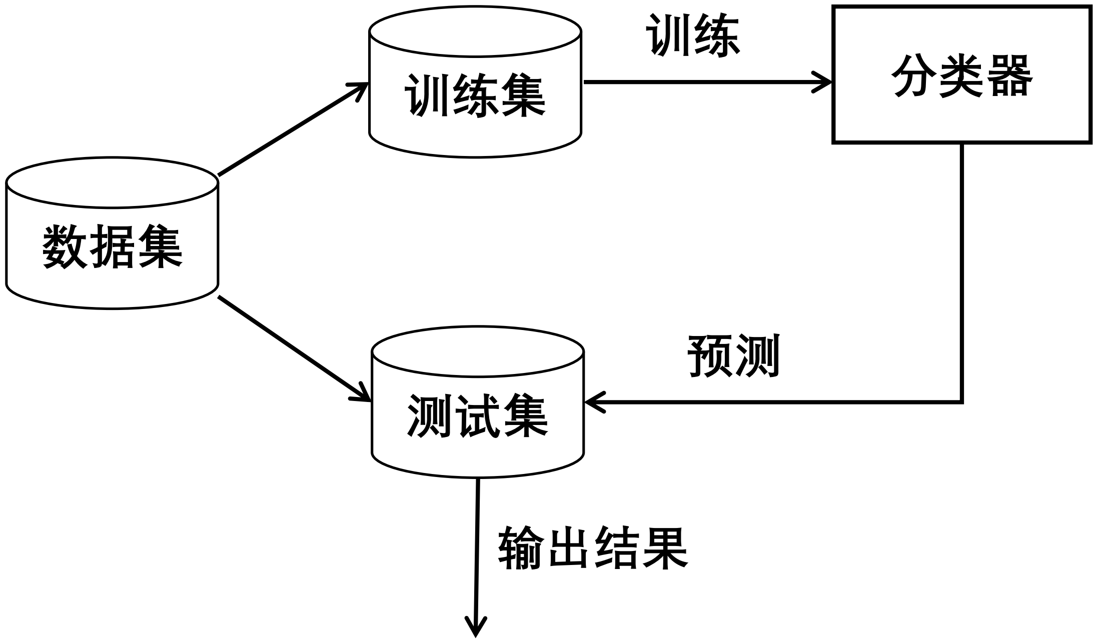
- 经验风险最小化 → 经验风险最小的模型就是最优的模型
  缺点：样本量小时，过拟合
- 结构风险最小化 → 在经验风险的基础上引入正则项：

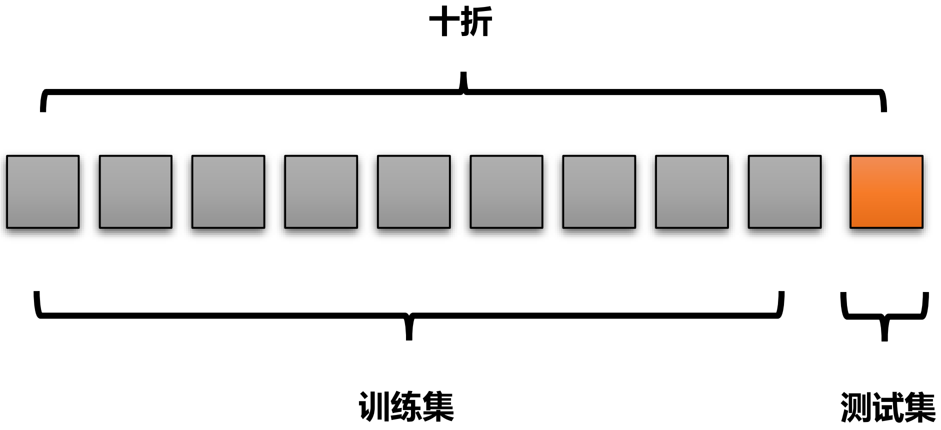
$$R(f) = \frac{1}{N}\sum_{i=1}^{N} L\big(y_i, f(x_1)\big) + \boxed{\lambda J(f)}$$

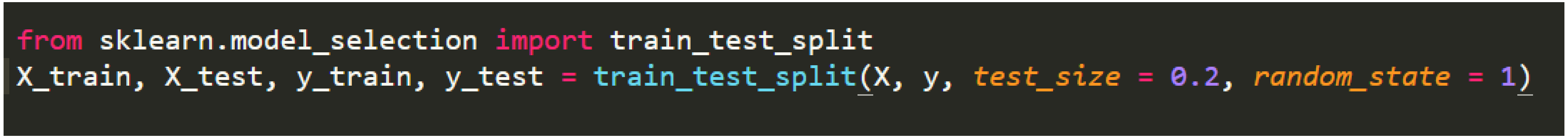
算法

最优化方法

理想状态下模型训练的一般流程







```python
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = 0.2, random_state = 1)
```

模型评估

- 训练误差 vs 测试误差

- 过拟合 vs 欠拟合

偏差(bias） vs 方差(variance)

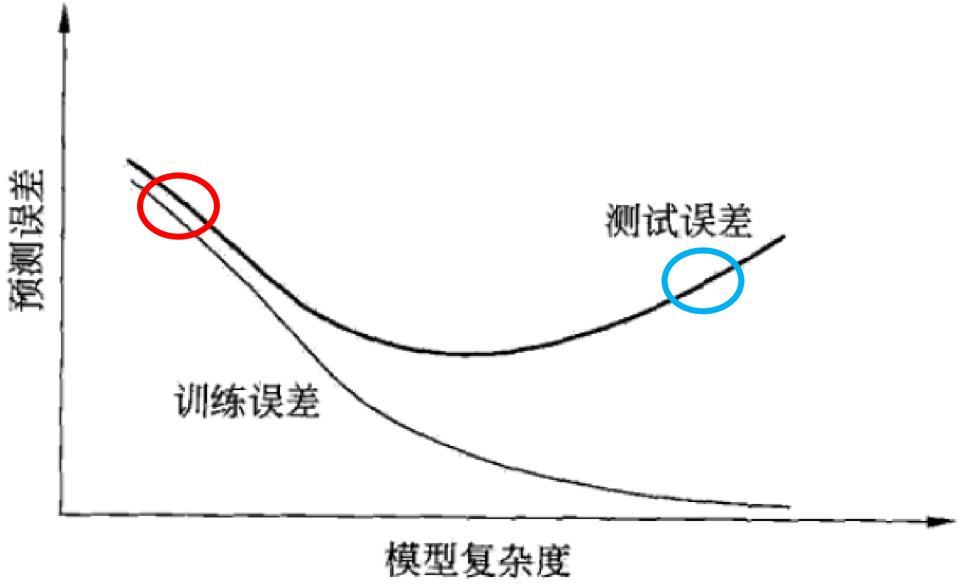


图 1.3 训练误差和测试误差与模型复杂度的关系

偏差：度量学习算法在训练集上的预测值与真实值的偏离程度，反映了模型的**拟合能力**。

方差：度量了同样大小的数据集的变动所导致的学习性能的变化。即反应了模型的**泛化能力**。

过拟合 和 欠拟合 的解决办法

过拟合

- 增加正则项
- 增大正则项系数
- 使用更多的训练样本
- 使用更少的特征（特征选择）
- 增加噪声
- 剪枝（决策树）
- Early stopping
- Dropout
- BN
- ……

欠拟合

- 尝试更多的特征（特征交叉、多项式特征）
- 减小正则项系数
- 使用更复杂的模型
- …….

机器学习中常用评估指标

常用的准确率（accuracy）有何缺点?

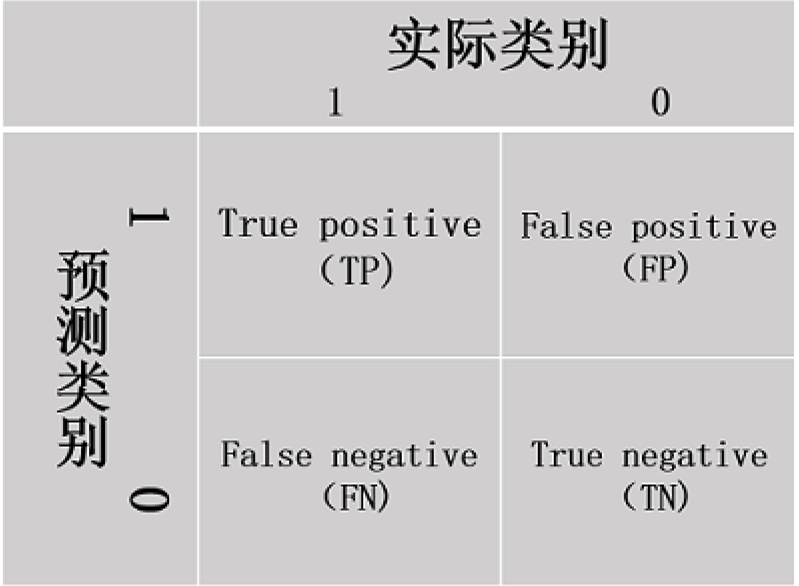
无法适用于**数据不平衡**现象。

Example：对于二分类而言，有99个正例（label 为 1），
1个负例（label 为0）。模型为：y=1，
accuracy=99/100=99%

机器学习中常用评估指标

precision、recall、F-Measure

混淆矩阵



	实际类别	
	1	0
预测类别 1	True positive （TP）	False positive （FP）
0	False negative （FN）	True negative （TN）

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

$$Recall = \frac{|TP|}{|TP| + |FN|}$$

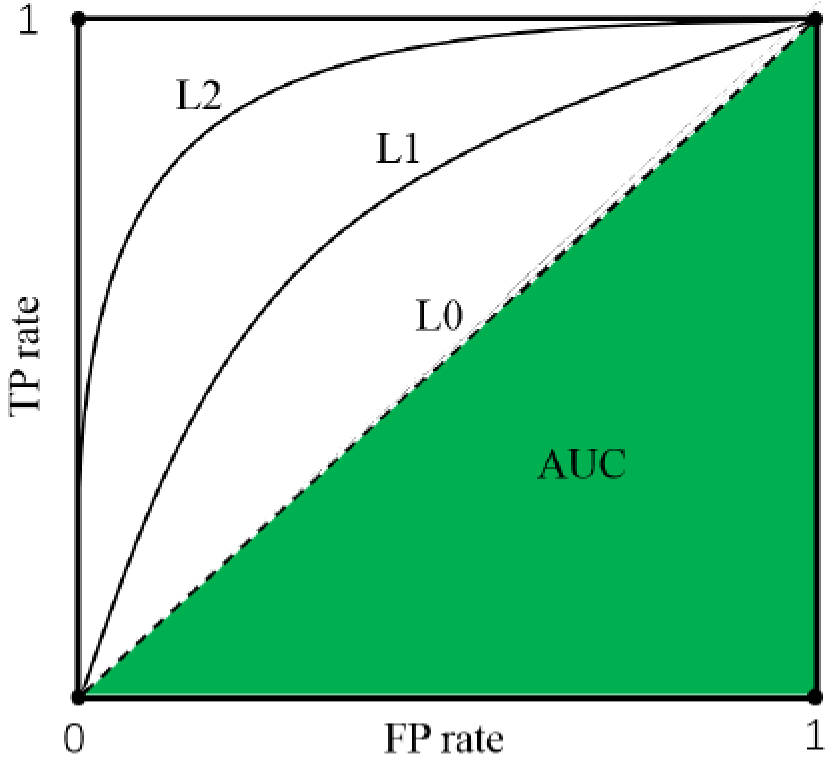
$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

β =1

$$F_{measure} = \frac{(\beta^2 + 1) recall \cdot precision}{recall + \beta^2 precision}$$

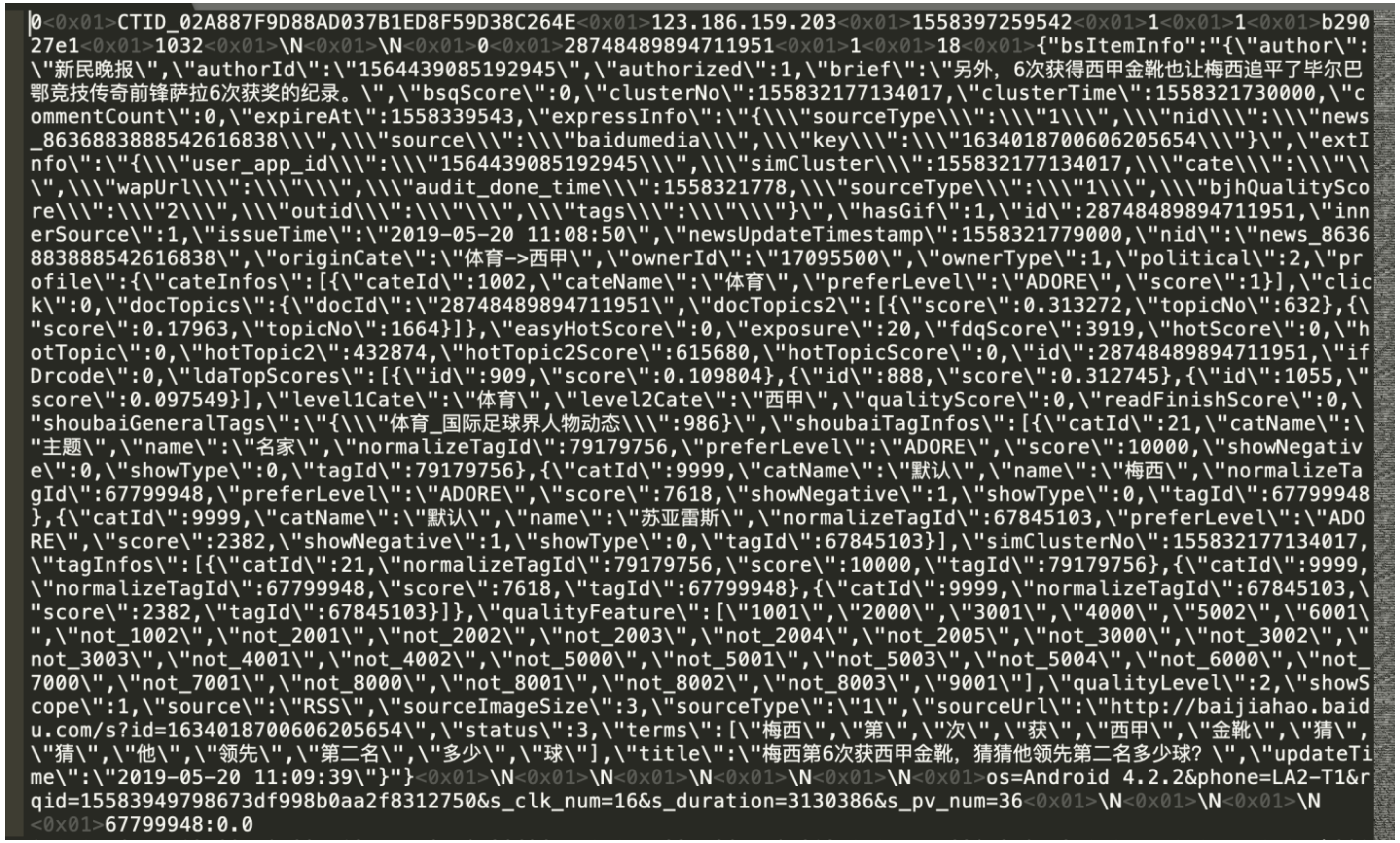
机器学习中常用评估指标

ROC(receiver operating characteristic)，受试者工作特征



非理想状态下模型训练的一般流程

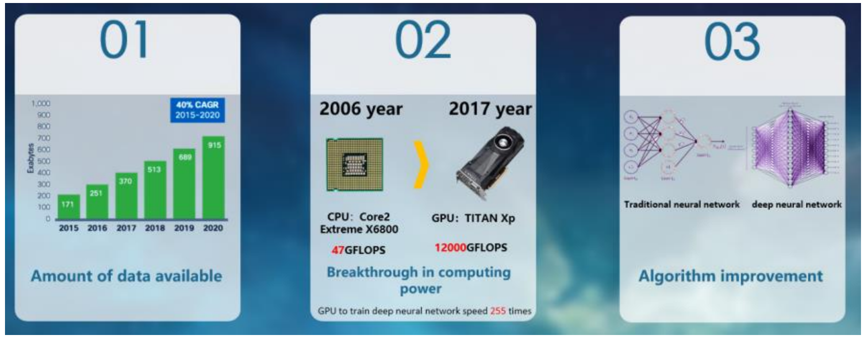
当我们只有用户的原始日志记录，那么该如何训练模型？



```
0<0x01>CTID_02A887F9D88AD037B1ED8F59D38C264E<0x01>123.186.159.203<0x01>1558397259542<0x01>1<0x01>1<0x01>b290
27e1<0x01>1032<0x01>\N<0x01>\N<0x01>0<0x01>28748489894711951<0x01>1<0x01>18<0x01>{"bsItemInfo":"{\"author\":
\"新民晚报\",\"authorId\":\"1564439085192945\",\"authorized\":1,\"brief\":\"另外，6次获得西甲金靴也让梅西追平了毕尔巴
鄂竞技传奇前锋萨拉6次获奖的纪录。\",\"bsqScore\":0,\"clusterNo\":15583217713401 7,\"clusterTime\":1558321730000,\"c
ommentCount\":0,\"expireAt\":1558339543,\"expressInfo\":\"{\\\"sourceType\\\":\\\"1\\\",\\\"nid\\\":\\\"news
_8636883888542616838\\\",\\\"source\\\":\\\"baidumedia\\\",\\\"key\\\":\\\"1634018700606205654\\\"}\",\"extI
nfo\":\"{\\\"user_app_id\\\":\\\"1564439085192945\\\",\\\"simCluster\\\":15583217713401 7,\\\"cate\\\":\\\"\\
\",\\\"wapUrl\\\":\\\"\\\",\\\"audit_done_time\\\":1558321778,\\\"sourceType\\\":\\\"1\\\",\\\"bjhQualitySco
re\\\":\\\"2\\\",\\\"outid\\\":\\\"\\\",\\\"tags\\\":\\\"\\\"}\",\"hasGif\":1,\"id\":28748489894711951,\"inn
erSource\":1,\"issueTime\":\"2019-05-20 11:08:50\",\"newsUpdateTimestamp\":1558321779000,\"nid\":\"news_8636
883888542616838\",\"originCate\":\"体育->西甲\",\"ownerId\":\"17095500\",\"ownerType\":1,\"political\":2,\"pr
ofile\":{\"cateInfos\":[{\"cateId\":1002,\"cateName\":\"体育\",\"preferLevel\":\"ADORE\",\"score\":1}],\"clic
k\":0,\"docTopics\":{\"docId\":\"28748489894711951\",\"docTopics2\":[{\"score\":0.313272,\"topicNo\":632},{\
"score\":0.17963,\"topicNo\":1664}]},\"easyHotScore\":0,\"exposure\":20,\"fdqScore\":3919,\"hotScore\":0,\"h
otTopic\":0,\"hotTopic2\":432874,\"hotTopic2Score\":615680,\"hotTopicScore\":0,\"id\":28748489894711951,\"if
Drcode\":0,\"ldaTopScores\":[{\"id\":909,\"score\":0.109804},{\"id\":888,\"score\":0.312745},{\"id\":1055,\"
score\":0.097549}],\"level1Cate\":\"体育\",\"level2Cate\":\"西甲\",\"qualityScore\":0,\"readFinishScore\":0,\
"shoubaiGeneralTags\":\"{\\\"体育_国际足球界人物动态\\\":986}\",\"shoubaiTagInfos\":[{\"catId\":21,\"catName\":\
"主题\",\"name\":\"名家\",\"normalizeTagId\":79179756,\"preferLevel\":\"ADORE\",\"score\":10000,\"showNegativ
e\":0,\"showType\":0,\"tagId\":79179756},{\"catId\":9999,\"catName\":\"默认\",\"name\":\"梅西\",\"normalizeTa
gId\":67799948,\"preferLevel\":\"ADORE\",\"score\":7618,\"showNegative\":1,\"showType\":0,\"tagId\":67799948
},{\"catId\":9999,\"catName\":\"默认\",\"name\":\"苏亚雷斯\",\"normalizeTagId\":67845103,\"preferLevel\":\"ADO
RE\",\"score\":2382,\"showNegative\":1,\"showType\":0,\"tagId\":67845103}],\"simClusterNo\":15583217713401 7,
\"tagInfos\":[{\"catId\":21,\"normalizeTagId\":79179756,\"score\":10000,\"tagId\":79179756},{\"catId\":9999,
\"normalizeTagId\":67799948,\"score\":7618,\"tagId\":67799948},{\"catId\":9999,\"normalizeTagId\":67845103,\
"score\":2382,\"tagId\":67845103}]},\"qualityFeature\":[\"1001\",\"2000\",\"3001\",\"4000\",\"5002\",\"6001\
",\"not_1002\",\"not_2001\",\"not_2002\",\"not_2003\",\"not_2004\",\"not_2005\",\"not_3000\",\"not_3002\",\"
not_3003\",\"not_4001\",\"not_4002\",\"not_5000\",\"not_5001\",\"not_5003\",\"not_5004\",\"not_6000\",\"not_
7000\",\"not_7001\",\"not_8000\",\"not_8001\",\"not_8002\",\"not_8003\",\"9001\"],\"qualityLevel\":2,\"showS
cope\":1,\"source\":\"RSS\",\"sourceImageSize\":3,\"sourceType\":\"1\",\"sourceUrl\":\"http://baijiahao.baid
u.com/s?id=1634018700606205654\",\"status\":3,\"terms\":[\"梅西\",\"第\",\"次\",\"获\",\"西甲\",\"金靴\",\"猜\",
\"猜\",\"他\",\"领先\",\"第二名\",\"多少\",\"球\"],\"title\":\"梅西第6次获西甲金靴，猜猜他领先第二名多少球？\",\"updateTi
me\":\"2019-05-20 11:09:39\"}"}<0x01>\N<0x01>\N<0x01>\N<0x01>\N<0x01>\N<0x01>os=Android 4.2.2&phone=LA2-T1&r
qid=15583949798673df998b0aa2f8312750&s_clk_num=16&s_duration=3130386&s_pv_num=36<0x01>\N<0x01>\N<0x01>\N
<0x01>67799948:0.0
```

我们需要手工设计特征，这也是传统机器学习必须的一步。

深度学习为什么能成功？



谢谢