

综合素质培养课：人工智能时代的材料设计与优化  
Materials Design and Optimization in the Era of Artificial Intelligence

## 4. 数据处理与模型建立

### Data Preprocessing and Model Building

任课教师：刘哲

材料学院

2021年秋季学期



1. 数据驱动  
材料设计的  
背景

2. 材料数  
据的高效  
获取

3. 材料  
数据库  
及应用

4. 数据处  
理与模型  
建立

5. 数据驱  
动材料设  
计的案例

6. 数据驱  
动方法的  
实践

## ➡ 4. 1 数据可视化和预处理方法

### Data Visualization and Preprocessing

4. 2 分类模型：建模及其“准确性”评估

4. 3 回归模型：建模及其预测“误差”计算

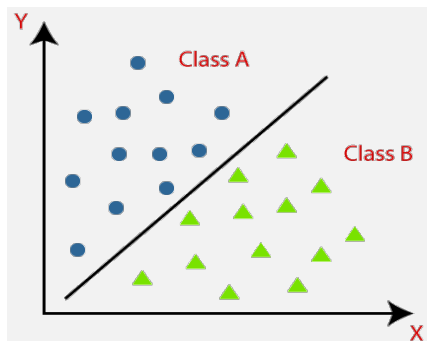
4. 4 模型剖析：打破机器学习的“黑箱”特性

# 机器学习模型能做什么事？



西北工业大学  
NORTHWESTERN POLYTECHNICAL UNIVERSITY

## 分类问题



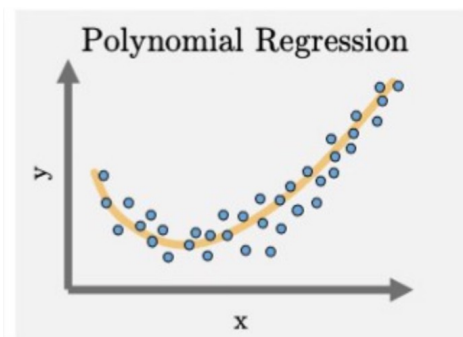
Classification

### 数字识别

将数字图像分为0-9共十个类别



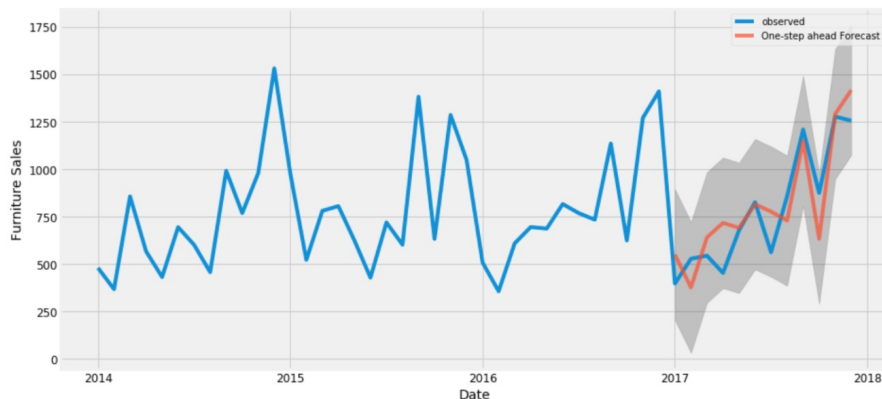
## 回归问题



Regression

### 对未来预测

根据以往数据对营业额走势进行预测



# 4.1 数据可视化和预处理方法



主要内容：

- 为什么要将数据可视化？
- 不同可视化方法有什么特点？
- 如何在Python中实现典型的图表方法？

- 数据可视化主旨在于：借助于图形化手段，清晰有效地传达与沟通信息
- 有效数据可视化的要素是：
  - 清楚地展示数据（尤其是展示数据的主要内涵）
  - 有目的性的制图
  - 以最简单地方式表示
  - 避免扭曲数据

Edward Tufte, The Visual Display of Quantitative Information, 1981.

# 数据可视化图表的基本构成



西北工业大学  
NORTHWESTERN POLYTECHNICAL UNIVERSITY

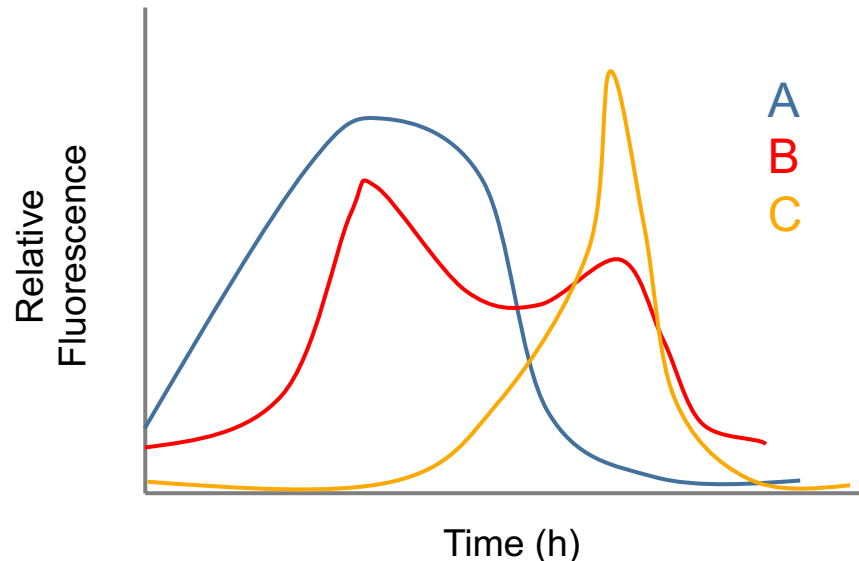
图表 = 信息+数据 (Figure = Message + Data)

选择数据  
(Choice of data)

展示手段  
(Presentation)

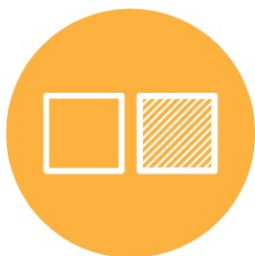
标题  
(Title)

注解  
(Caption)



**Fig. 1: A, B, and C have different dynamics under Condition X.** A, B, and C were sampled using Method 1 and their fluorescence quantified with Method 2. Fluorescence data normalized to negative control.

# 制图目的性：要显示什么？



比较



比例



关系



层次结构



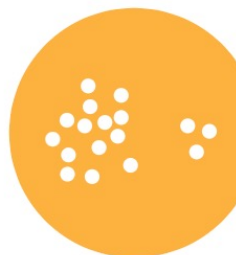
概念



位置



部分对整体



分布



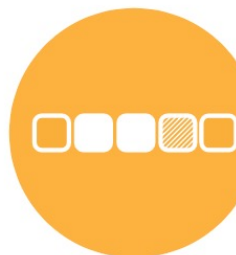
运作模式



过程与方法

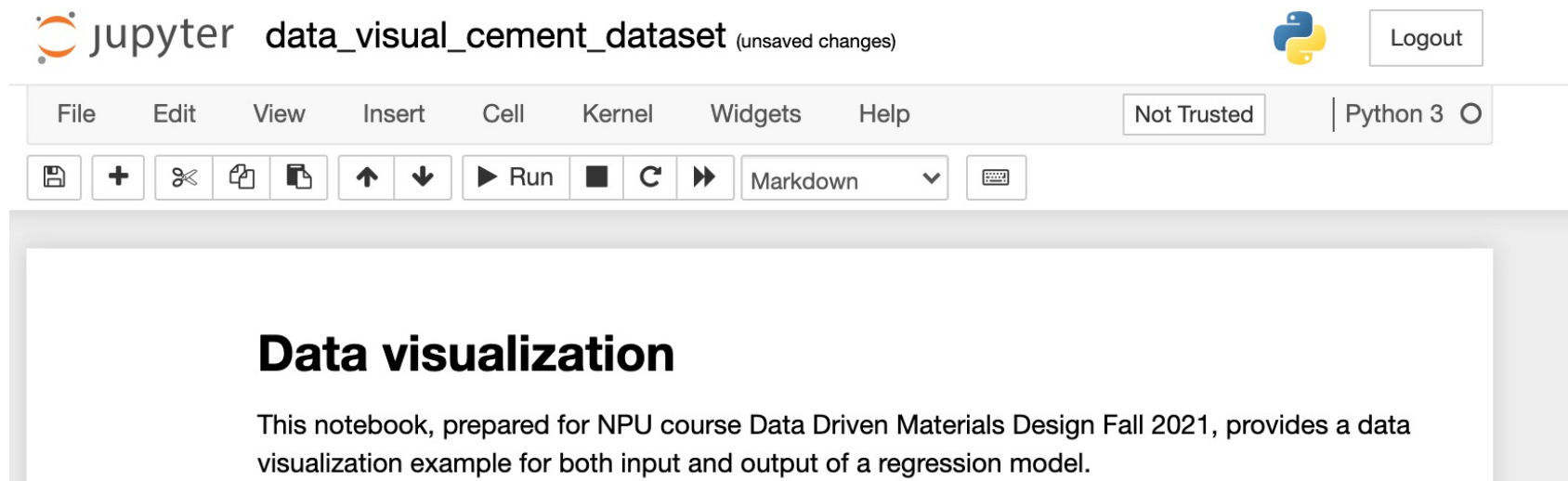



移动或流程



模式

- Jupyter Notebook演示



jupyter data\_visual\_cement\_dataset (unsaved changes)  Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted | Python 3

Save Add Cut Copy Paste Up Down Run Stop Restart Markdown

## Data visualization

This notebook, prepared for NPU course Data Driven Materials Design Fall 2021, provides a data visualization example for both input and output of a regression model.



- 在`sklearn.preprocessing`<sup>1</sup>中的常用方法
  - 归一化 `MinMaxScaler`: 把数据变为 0 – 1 的区间分布
  - 标准化 `StandardScaler`: 适用于近高斯分布，把分布变为均值为0，方差变为1
  - 注意：当数据有离群值（outlier）时，可能会出现问题<sup>2</sup>
  - 非连续数据数据处理 – 自然编码 vs 独热码（one hot encoding）
    - 红 = 1, 黄 = 2, 蓝 = 3
    - 红色: 1 0 0 , 黄色: 0 1 0, 蓝色: 0 0 1

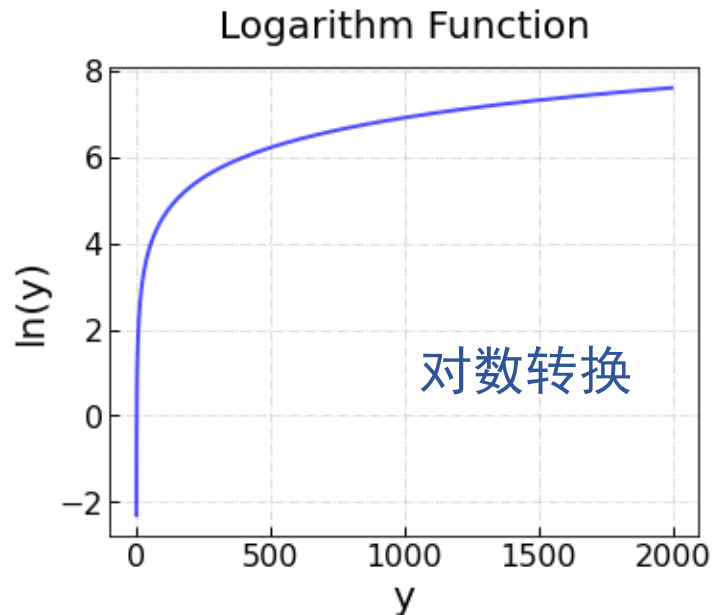
<sup>1</sup> <https://scikit-learn.org/0.20/modules/classes.html#module-sklearn.preprocessing>, accessed on Oct 14 2021

<sup>2</sup> [https://scikit-learn.org/0.20/auto\\_examples/preprocessing/plot\\_all\\_scaling.html#sphx-gl-auto-examples-preprocessing-plot-all-scaling-py](https://scikit-learn.org/0.20/auto_examples/preprocessing/plot_all_scaling.html#sphx-gl-auto-examples-preprocessing-plot-all-scaling-py), accessed on Oct 14 2021

- 自定义方程：

```
def log_transform(x):  
    return log (x)
```

（伪代码）



- 注意别忘了定义：逆转换方程：

```
def log_inverse_transform(x'):  
    return exp(x')
```

（伪代码）

对数逆转换 → 指数

- 什么模型下需要或不需要归一化（normalization）？
- 将输入变量 $X$  改为0-1的区间

Yes, 需要	No, 不需要
基于参数的模型或基于距离的模型，通常都是要进行特征的归一化。例如，神经网络，高斯过程回归等	基于树的方法是不需要进行特征的归一化，例如，随机森林，梯度提升等。

### 数据可视化和预处理方法

- 数据可视化是了解数据特点的必要手段
- 常用的图表工具包括：Matplotlib、Seaborn、Plotly
- 数据可视化既是一门艺术，也是一门科学
- 根据数据特点和模型需求，选择合适的预处理方式
  - 考量数据分布、模型特性

1. 数据驱动  
材料设计的  
背景

2. 材料数  
据的高效  
获取

3. 材料  
数据及  
应用

4. 数据处  
理与模型  
建立

5. 数据驱  
动材料设  
计的案例

6. 数据驱  
动方法的  
实践

## 4. 1 数据可视化和预处理方法

## ➔ 4. 2 分类模型：建模及其“准确性”评估

## 4. 3 回归模型：建模及其“误差”计算

## 4. 4 模型剖析：打破机器学习的“黑箱”特性

## 4.2 分类模型 (Classifier)



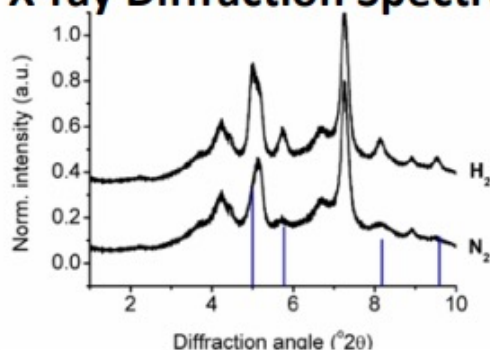
主要内容：

- 在材料领域，分类模型能做什么事？
- 如何快速建立一个机器学习分类模型？
- 怎样评价一个分类模型的优劣？

## ➤ 材料学科领域的分类问题

- 通过XRD光谱，将材料结构分类为230个空间群

X-ray Diffraction Spectrum



Classify into one of 230 Space Groups



- 预测反溶剂是否实现溶质析出结晶（两类）

反溶剂特征描述

	Solvent	Formula	$\delta D$ [Disperse or non-Polar]	$\delta P$ [Polar bond]	$\delta H$ [Hydrogen bond]	
						Order
1	Pentane	<chem>CH3(CH2)3CH3</chem>	14.5	0	0	
2	Hexane	<chem>CH3(CH2)4CH3</chem>	14.9	0	0	
3	Heptane	<chem>CH3(CH2)5CH3</chem>	15.3	0	0	
4	Octane	<chem>CH3(CH2)6CH3</chem>	15.5	0	0	



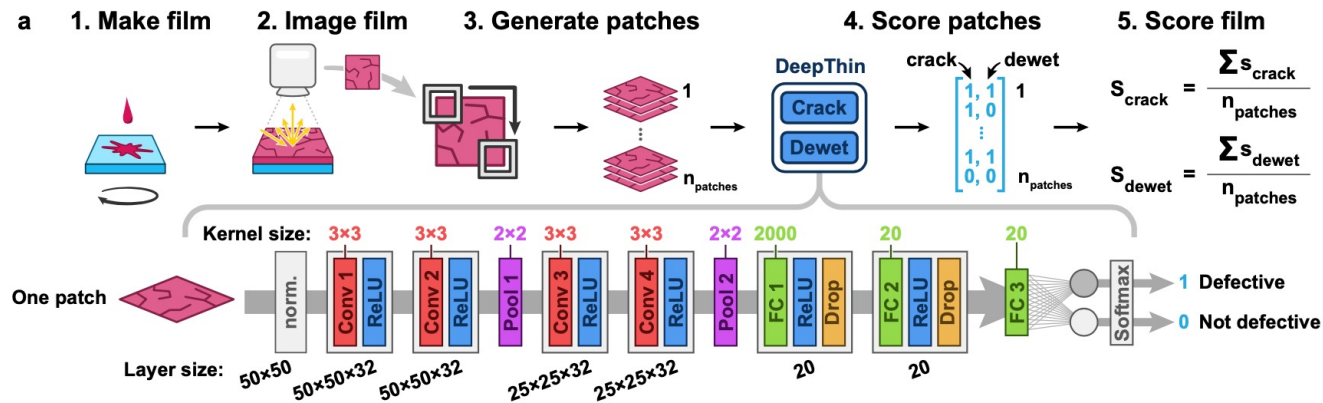
能结晶 ✓

不能结晶 ✗

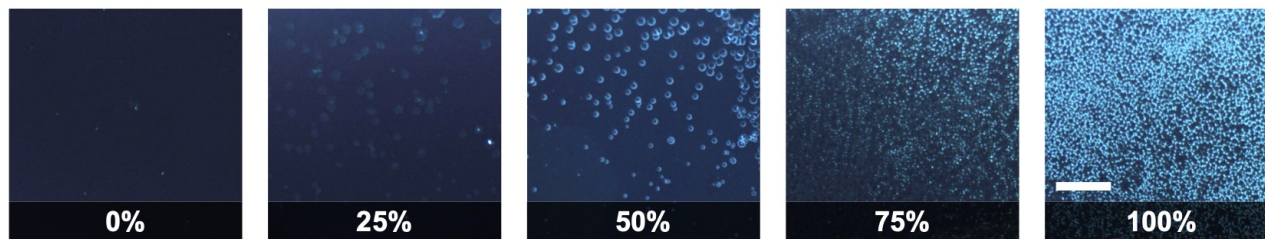
ARTICLE OPEN



## Quantifying defects in thin films using machine vision



**b** Example of dewet scores










- 数字识别分类 - Jupyter Notebook演示

 Jupyter Classification\_demo (unsaved changes)

File Edit View Insert Cell Kernel Widgets Help

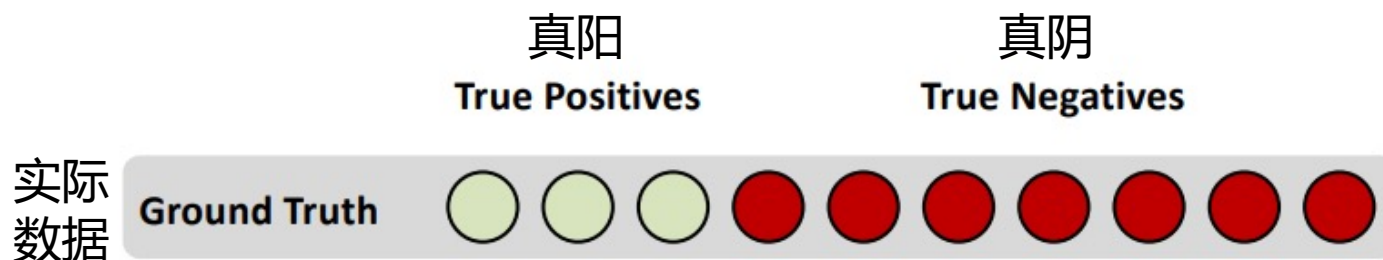
        Run    Code  

## Classification of hand-written digits using scikit-learn

An example showing how we can classify hand-written digits using scikit-learn. The example is adapted from

[https://scikit-learn.org/stable/auto\\_examples/classification/plot\\_digits\\_classification.html](https://scikit-learn.org/stable/auto_examples/classification/plot_digits_classification.html)

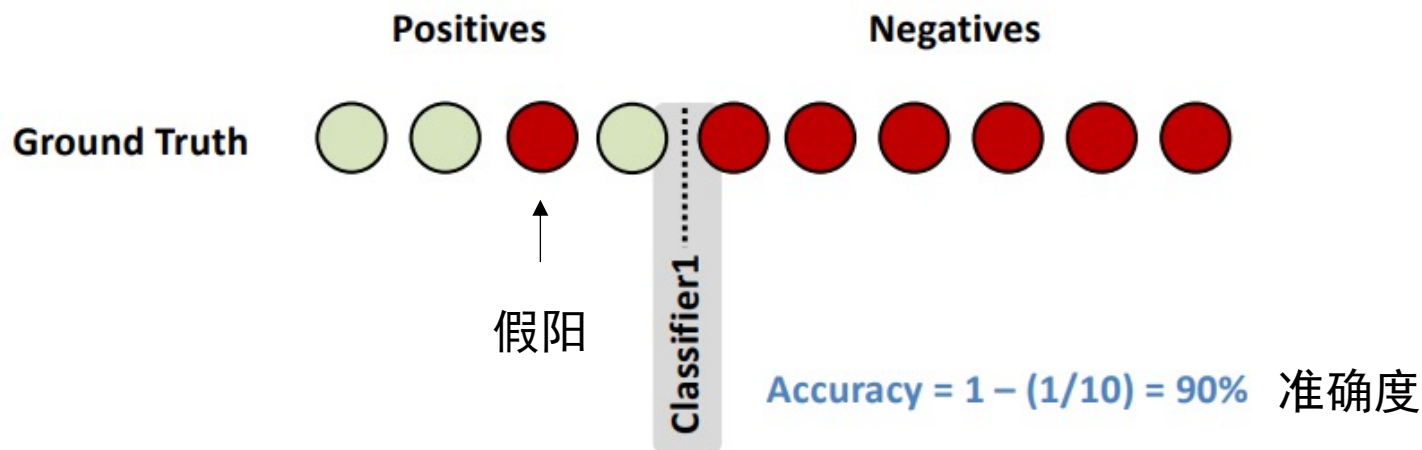
- 假设我们研发了一种通过测试新冠病毒的新方法
- 为加快数据分析，我们建立了一个分析数据的机器学习分类器



# 评估模型（一）：准确率

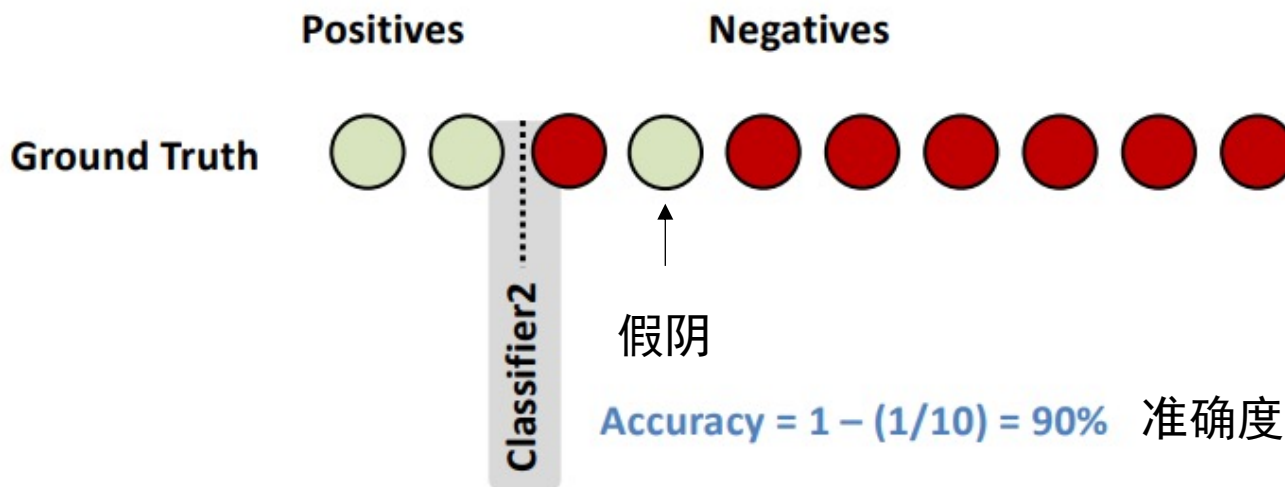


西北工业大学  
NORTHWESTERN POLYTECHNICAL UNIVERSITY



都是90%的准确度，

哪个模型更好？

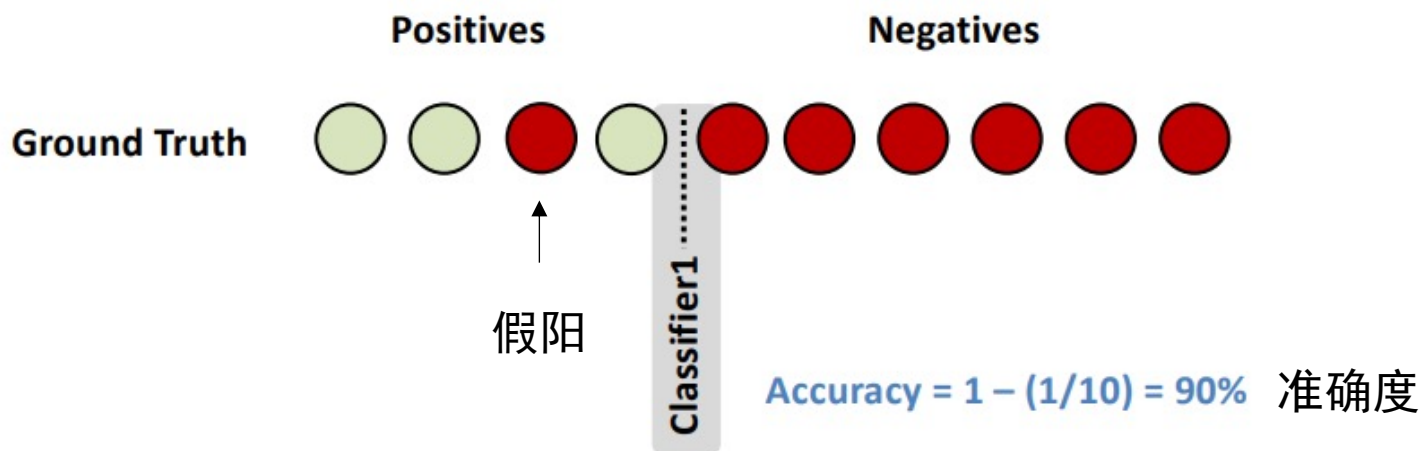


## 评估模型（二）：精确率



➤ 概念定义：测出的阳性案例中，真阳性的比例是多少？

➤ 数学定义：精确率(precision) =  $\frac{\text{真阳 (3)}}{\text{真阳 (3)} + \text{假阳 (1)}}$



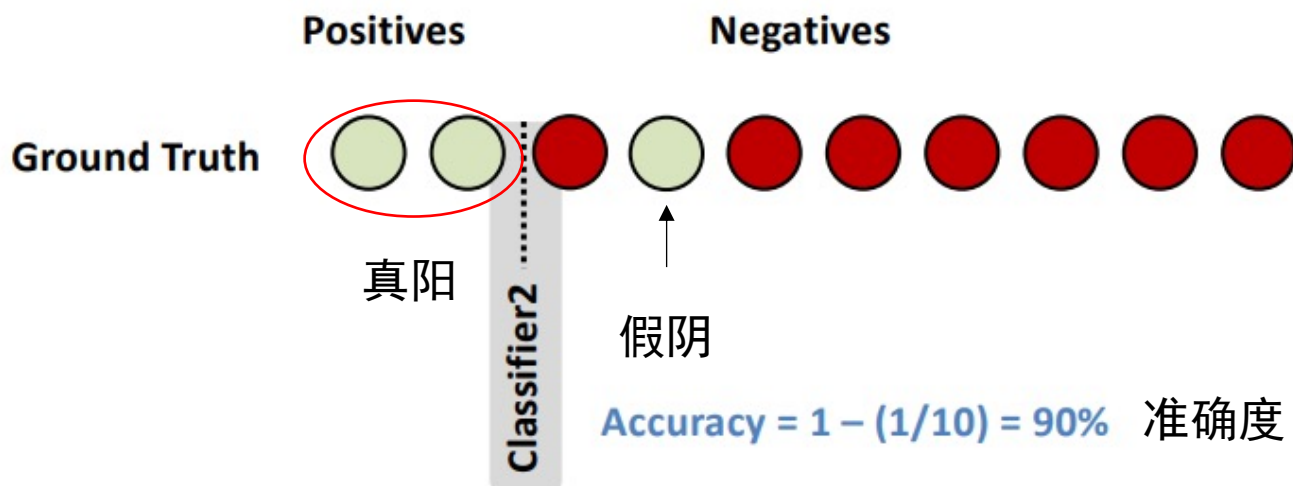
➤ 因此，第一个分类器模型的精确率为75%。意味着：当一个人测试的结果为阳性时，这个人有25%的可能性没有携带新冠病毒。

# 评估模型（三）：召回率



- 概念定义：能够检测出的真阳性的比例是多少？

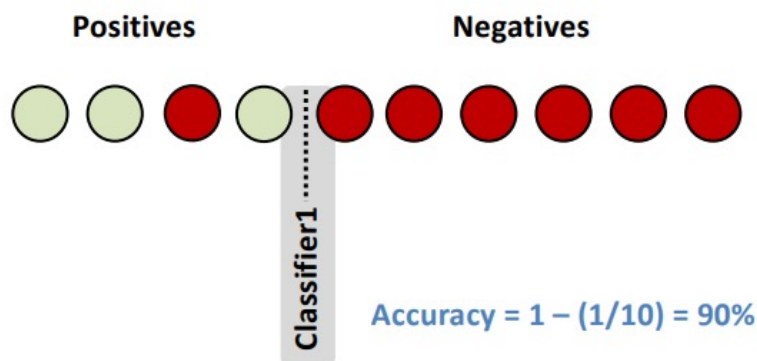
- 数学定义：召回率(recall) = 
$$\frac{\text{真阳 (2)}}{\text{真阳 (2)} + \text{假阴 (1)}}$$



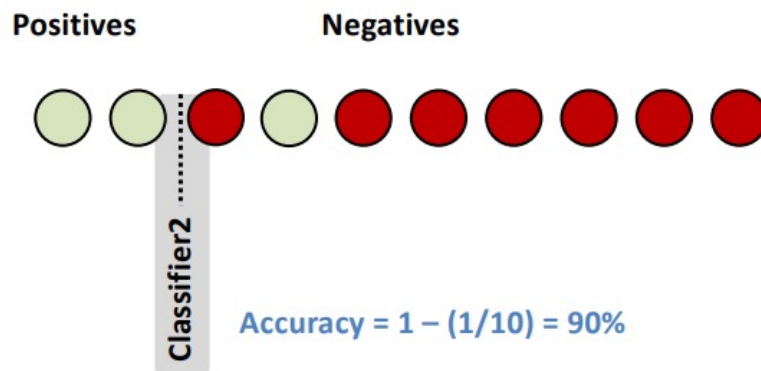
- 因此，第二个分类器模型的召回率是67%。意味着：人群中的新冠病毒携带者仅有67%将被检出，而33%的携带者将被漏掉。

- 对内（研发团队）：提供量化指标，方便继续优化和提高
- 对外（使用者、决策者）：说明模型结果的实际含义（没有100%的模型

反思：  
对社会的影响？



精确率 = 75%  
25%的可能性误判为阳性



召回率 = 67%  
33%的携带者将被漏掉

## ➤ 讨论不同测试方法中的精确率、召回率对结论的影响？

90%  
的  
准  
确  
度

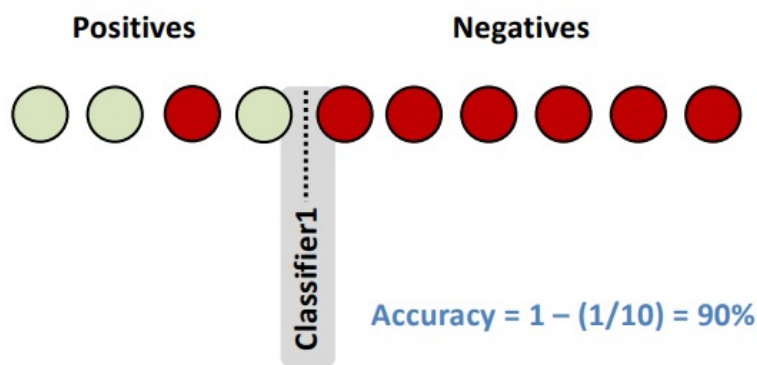
- 模型1（精确率75%）：当结果为阳性时，有25%的可能性没有携带新冠病毒
- 模型2（召回率67%）：新冠病毒携带者中有33%将被漏掉。
- 如何选择模型？考虑那些因素？
- 预测结果导致怎样的“不公平”对待？
- 其他材料领域的类似问题？有何技术影响？有何社会影响？

（课堂讨论）

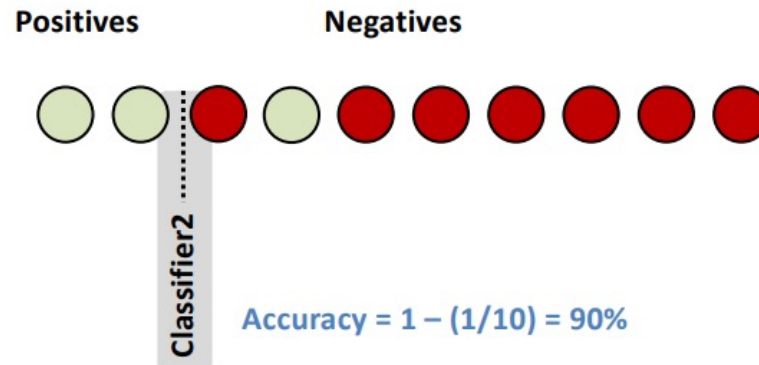
# 评估模型（四）：F1得分



- 概念定义：精确率和召回率的调和平均值（没有偏好时的选择）
- 数学定义：
$$F1\text{得分} = 2 * \frac{\text{精确率} * \text{召回率}}{\text{精确率} + \text{召回率}}$$
- 在对精确率和召回率没有特殊要求时，F1得分是一个常用的选择



$$\begin{aligned} F1\text{得分} &= 2 * 75\% * 100\% / (75\% + 100\%) \\ &= 86\% \end{aligned}$$



$$\begin{aligned} F1\text{得分} &= 2 * 100\% * 67\% / (67\% + 100\%) \\ &= 80\% \end{aligned}$$



## 4.2 要点小结



### 分类模型：建模及其“准确性”评估

- 分类模型是针对目标变量为离散/非连续的“类别”
- 机器学习分类模型能解决许多材料表征数据分析问题
- 评估指标的意义：准确率、精确率、召回率和F1得分
- 不要只看一个模型的单一指标，准确率有时非常片面

1. 数据驱动  
材料设计的  
背景

2. 材料数  
据的高效  
获取

3. 材料  
数据及  
应用

4. 数据处  
理与模型  
建立

5. 数据驱  
动材料设  
计的案例

6. 数据驱  
动方法的  
实践

4. 1 数据可视化和预处理方法

4. 2 分类模型：建模及其“准确性”评估

➡ 4. 3 回归模型：建模及其预测“误差”计算

4. 4 模型剖析：打破机器学习的“黑箱”特性

## 4.3 回归模型 (Regressor)



主要内容：

- 如何建立回归模型？建模流程是什么？
- 如何评估和比较不同模型的优劣？
- 如何结合不同模型？

数据可视化分析

- 关注 $X$  和  $y$  数据分布
- 利用直方图、t-SNE降维散点图

数据预处理

- 数据均匀化、正态化、归一化
- 预处理方法对建模影响

模型筛选与集成学习

- 模型自测误差对比及交叉验证优化
- 集成学习 - 多模型融合预测

- A、B、C三组数据集，分别含：
  - 训练集1000个样本（已知 $X$ 和 $y$ ）
  - 测试集200个样本（已知 $X$ ）
  - A添加0%扰动， B添加1%扰动， C添加5%扰动
  - $X$ ：5维数据变量； $y$ ： 单一目标变量
- 评估标准：
  - 降低模型预测的平均相对误差
  - 怎样计算误差？

- Mean Absolute Error 平均绝对误差

$$MAE = \frac{1}{N} \sum_{i=0}^N |y_{\text{pred}} - y_{\text{true}}|$$

- Mean Squared Error 平均平方误差

$$MSE = \frac{1}{N} \sum_{i=0}^N (y_{\text{pred}} - y_{\text{true}})^2$$

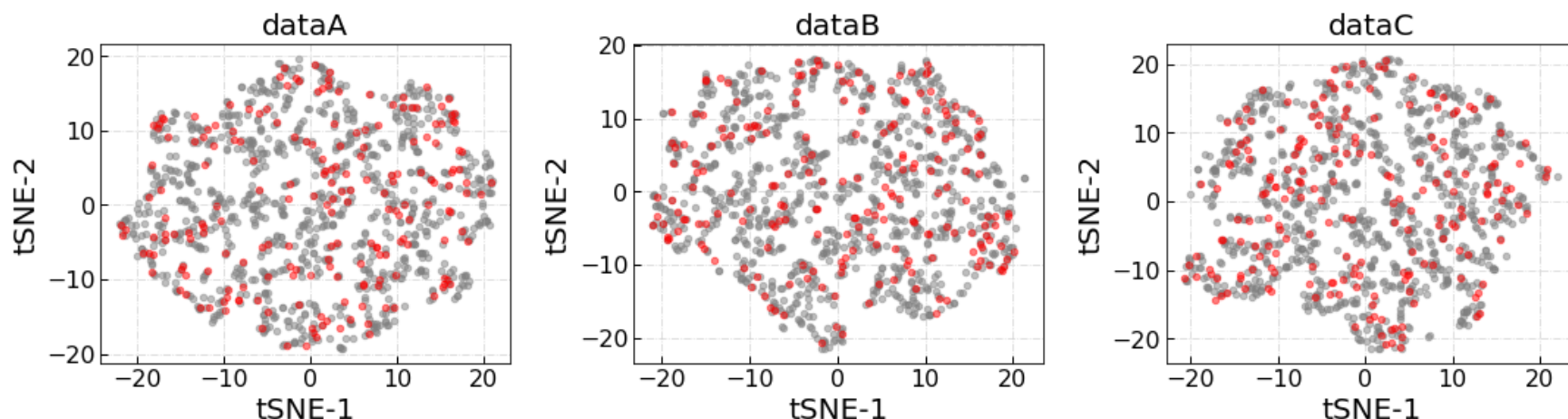
- Mean Absolute Percentage Error 平均相对误差

$$MAPE = \frac{1}{N} \sum_{i=0}^N \frac{|y_{\text{pred}} - y_{\text{true}}|}{y_{\text{true}}} \times 100\%$$

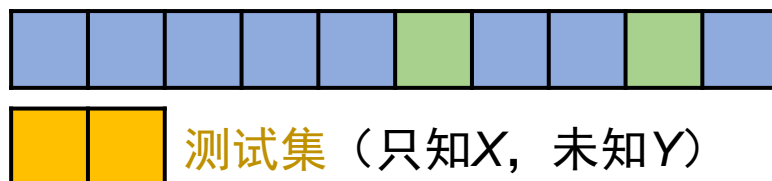
`sklearn.metrics`

## 训练集与测试集 $X$ 的分布

- 目的：训练集与测试集的数据分布是否一致？
- 用t-SNE方法降为后（5维 $X$ 到2维），画出分布图如下：



因此，可在1000个已知数据中随机选取20%作为自测集。



训练集 | 自测集（已知 $X$ 和 $Y$ ）

测试集（只知 $X$ ，未知 $Y$ ）

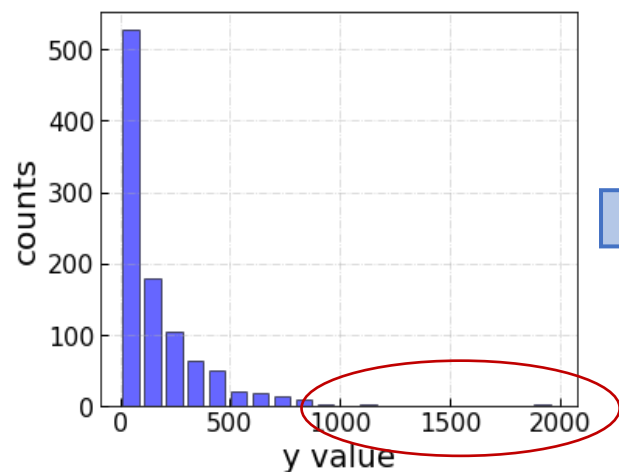
# 可视化分析（二）



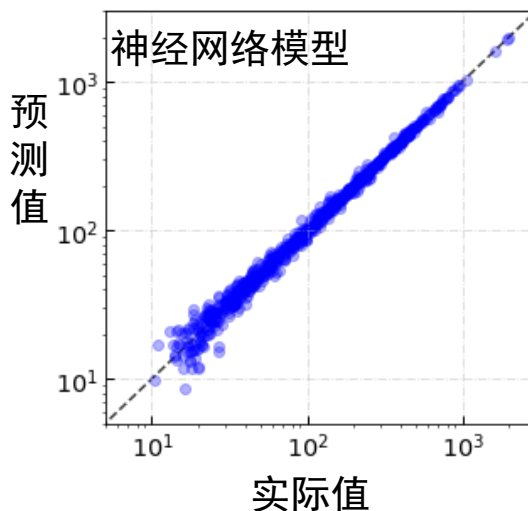
西北工业大学  
NORTHWESTERN POLYTECHNICAL UNIVERSITY

## 数据中 $y$ 值呈偏态分布

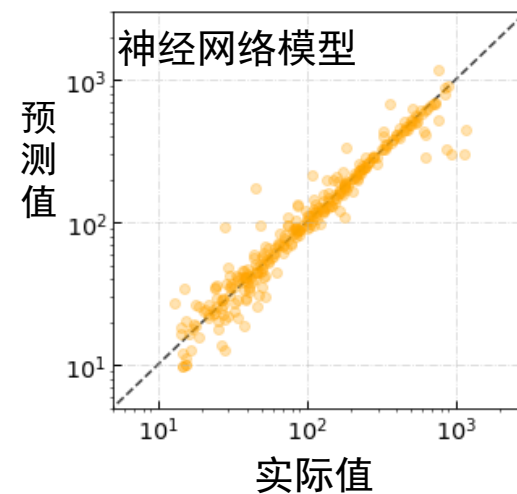
$y$  原始数据的直方图分布



80%训练集  
相对均误差 = 7%



20%自测集  
相对均误差 = 17%

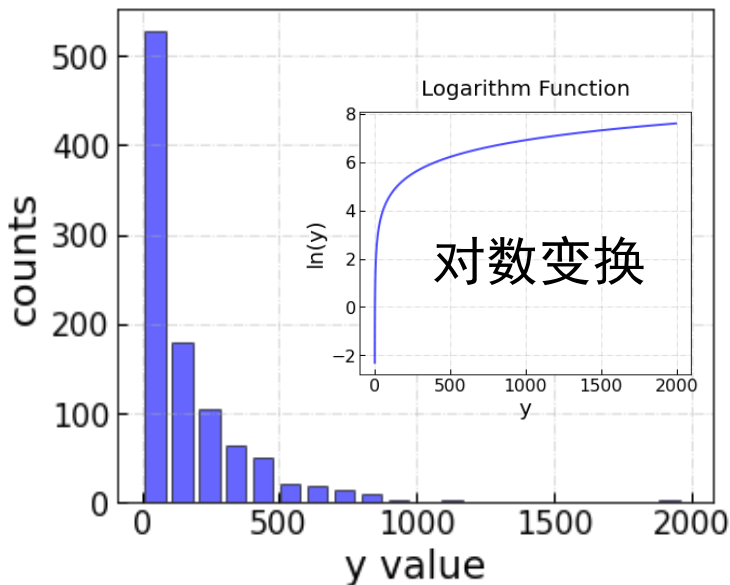


- $y$  值的偏态分布导致建模误差较大，尤其是较小的 $y$ 值

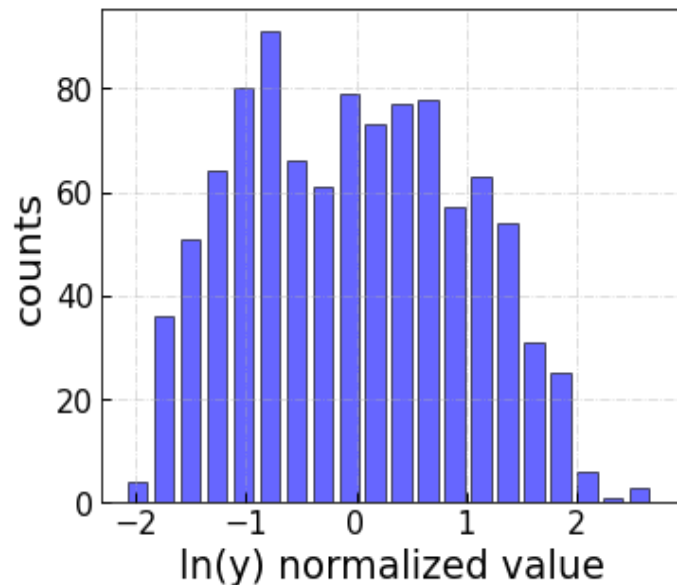


## 对数变换

$y$  原始数据的直方图分布



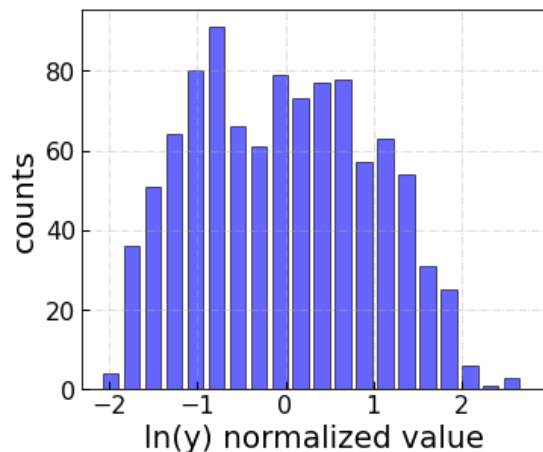
$\ln(y)$ +正态化数据的直方图分布



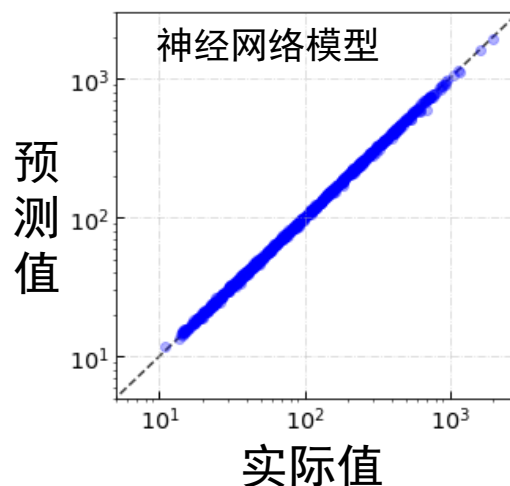
对数变换使得 $y$  值的分布均匀化（近正态分布）

## 对数变换的建模效果示例

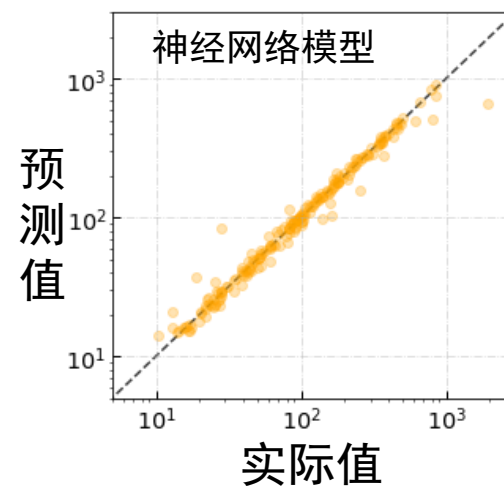
$\ln(y)$  + 正态化数据的  
分布直方图



80%训练集  
相对均误差 = 2%



20%自测集  
相对均误差 = 9%



- 对数变换显著提高建模准确度：自测集误差从17%降至9%

## ➤ Jupyter Notebook演示

 jupyter Building up regression model (unsaved changes)



Logout

File Edit View Insert Cell Kernel Widgets Help

Not Trusted

Python 3



### ML Regression Model

This notebook, prepared for NPU course Data Driven Materials Design Fall 2021, providing an example for building different ML regression models.

表一：预留自测数据的误差对比

模型类别	20%自测集 - 百分比相对误差均值MAPE		
	数据A	数据B	数据C
<b>X</b> 线性	37%	35%	31%
多项式（最高5次项）	11%	10%	12%
随机森林	14%	12%	14%
梯度提升决策树	10%	8%	10%
神经网络（三层）	9%	8%	10%
高斯过程（RBF内核）	11%	10%	11%
高斯过程（Matern52内核）	<b>5%</b>	<b>6%</b>	<b>9%</b>

- 含Matern52内核的高斯过程模型，实现自测误差最低

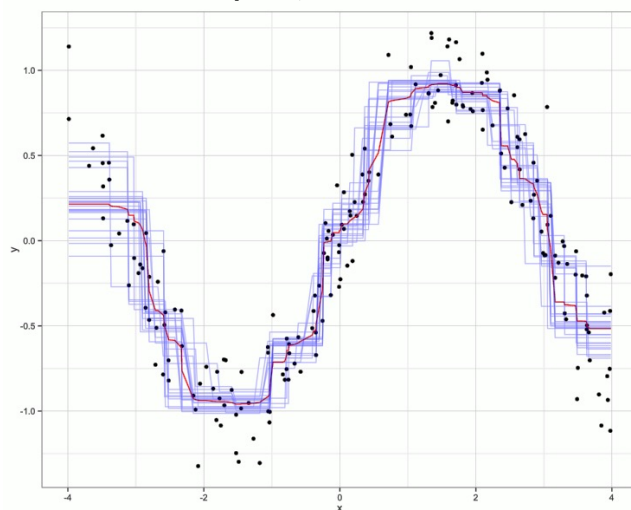
# 两种类型的模型比较



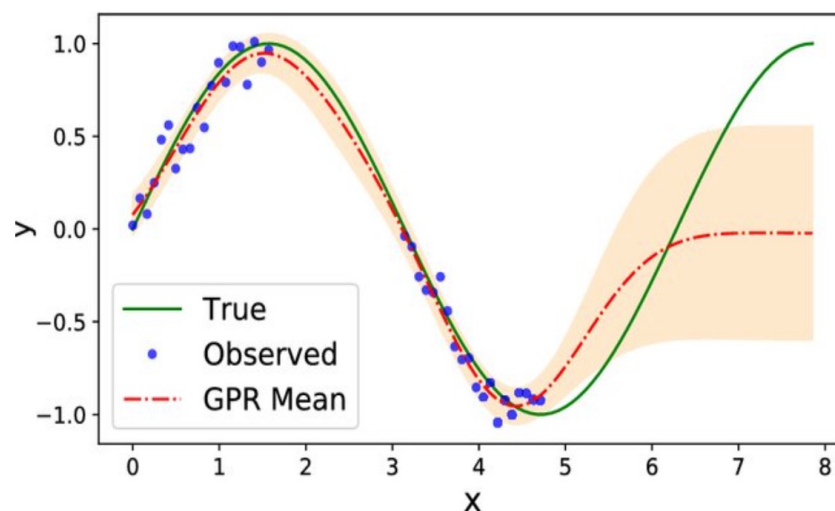
## 随机森林(RF)回归 vs 高斯过程(GP)回归

### ➤ $\sin(x)$ 拟合举例

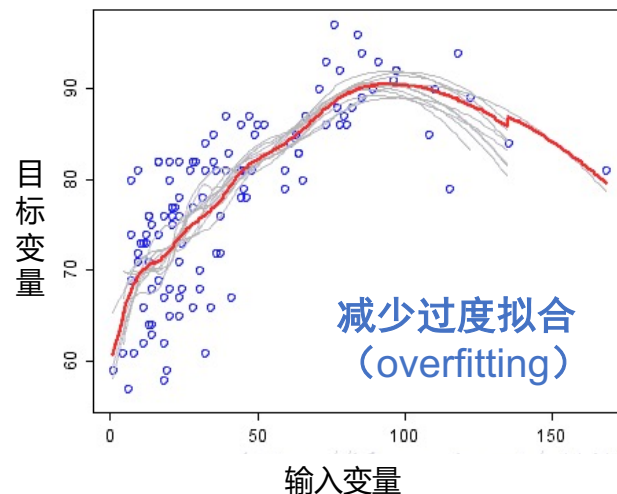
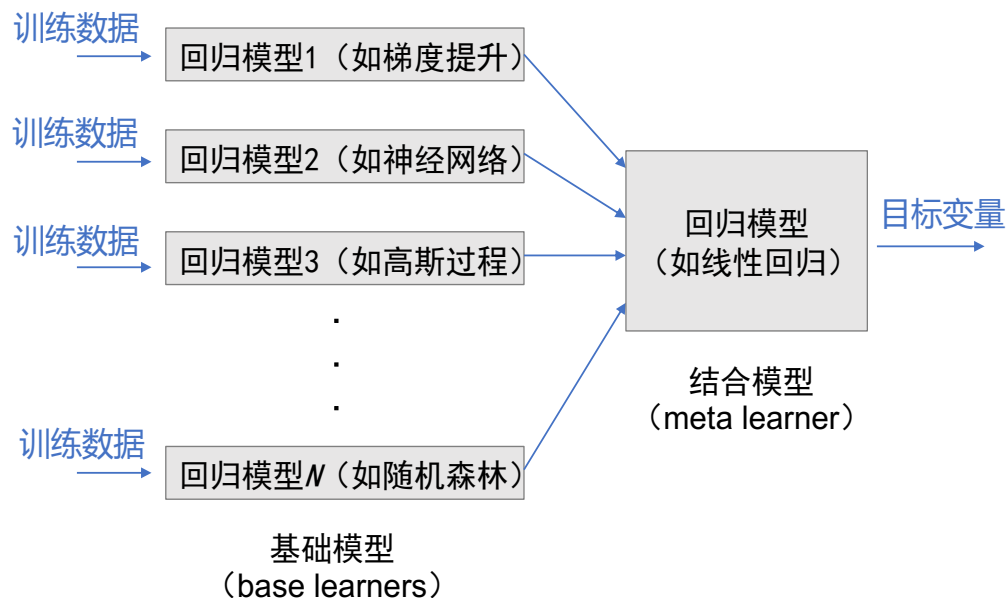
随机森林



高斯过程



# 集成学习(Ensemble Learning)



模型类别	20%自测集 - 百分比相对误差均值MAPE		
	数据A	数据B	数据C
集合学习 (线性结合)	5%	6%	8%
集合学习 (随机森林结合)	5%	5%	8%

### 4.3 回归模型

- 根据一组5对1的数据，进行了回归建模
- 比较了7种常用模型，并计算了每个模型的误差
- 集成学习可用于结合多种模型，防止单一模型过度拟合

1. 数据驱动  
材料设计的  
背景

2. 材料数  
据的高效  
获取

3. 材料  
数据及  
应用

4. 数据处  
理与模型  
建立

5. 数据驱  
动材料设  
计的案例

6. 数据驱  
动方法的  
实践

4. 1 数据可视化和预处理方法

4. 2 分类模型：建模及其“准确性”评估

4. 3 回归模型：建模及其预测“误差”计算

➡ 4. 4 模型剖析：打破机器学习的“黑箱”特性



## 4.4 模型剖析



主要内容：

- 打开黑箱：如何查看模型在“想”什么？
- 提取输入特征（即输入变量）的重要性
- 理解**模型中**输入变量与输出变量的关系

## 简单函数举例

### ➤ 线性模型

$$y = -10x_1 + 0.1x_2$$

$$x_1 = 1, x_2 = 1 \Rightarrow y = -9.9$$

$$x_1 = 1, x_2 = 100 \Rightarrow y = 0$$

$$x_1 = 100, x_2 = 1 \Rightarrow y = -999.9$$

### ➤ 二次多项式

$$y = \sqrt{x_1} + x_2^2$$

$$x_1 = 1, x_2 = 1 \Rightarrow y = 2$$

$$x_1 = 1, x_2 = 100 \Rightarrow y = 10001$$

$$x_1 = 100, x_2 = 1 \Rightarrow y = 11$$

## SHAP分析

- SHAP = Shapley Additive Explanation

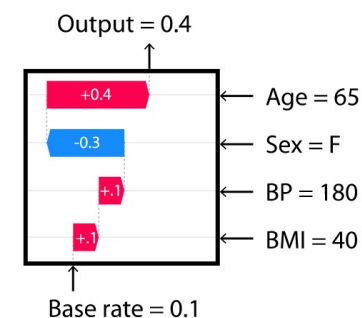
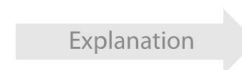
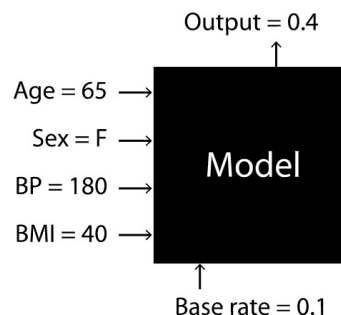


沙普利 (Shapley)

- 美国数学家、经济学家
- 博弈论专家
- 2012年获诺贝尔经济学奖



SHAP



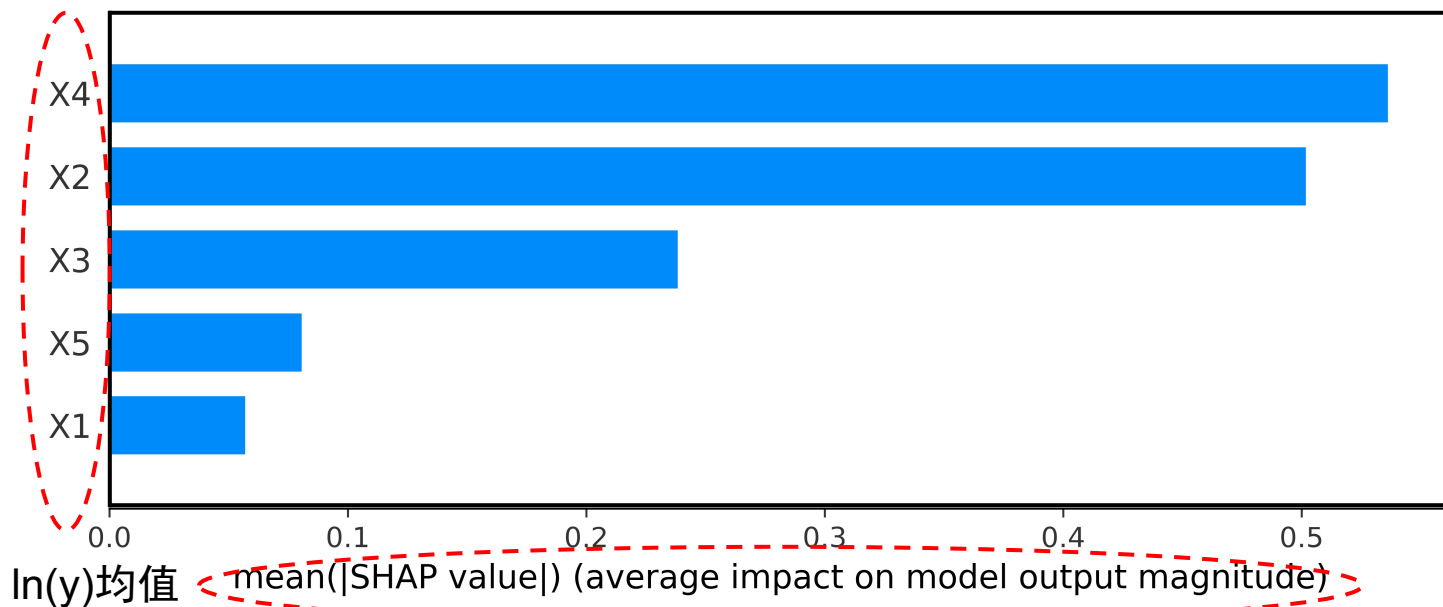
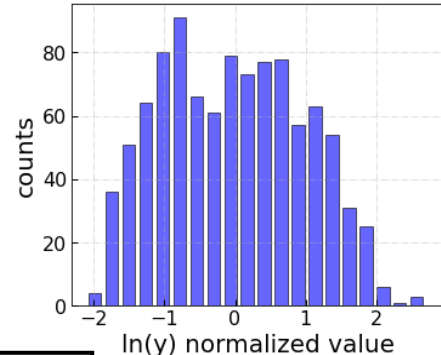
1. Lundberg & Lee. "A unified approach to interpreting model predictions." Advances in Neural Information Processing Systems. 2017.
2. SHAP: <https://github.com/slundberg/shap>, accessed on Oct 14 2021
3. 中文博客: <https://cloud.tencent.com/developer/news/624937>, accessed on Oct 14 2021

# 输入变量的重要性



## 平均SHAP值：特征重要性

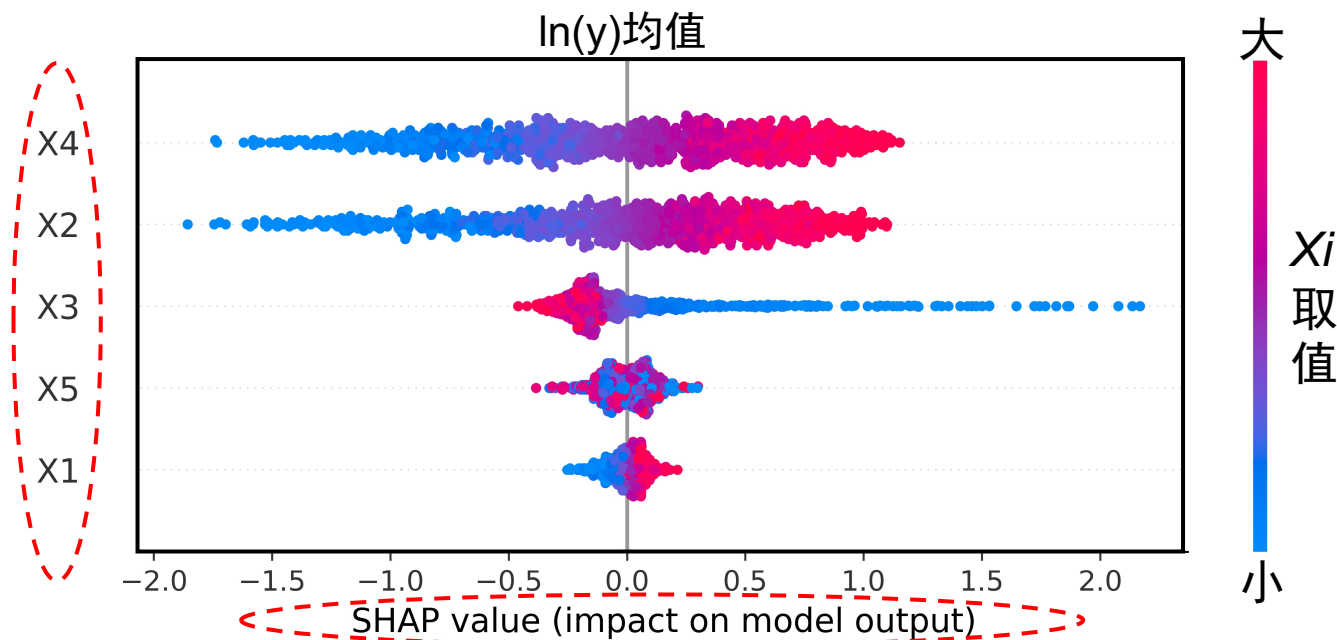
➤ 输入变量  $X_i$  对 目标变量  $\ln(y)$  的影响



$X_4$  和  $X_2$  对  $y$  预测比较重要

## SHAP值：单点影响

➤ 输入变量  $X_i$  的取值对目标变量  $\ln(y)$  的关系趋势

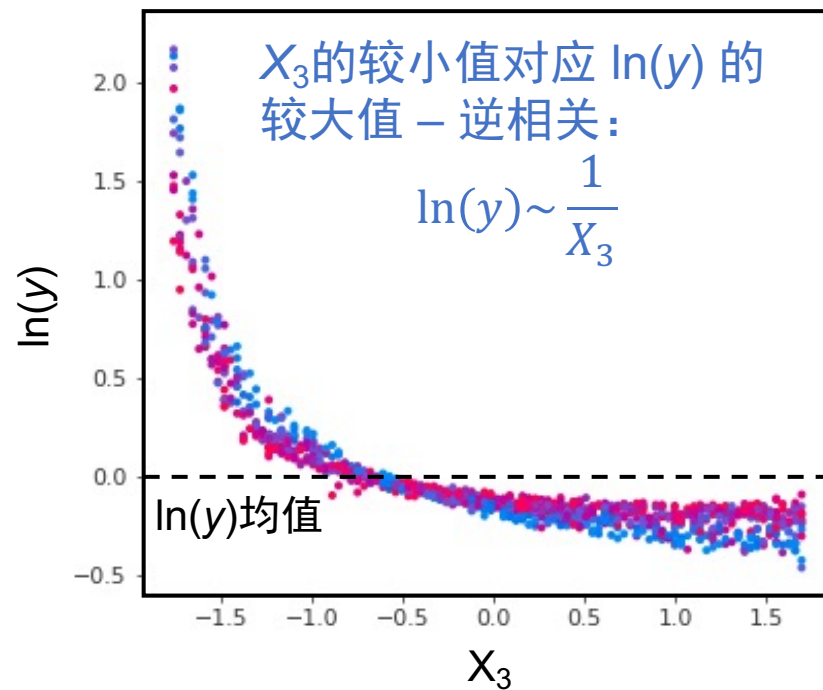
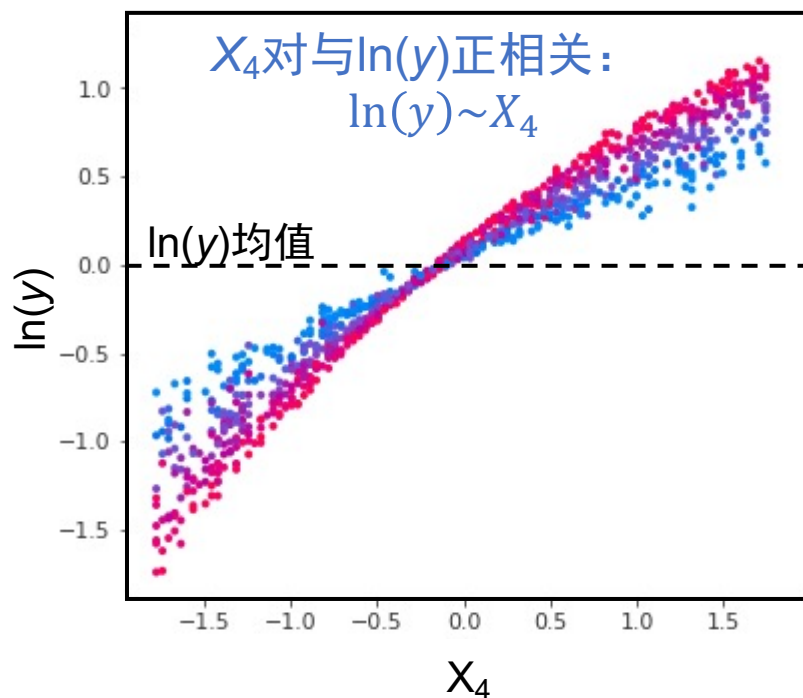


- $X_4$  和  $X_2$  与  $\ln(y)$  成正相关
- $X_3$  的较小值对应  $\ln(y)$  的较大值
- 具体函数关系怎样？

# X 各项与 $\ln(y)$ 的关系



## ➤ 查看单一变量 $X$ 与 $\ln(y)$ 的变量关系

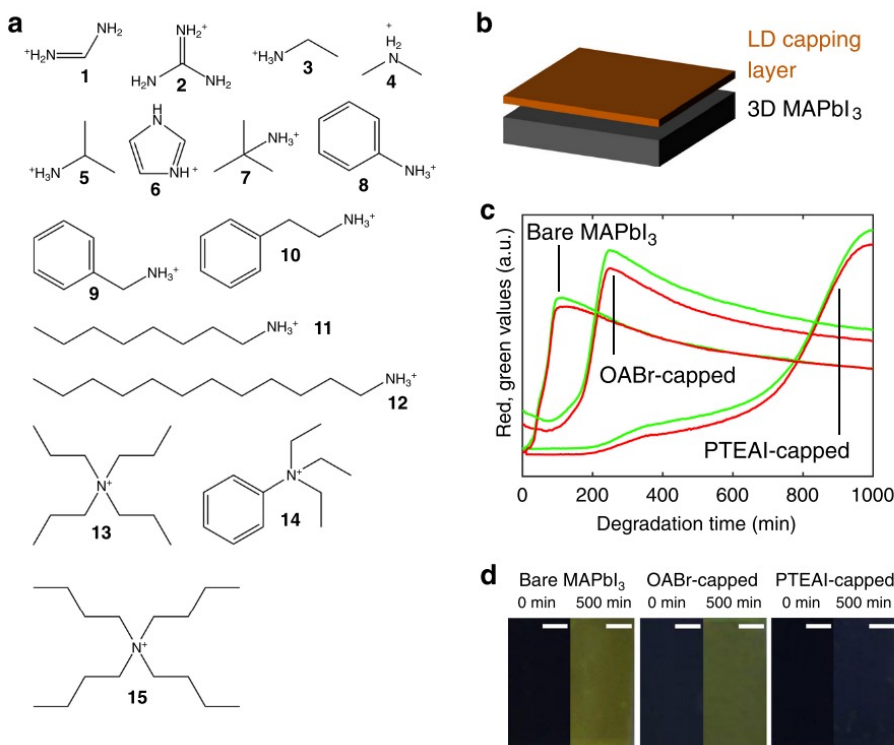


## ARTICLE

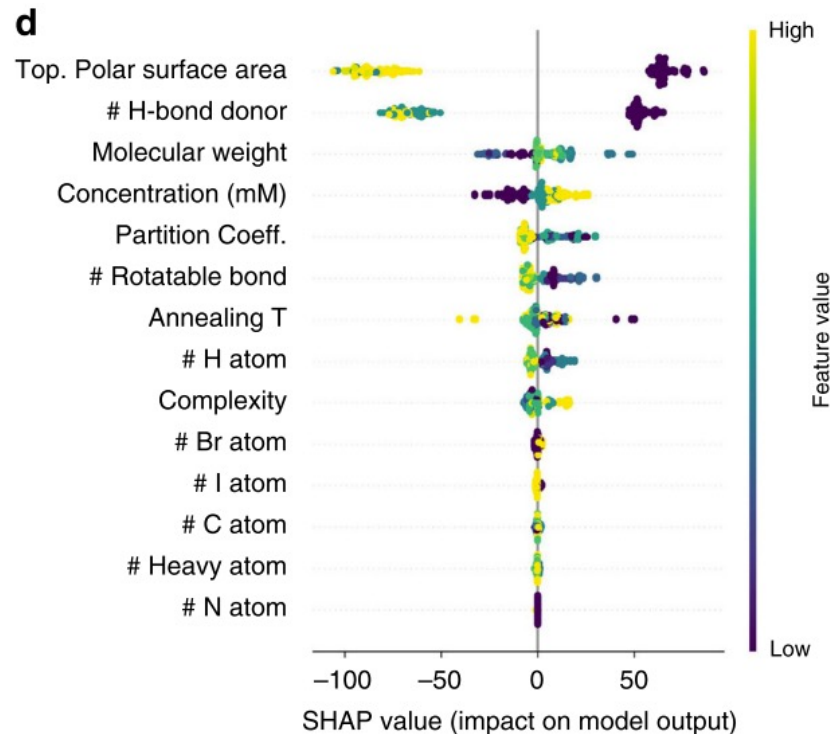


<https://doi.org/10.1038/s41467-020-17945-4> OPEN

How machine learning can help select capping layers to suppress perovskite degradation



## SHAP值分析：单点影响



### 模型剖析：打破机器学习的“黑箱”特性

- SHAP分析是一个有效工具（除此之外还有LIME<sup>1</sup>等）
- 剖析机器学习模型的意义：
  - 考察模型的正确性，避免低级错误
  - 找出影响目标变量的最重要参数
  - 能够帮助我们找出变量之间的近似关系
- 注意：SHAP分析是**针对模型的**，而不是**针对数据**
  - 如果模型误差较大，则找出的变量关系也不准确

<sup>1</sup> LIME: <https://github.com/marcotcr/lime>, accessed on Oct 14, 2021.