



# AI机器学习模型预测实战

日期：2021.01

---

## 目录 / CONTENTS

01 Scikit-learn

02 股票预测

03 糖尿病预测

04 手写数字识别





PART 01

# Scikit-learn

---

Scikit-learn (以前称为scikits.learn, 也称为sklearn) 是针对Python 编程语言的免费软件机器学习库。它具有各种分类, 回归和聚类算法, 包括支持向量机, 随机森林, 梯度下降, Kmeans。

网址:

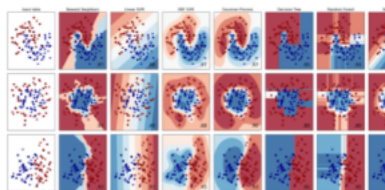
<https://scikit-learn.org/stable/index.html>

## 分类

标识对象所属的类别。

应用范围: 垃圾邮件检测, 图像识别。

算法: SVM, 最近的邻居, 随机森林, 和更多...



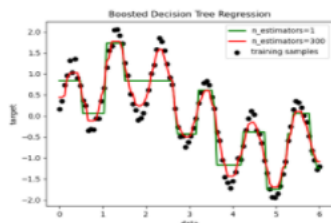
例子

## 回归

预测与对象关联的连续值属性。

应用范围: 药物反应, 股票价格。

算法: SVR, 最近的邻居, 随机森林, 和更多...



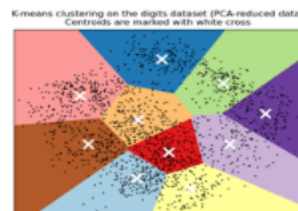
例子

## 聚类

自动将相似对象归为一组。

应用: 客户细分, 分组实验成果

算法: k-均值, 谱聚类, 平均移动, 和更多...



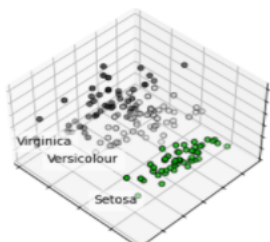
例子

## 降维

减少要考虑的随机变量的数量。

应用: 可视化, 提高效率

的算法: K-手段, 特征选择, 非负矩阵分解, 以及更多...

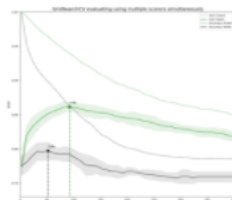


## 选型

比较, 验证和选择参数和模型。

应用: 通过参数调整改进精度

算法: 网格搜索, 交叉验证, 指标, 和更多...

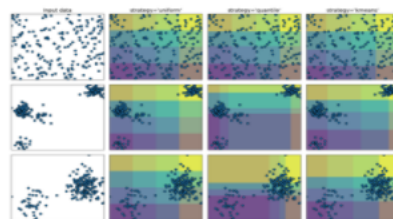


## 预处理

特征提取和归一化。

应用程序: 转换输入数据, 例如文本, 以供机器学习算法使用。

算法: 预处理, 特征提取, 以及更多...





# 导入数据集, sklearn自带数据集

- ✓ **from sklearn.datasets import load\_boston** # 波士顿房价数据集
- ✓ **from sklearn.datasets import load\_breast\_cancer** # 乳腺癌数据集
- ✓ **from sklearn.datasets import load\_iris** # 鸢尾花数据集

# 导入数据集切分工具

- ✓ **from sklearn.model\_selection import train\_test\_split** # 数据切分

# 导入模型

- ✓ **from sklearn.linear\_model import LinearRegression** # 线性回归模型
- ✓ **from sklearn.linear\_model import LogisticRegression** # 逻辑回归模型

# 导入数据预处理数据

- ✓ **from sklearn.preprocessing import PolynomialFeatures** # 多项式特征
- ✓ **from sklearn.preprocessing import StandardScaler** # 标准化
- ✓ **from sklearn.pipeline import Pipeline** # 管道





PART 02

## 股票预测

# 股票数据集含义

特征	解释
date	日期
open	开盘价
high	最高价
close	收盘价
low	最低价
volume	成交量
price_change	价格变动
p_change	涨跌幅
ma5	5日均价
ma10	10日均价
ma20	20日均价
v_ma5	5日均量
v_ma10	10日均量
v_ma20	20日均量
turnover	换手率



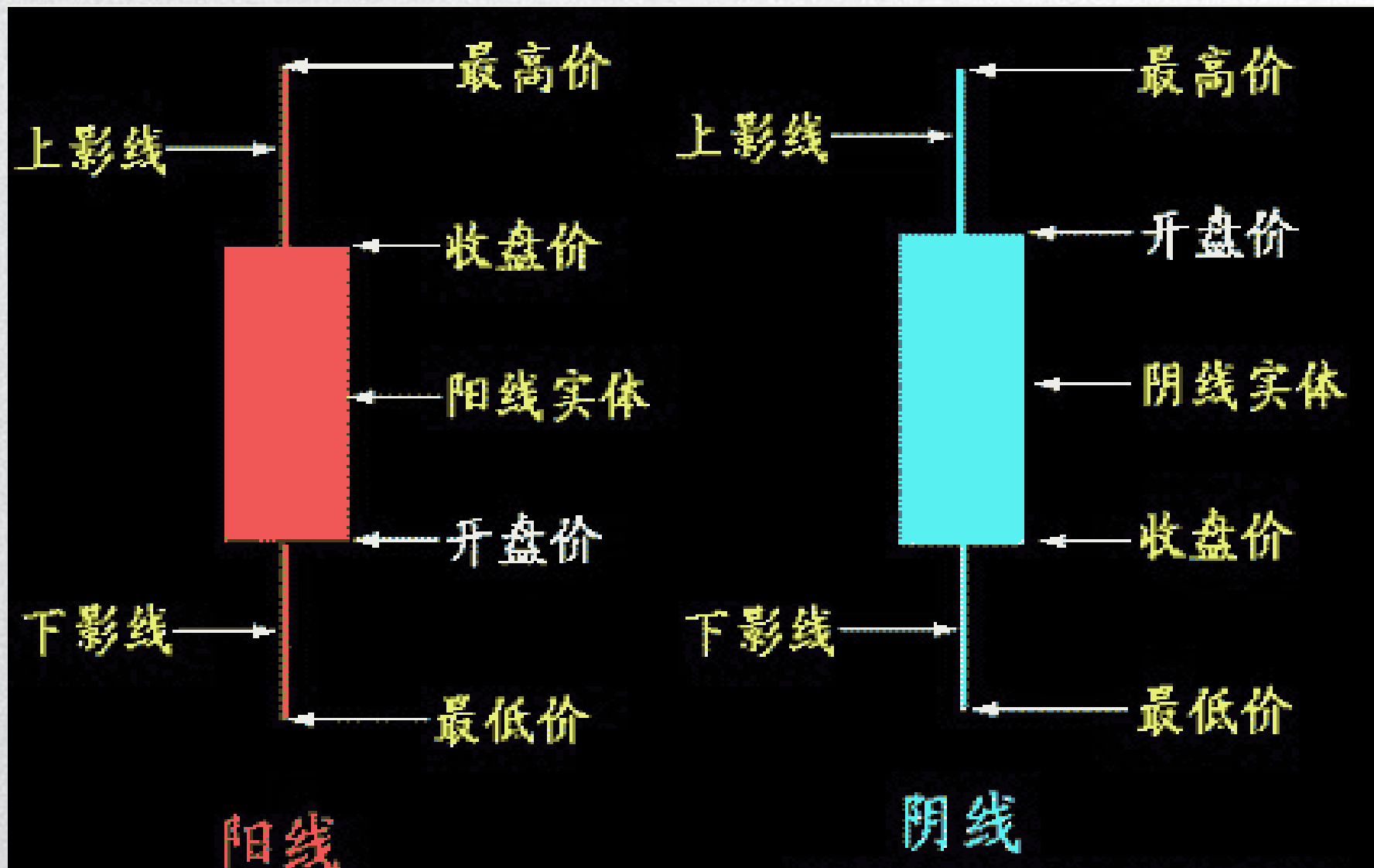
安装库，通过tushare包可以获取到股票数据

✓ pip install tushare

✓ win+R 输入CMD

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
date	open	high	close	low	volume	price_change	p_change	ma5	ma10	ma20	v_ma5	v_ma10	v_ma20	turnover
2020/9/23	1635	1666.01	1649.98	1621.02	35725.52	-15.71	-0.94	1669.598	1703.789	1729.764	38463.26	32257.55	32590.67	0.2
2020/9/22	1650	1686.98	1665.69	1650	23265.47	-1.11	-0.07	1684.622	1707.591	1733.615	35821.35	31281.35	32115.18	0.1
2020/9/21	1692.76	1692.76	1666.8	1666.66	22444.07	-28.2	-1.66	1703.484	1712.162	1736.68	35332.43	32402.04	33045.54	0.1
2020/9/18	1665	1695	1695	1635.01	47201.4	24.48	1.47	1723.324	1717.832	1738.158	37054.82	33268.84	33470	0.3
2020/9/17	1700	1700	1670.52	1658	63679.85	-54.58	-3.16	1730.924	1725.332	1737.208	33832.14	31634.5	32364.05	0.5
2020/9/16	1760	1764.81	1725.1	1719	22515.98	-34.9	-1.98	1737.98	1737.58	1736.882	26051.84	28310.25	30299.18	0.1
2020/9/15	1769.99	1769.99	1760	1747.75	20820.87	-6	-0.34	1730.56	1744.57	1734.977	26741.35	29259.58	30519.13	0.1
2020/9/14	1745	1769	1766	1730.58	31056.01	33	1.9	1720.84	1748.768	1732.227	29471.65	30341.48	30999.23	0.2
2020/9/11	1688	1736	1733	1688	31088	27.2	1.59	1712.34	1750.818	1728.427	29482.86	32016.92	31625.03	0.2
2020/9/10	1703.74	1720	1705.8	1700	24778.33	17.8	1.05	1719.74	1753.218	1724.827	29436.85	33138.08	31953.28	0.1
2020/9/9	1699.67	1711	1688	1680.04	25963.55	-23.4	-1.37	1737.18	1755.738	1721.287	30568.65	32923.8	31741.44	0.2
2020/9/8	1732	1737.8	1711.4	1677.07	34472.38	-12.1	-0.7	1758.58	1759.638	1718.234	31777.8	32949.02	31839.95	0.2
2020/9/7	1760	1777.99	1723.5	1703.97	31112.02	-46.5	-2.63	1776.696	1761.198	1714.79	31211.31	33689.03	31907.24	0.2
2020/9/4	1766	1776.99	1770	1746	30857.96	-23	-1.28	1789.296	1758.483	1710.314	34550.98	33671.17	31532.82	0.2
2020/9/3	1795	1812	1793	1779.7	30437.35	-2	-0.11	1786.696	1749.083	1703.364	36839.31	33093.61	31702.02	0.2
2020/9/2	1825	1828	1795	1770	32009.29	-6.98	-0.39	1774.296	1736.183	1695.612	35278.94	32288.11	31812.6	0.2





## 多元1次方多项式模型表达式

$$y = a_1 \cdot x_1 + a_2 \cdot x_2 + a_3 \cdot x_3 + a_4 \cdot x_4$$

.....

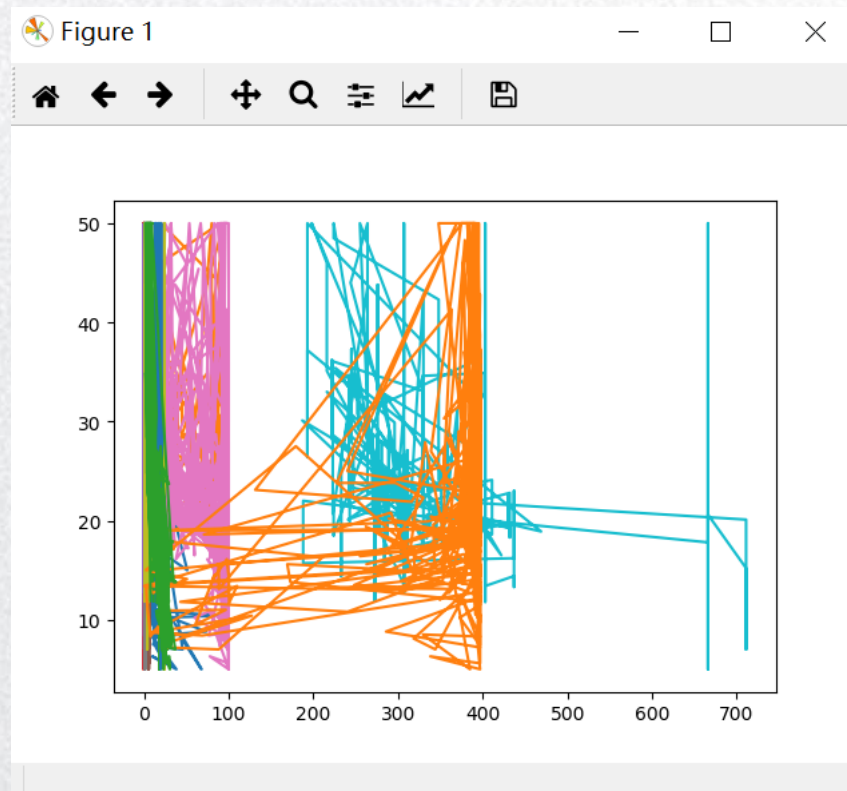
$$+ a_n \cdot x_n + b$$

## 2元1次方多项式模型表达式

$$y = a_1 \cdot x_1 + a_2 \cdot x_2 + b$$

## 3元1次方多项式模型表达式

$$y = a_1 \cdot x_1 + a_2 \cdot x_2 + a_3 \cdot x_3 + b$$





## 数据构成以及数据集切分

# 根据训练数据的已知X和Y，找到模型参数

✓ **训练集 Training Set Data**

# 根据训练好的模型，在测试集上做推断，验证

✓ **测试集 Testing Set Data**

# 加载数据集

✓ **`stock = ts.get_hist_data('600519')`**

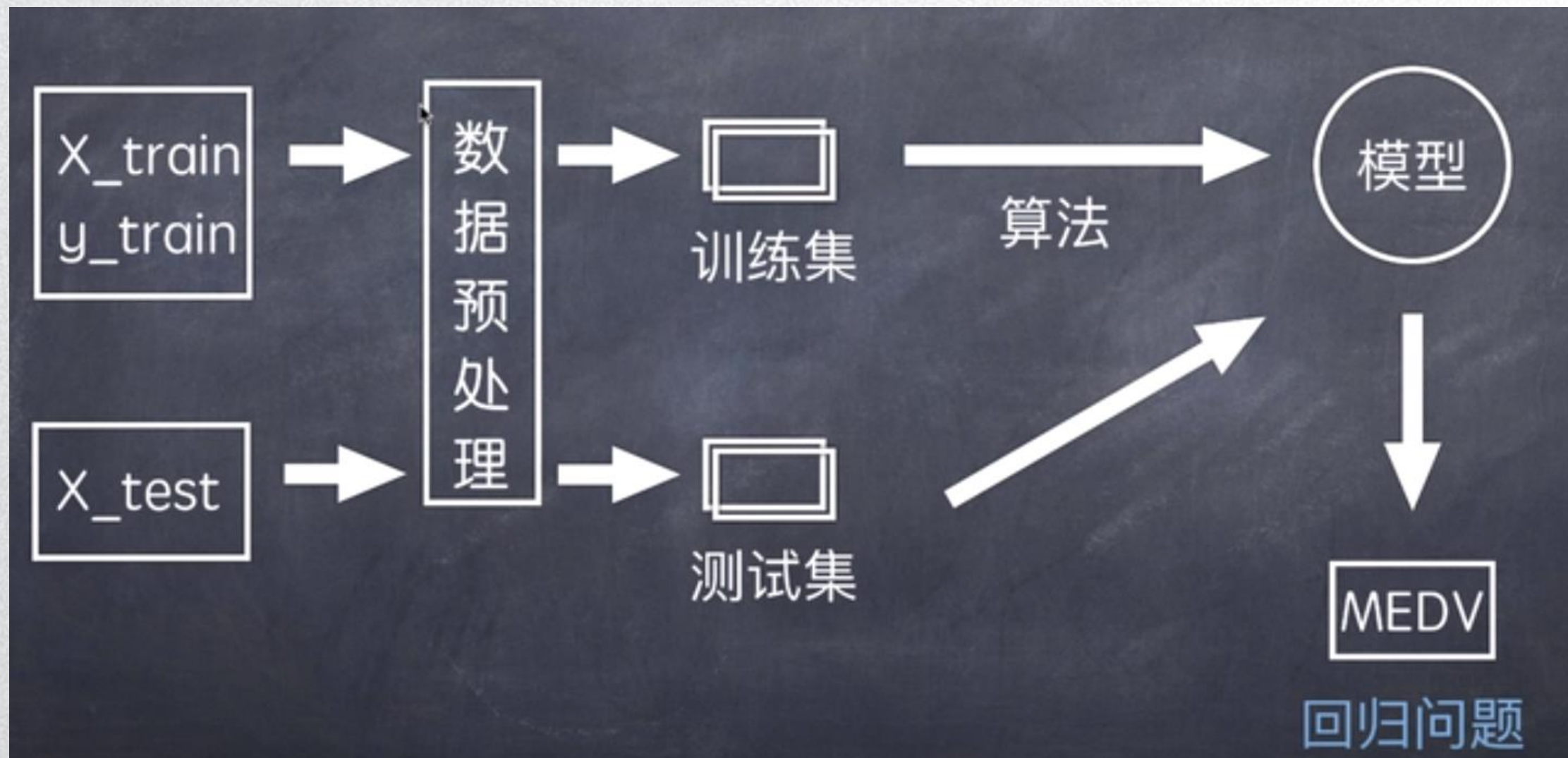
# 时序数据切分

✓ **`X_train = X[:-120, :]`**

✓ **`y_train = y[:-120, :]`**

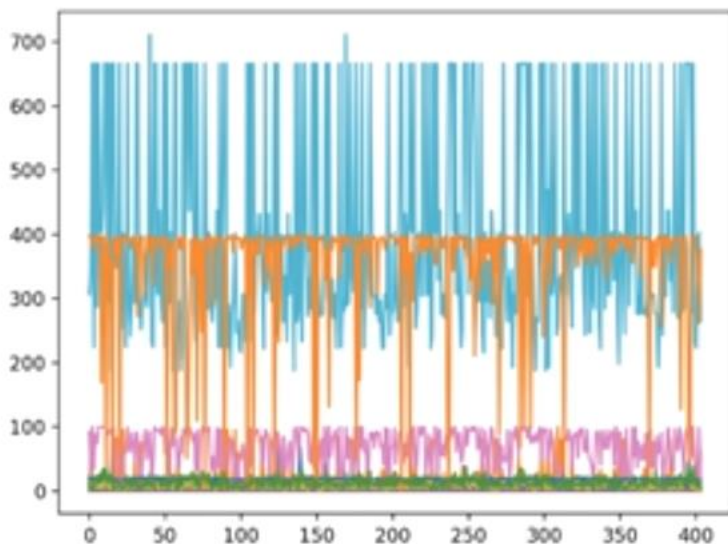
✓ **`X_test = X[-120:, :]`**

✓ **`y_test = y[-120:, :]`**

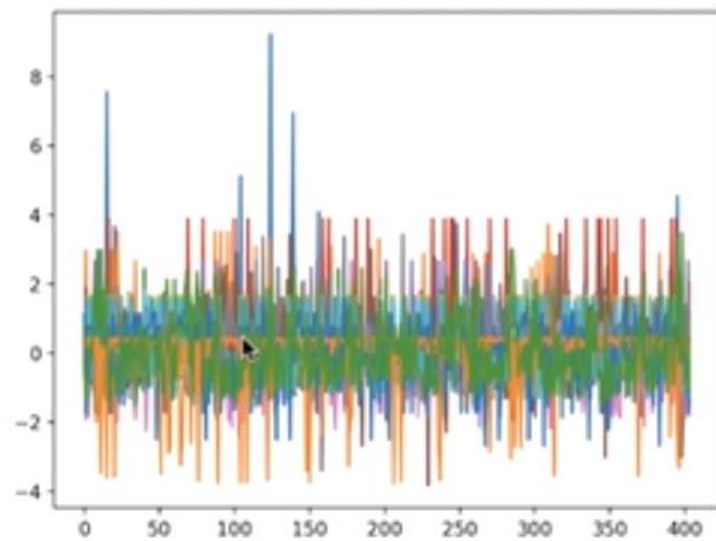




## 数据预处理

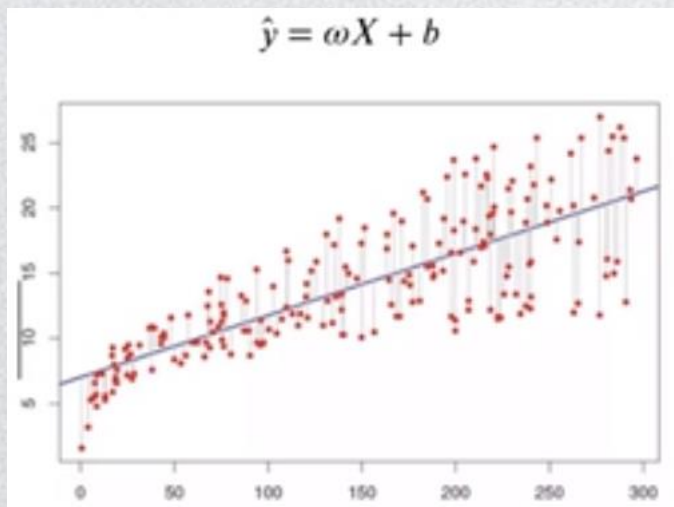


标准化处理

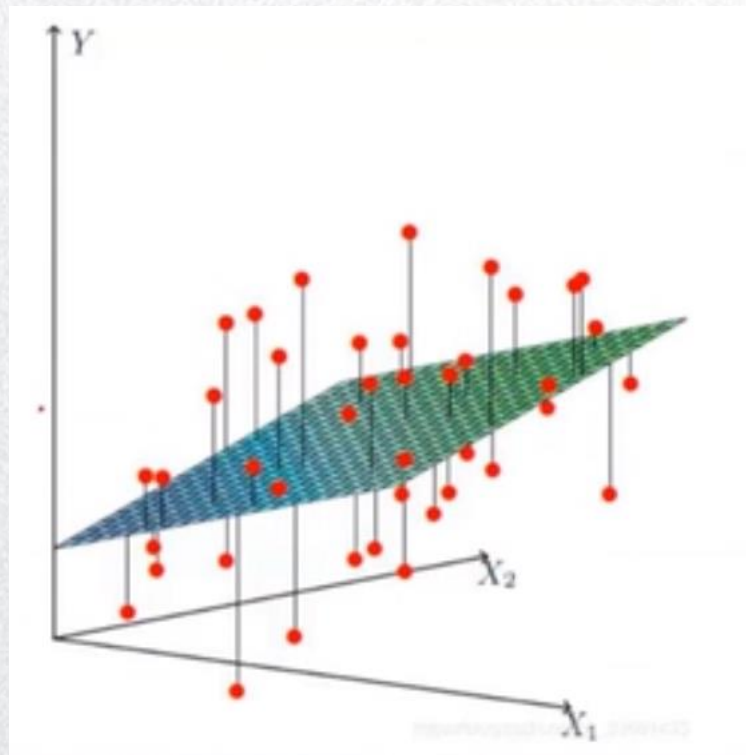


# 一元、多元线性回归模型LOSS函数对比

## 一元线性回归LOSS函数图解 (点到线的距离)



## 多元线性回归LOSS函数图解 (点到超平面的距离)







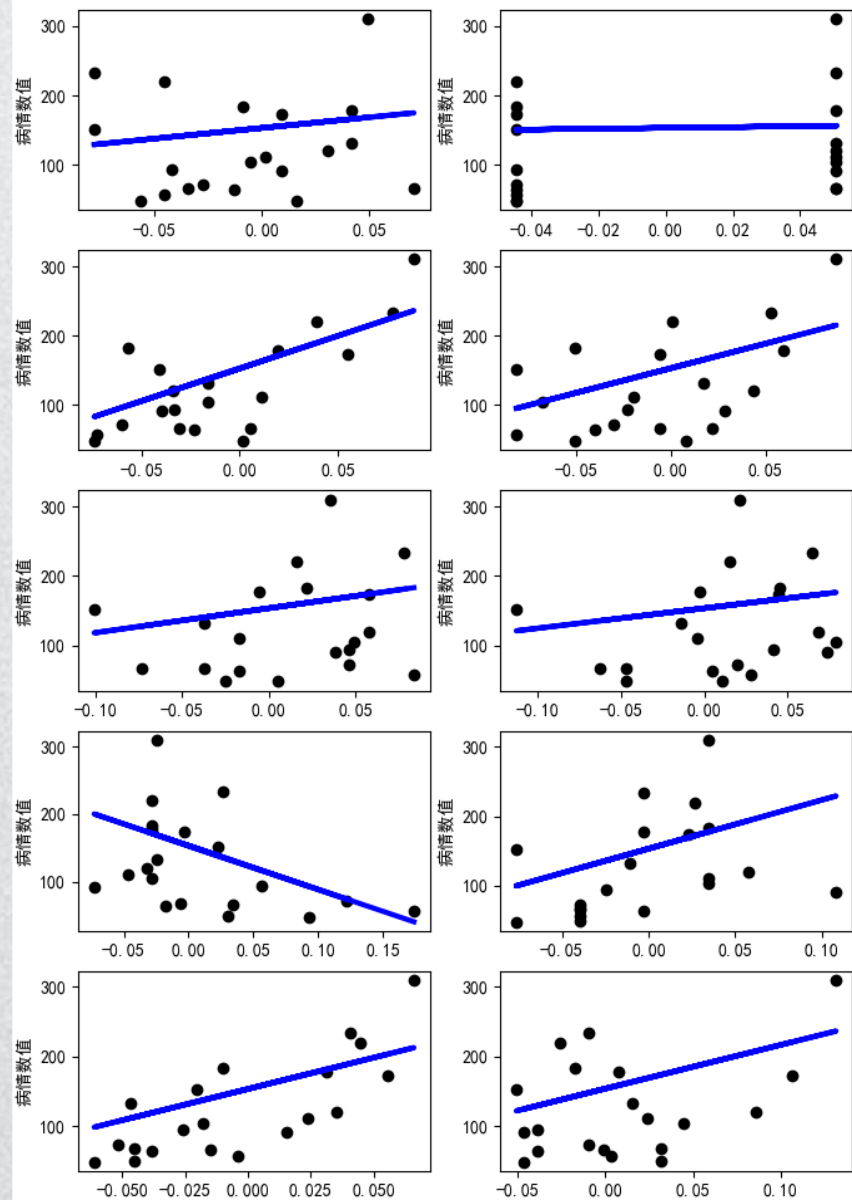
PART 03

## 糖尿病预测

特征一共有442组10维：

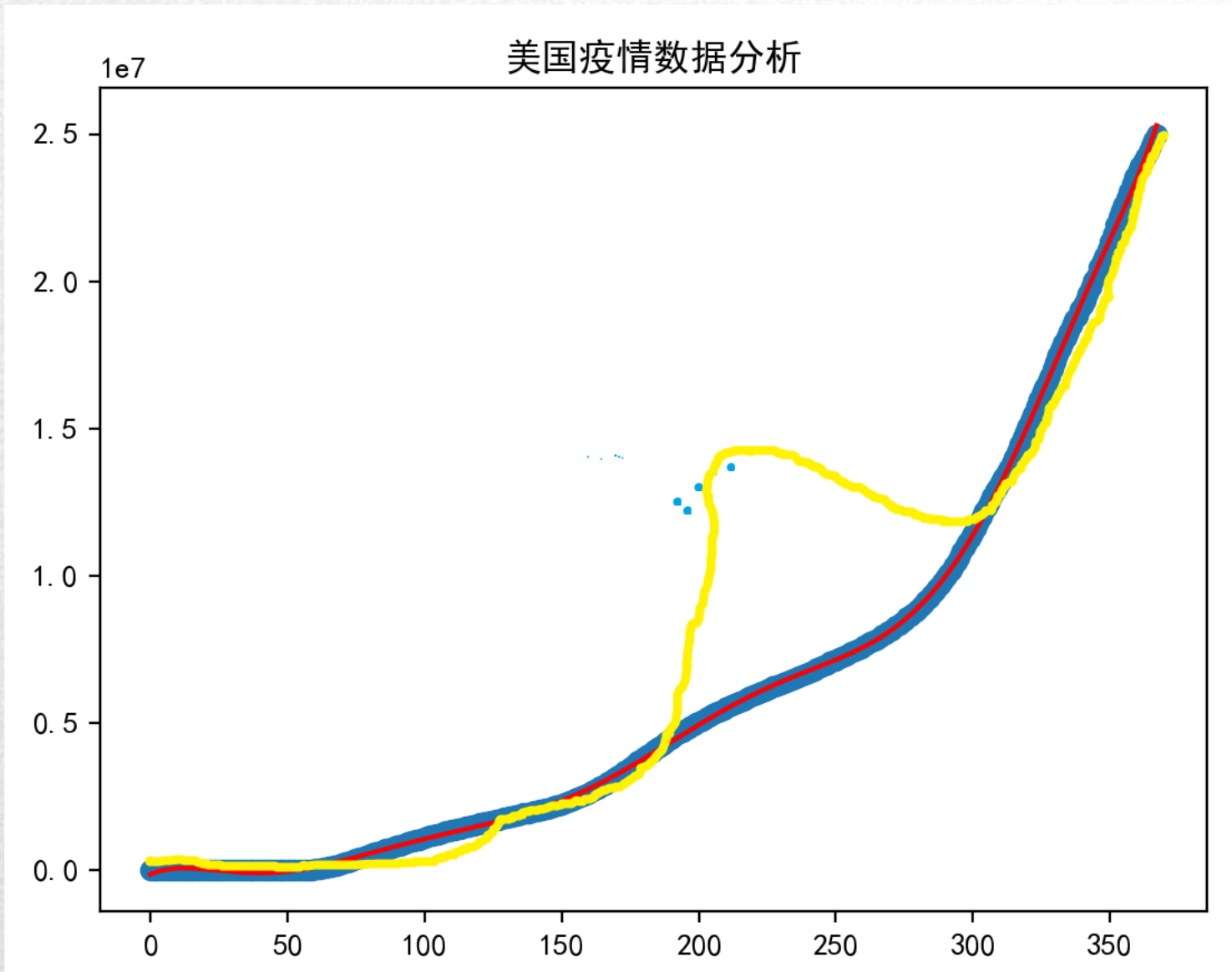
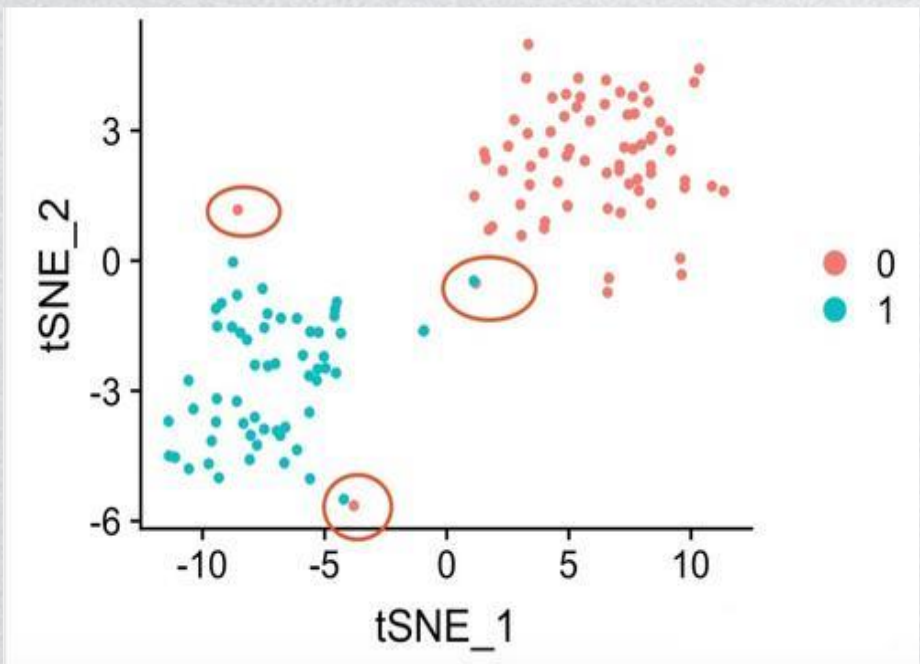
类型	个数
年龄	442
性别	
体质指数	
血压	
S1(血清的化验数据)	
S2(血清的化验数据)	
S3(血清的化验数据)	
S4(血清的化验数据)	
S5(血清的化验数据)	
S6(血清的化验数据)	





## 数据噪声、离群点

函数拟合噪声，模型过拟合








数据来源-医院检测结果

目标-根据已有结果，预测患病概率



暨南大学附属第一医院临床医学检验中心  
血液室报告单

姓名: 性别:女 年龄:57 岁 床号:30床 科别:胃肠外一区病房 标本号:553 标本类型:全血 打印时间:2018/09/18 11:26:12 病区:胃肠外病区(9F东) 病历号:476584

项目名称	结果	单位	参考范围
1 糖化血红蛋白(HbA1c)	7.2	↑ %	4--6.1

无可见异常

备注:

送检日期:2018/09/18 检验师:查显丰 检验日期:2018/09/18 11:26 审核者:李莉 送检医生: 本报告单仅对所检测的标本负责,如有疑问请及时与检验科联系!

病区:	床号:	信息提示:
检验项目	结果	参考值
空腹葡萄糖	4.60	3.90-6.10 mmol/L
餐后一小时葡萄糖	9.70	<10.00 mmol/L
餐后二小时葡萄糖	8.60	↑ <8.50 mmol/L



PART 04

## 手写数字识别

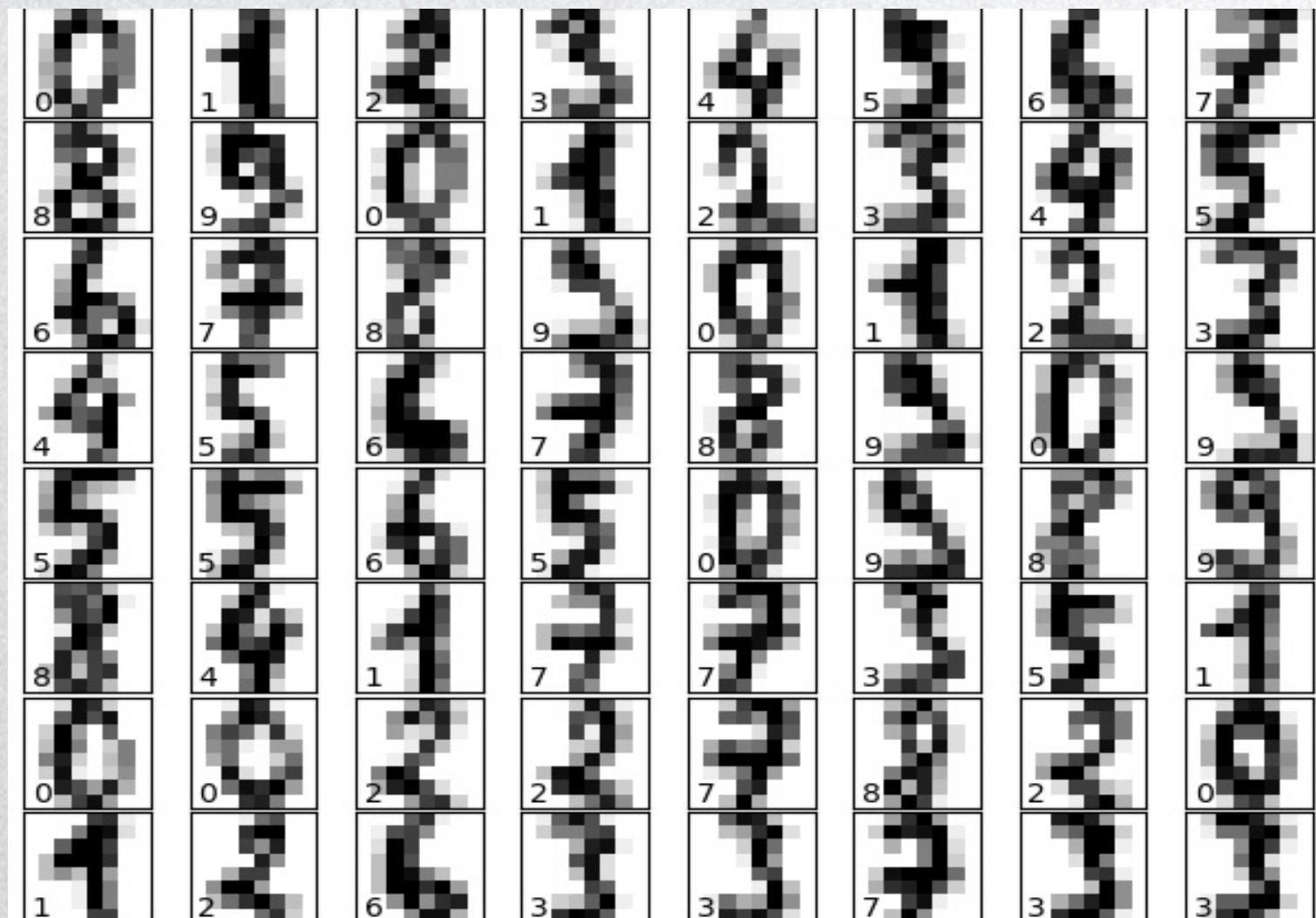


# 手写数字识别数据集含义

每个数据点是一个数字的8x8图像。

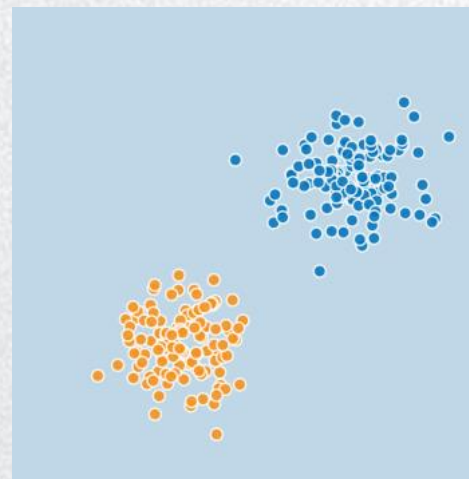
```
=====
分类数目                                10
每个分类的样本数                      大约180
样本总数                              1797
维度                                  64
=====
```

	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM	AN	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB	BC	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM	BN
	0	4	12	0	0	8	8	0	0	5	8	0	0	9	8	0	0	4	11	0	1	12	7	0	0	2	14	5	10	12	0	0	0	0	6	13	10	0	0	0	0	
	0	7	15	16	16	2	0	0	0	0	1	16	16	3	0	0	0	0	1	16	16	6	0	0	0	0	1	16	16	6	0	0	0	0	0	11	16	10	0	0	1	
	0	0	1	6	15	11	0	0	0	1	8	13	15	1	0	0	0	9	16	16	5	0	0	0	0	3	13	16	16	11	5	0	0	0	0	3	11	16	9	0	2	
	0	0	2	15	11	1	0	0	0	0	0	1	12	12	1	0	0	0	0	0	1	10	8	0	0	0	8	4	5	14	9	0	0	0	7	13	13	9	0	3		
	0	0	7	15	0	9	8	0	0	5	16	10	0	16	6	0	0	4	15	16	13	16	1	0	0	0	0	3	15	10	0	0	0	0	0	2	16	4	0	4		
	0	0	11	16	16	7	0	0	0	0	0	4	7	16	7	0	0	0	0	0	4	16	9	0	0	0	5	4	12	16	4	0	0	0	9	16	16	10	0	5		
	0	0	14	13	0	0	0	0	0	0	15	12	7	2	0	0	0	0	13	16	13	16	3	0	0	0	7	16	11	15	8	0	0	0	1	9	15	11	3	6		
	0	4	8	8	15	15	6	0	0	2	11	15	15	4	0	0	0	0	0	16	5	0	0	0	0	0	9	15	1	0	0	0	0	0	13	5	0	0	7			
	0	0	3	16	12	14	2	0	0	0	4	16	16	2	0	0	0	3	16	8	10	13	2	0	0	1	15	1	3	16	8	0	0	0	11	16	15	11	1	8		
	0	1	16	1	12	15	0	0	0	0	13	16	9	15	2	0	0	0	0	3	0	9	11	0	0	0	0	0	9	15	4	0	0	0	9	12	13	3	0	9		
	0	1	16	4	0	8	8	0	0	4	16	4	0	8	8	0	0	1	16	5	1	11	3	0	0	0	12	12	10	10	0	0	0	0	1	10	13	3	0	0		
	0	1	10	16	16	12	0	0	0	3	12	14	16	9	0	0	0	0	0	5	16	15	0	0	0	0	0	4	16	14	0	0	0	0	0	1	13	16	1	1		
	0	2	10	0	14	0	0	0	0	0	2	0	16	1	0	0	0	0	0	6	15	0	0	0	0	0	9	16	15	9	8	2	0	0	3	11	8	13	12	4	2	
	0	0	0	11	14	2	0	0	0	0	0	2	15	11	0	0	0	0	0	0	2	15	4	0	0	1	5	6	13	16	6	0	0	2	12	12	13	11	0	3		
	0	7	16	7	1	16	8	0	0	9	16	13	14	16	5	0	0	1	10	15	16	14	0	0	0	0	0	1	16	10	0	0	0	0	10	15	4	0	4			
	0	8	16	16	14	0	0	0	0	1	6	6	16	0	0	0	0	0	0	5	16	3	0	0	0	1	5	15	13	0	0	0	4	15	16	2	0	0	5			
	0	0	6	16	2	0	0	0	0	0	7	16	16	13	5	0	0	0	15	16	9	9	14	0	0	0	3	14	9	2	16	2	0	0	0	7	15	16	11	0	6	





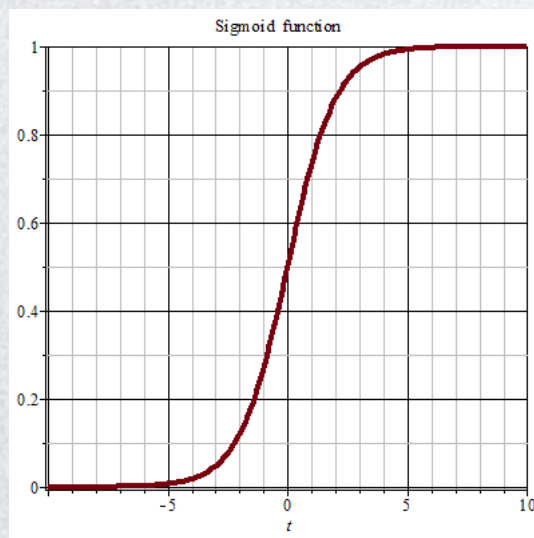
逻辑回归**Logistic Regression**，是一种广义的线性回归分析模型，主要是解决分类问题，常用于数据挖掘，疾病自动诊断，经济预测等领域。例如，探讨引发疾病的危险因素，并根据危险因素预测疾病发生的概率等。以胃癌病情分析为例，选择两组人群，一组是胃癌组，一组是非胃癌组，两组人群必定具有不同的体征与生活方式等。因此因变量就为是否胃癌，值为“是”或“否”，自变量就可以包括很多了，如年龄、性别、饮食习惯、幽门螺杆菌感染等。根据不同的特征值作为不同的致病因素预测一个人患癌症的可能性。



逻辑回归与线性回归有很多相同之处。它们的模型形式基本上相同，都具有  $w*x+b$ ，其中  $w$  和  $b$  是待求参数，其区别在于他们的因变量不同，线性回归直接将  $w*x+b$  作为因变量，即  $y = w*x+b$ ，而 logistic 回归则通过函数  $L$  将  $w*x+b$  对应一个隐状态  $p$ ， $p = L(w*x+b)$ ，然后根据  $p$  与  $1-p$  的大小决定因变量的值。逻辑回归在线性回归的基础上，增加了激活函数  $y = \text{sigmoid}(w*x+b)$

Sigmoid函数：隐藏层神经元输出，将输出的Y值映射到(0,1)的区间，主要用来做二分类算法。

$$S(x) = \frac{1}{1 + e^{-x}}$$





## 逻辑回归的损失(LOSS)函数

损失函数交叉熵公式：

$$\sum_x p(x) \cdot \log \left( \frac{1}{q(x)} \right)$$

# 真实值: [p1,p2]

# 预测值: [q1,q2]

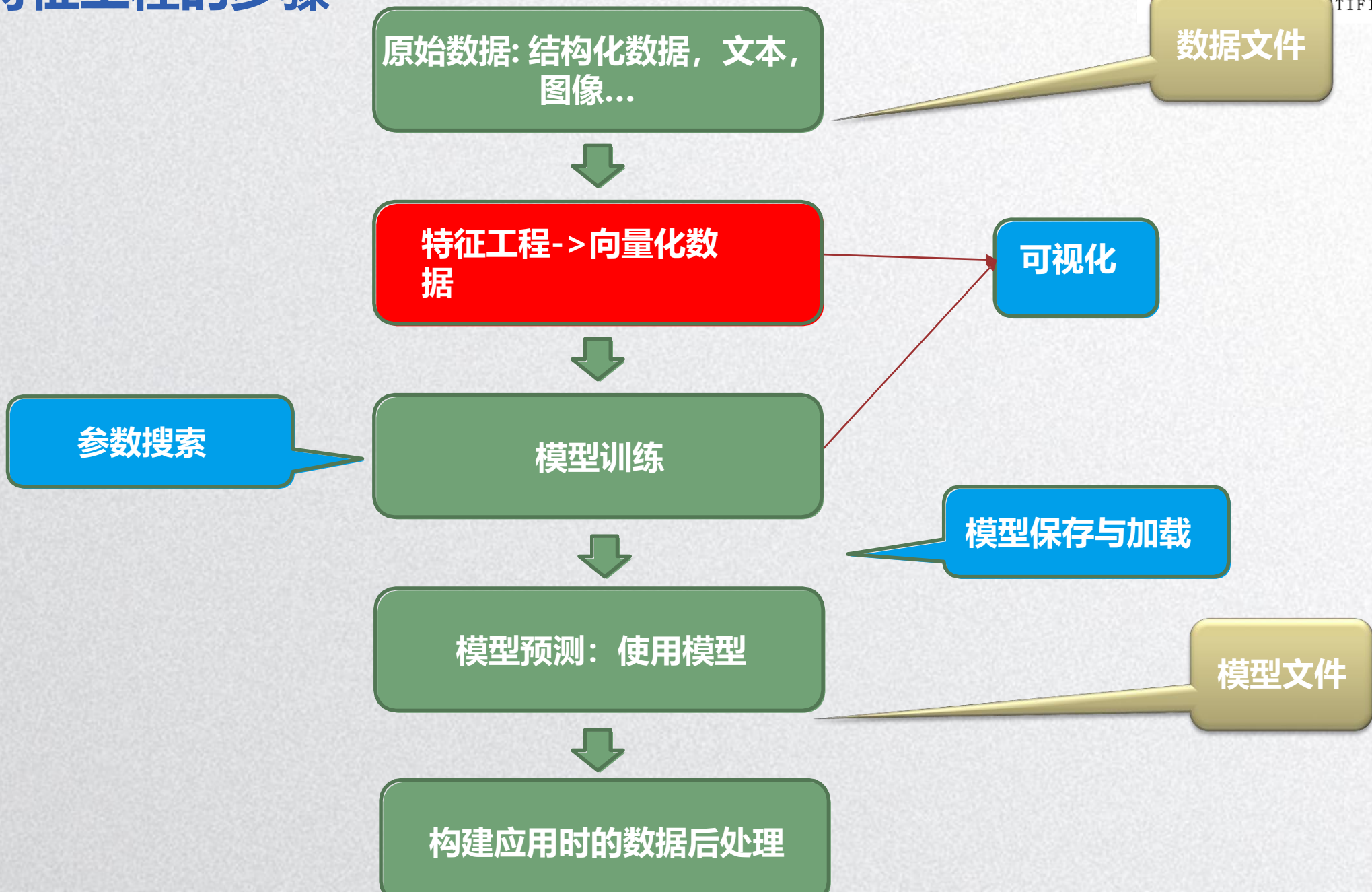
# -(p1\*np.log(q1)+p2\*np.log(q2))

**业界广为流传的一句话：  
“数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限而已。”**

**其本质是一项工程活动，目的是最大限度地  
从原始数据中提取特征以供算法和模型使用。**



# 特征工程的步骤



谢谢

