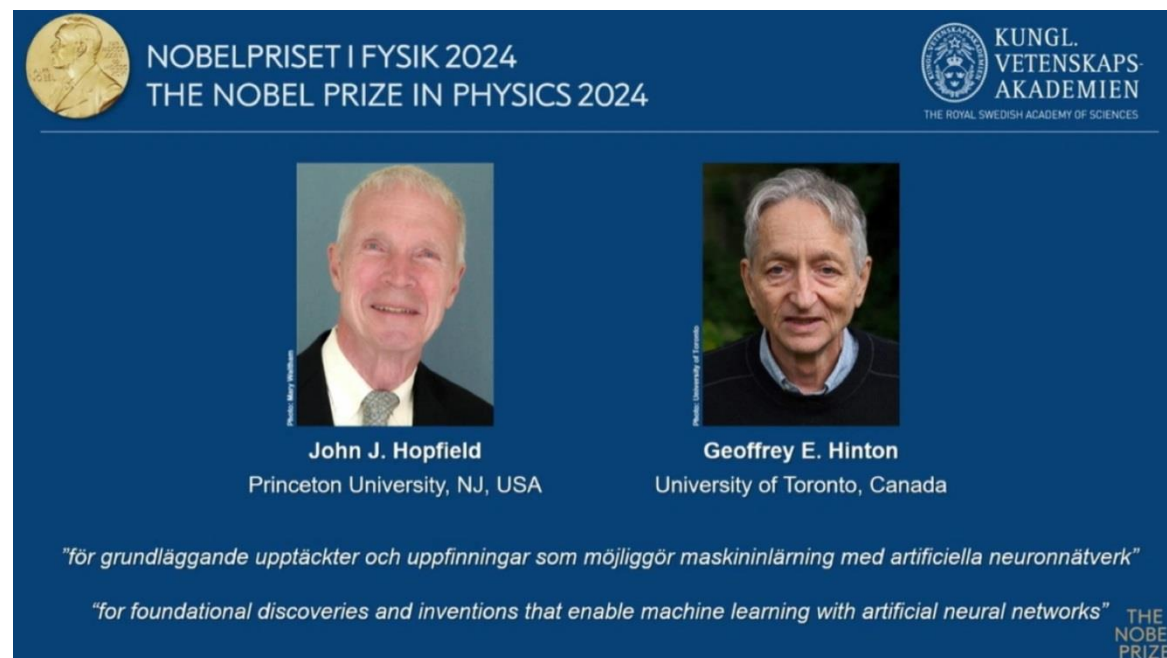


Kaggle Pro 银牌计划

机器学习基础与Kaggle平台简介

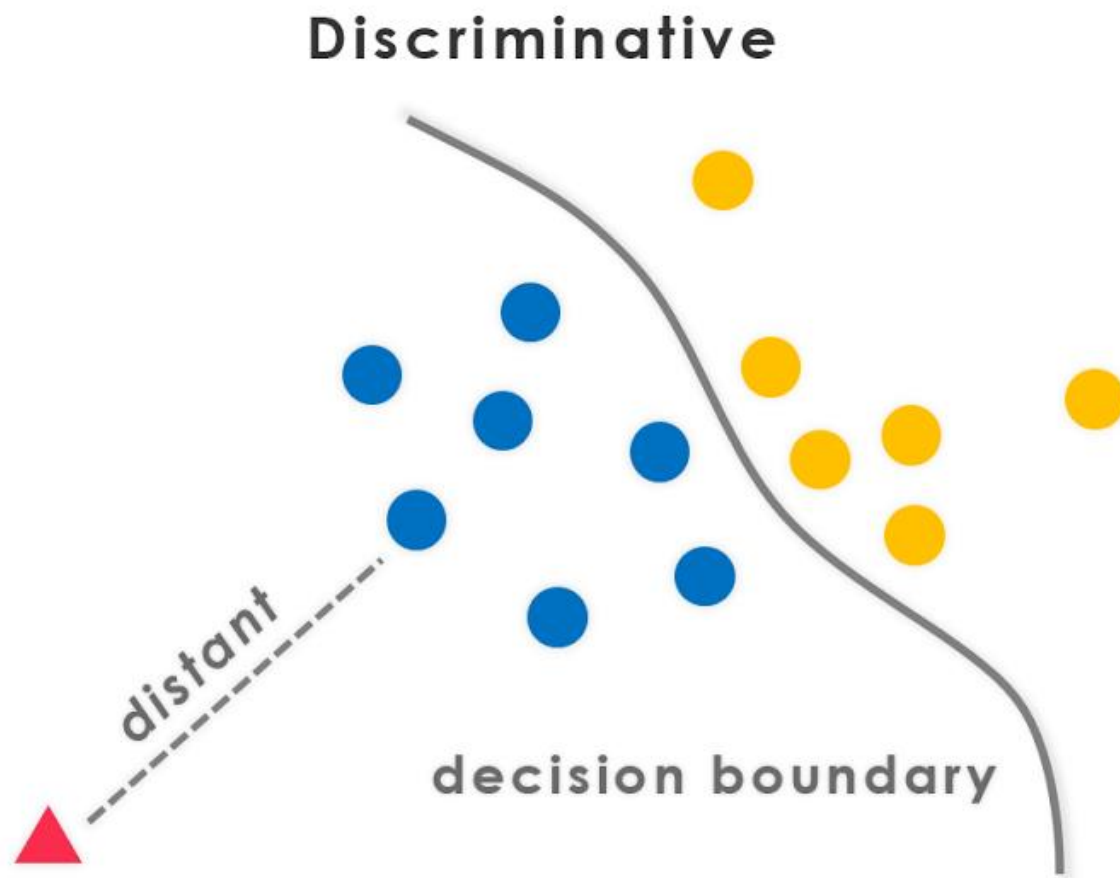
主讲人：黄老师

什么是人工智能？ 什么是机器学习？

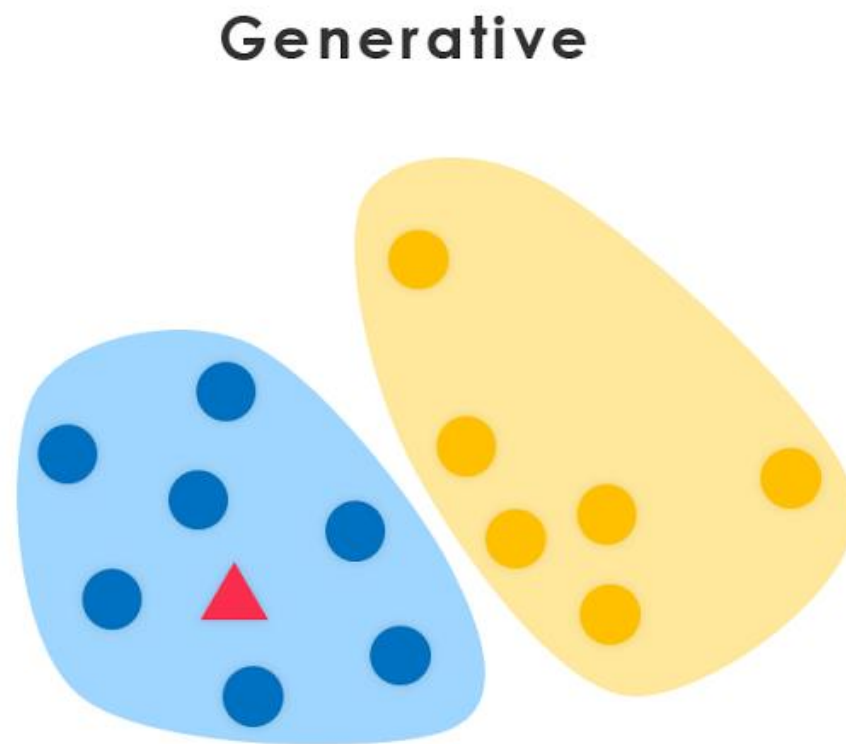


- CNN / RNN/ Transformer ...
- ChatGPT / Llama / Sora / Stable Diffusion ...

判别式AI vs 生成式AI



Before 2023



After 2023

找规律

n	1	2	3	4	5
F(n)	1	8	27	64	?

n	1	2	3	4	5
F(n)	1	3	6	10	?

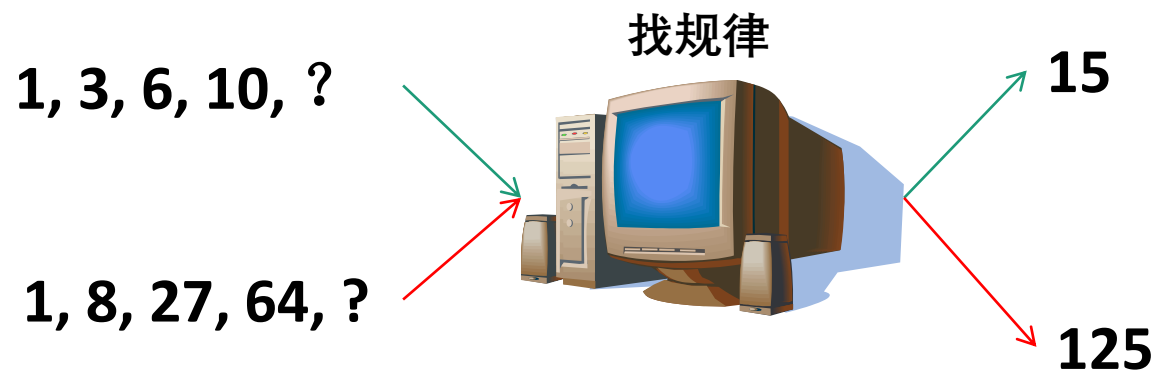
- 125

- $f(n) = n^3$

- 15

- $f(n) = f(n-1) + n$

- $f(n) = (n^2 + n) / 2$

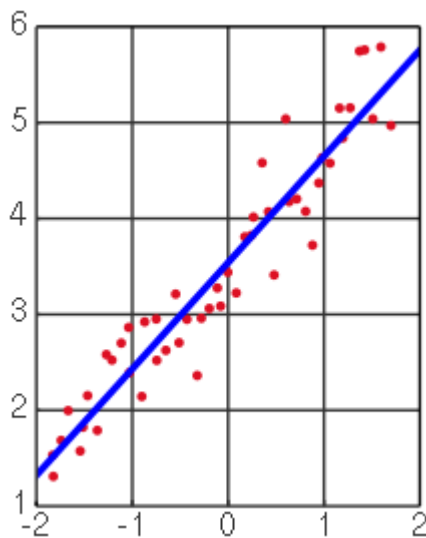


什么是机器学习?

- 根据现有数据找规律
- **基于大数据的统计学 Statistics based on Big Data**

- 维基百科：机器学习算法是一类从数据中**自动分析获得规律**，并利用规律对未知数据进行预测的算法。
- 给定数据 $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ ，机器自动学习X和Y之间的关系，**从而对新的 X_i ，能够预测 Y_i**
- 垃圾邮件识别：(邮件1, 垃圾), (邮件2, 正常), (邮件3, 垃圾), \dots (邮件N, 正常)
 - 邮件X \Rightarrow 垃圾or正常?

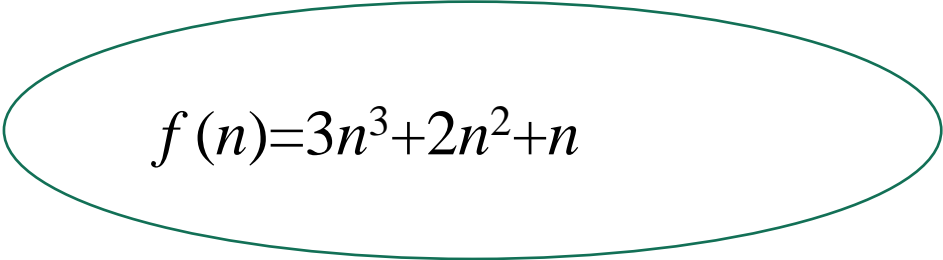
线性回归就是最简单的机器学习模型之一



为什么需要机器学习?

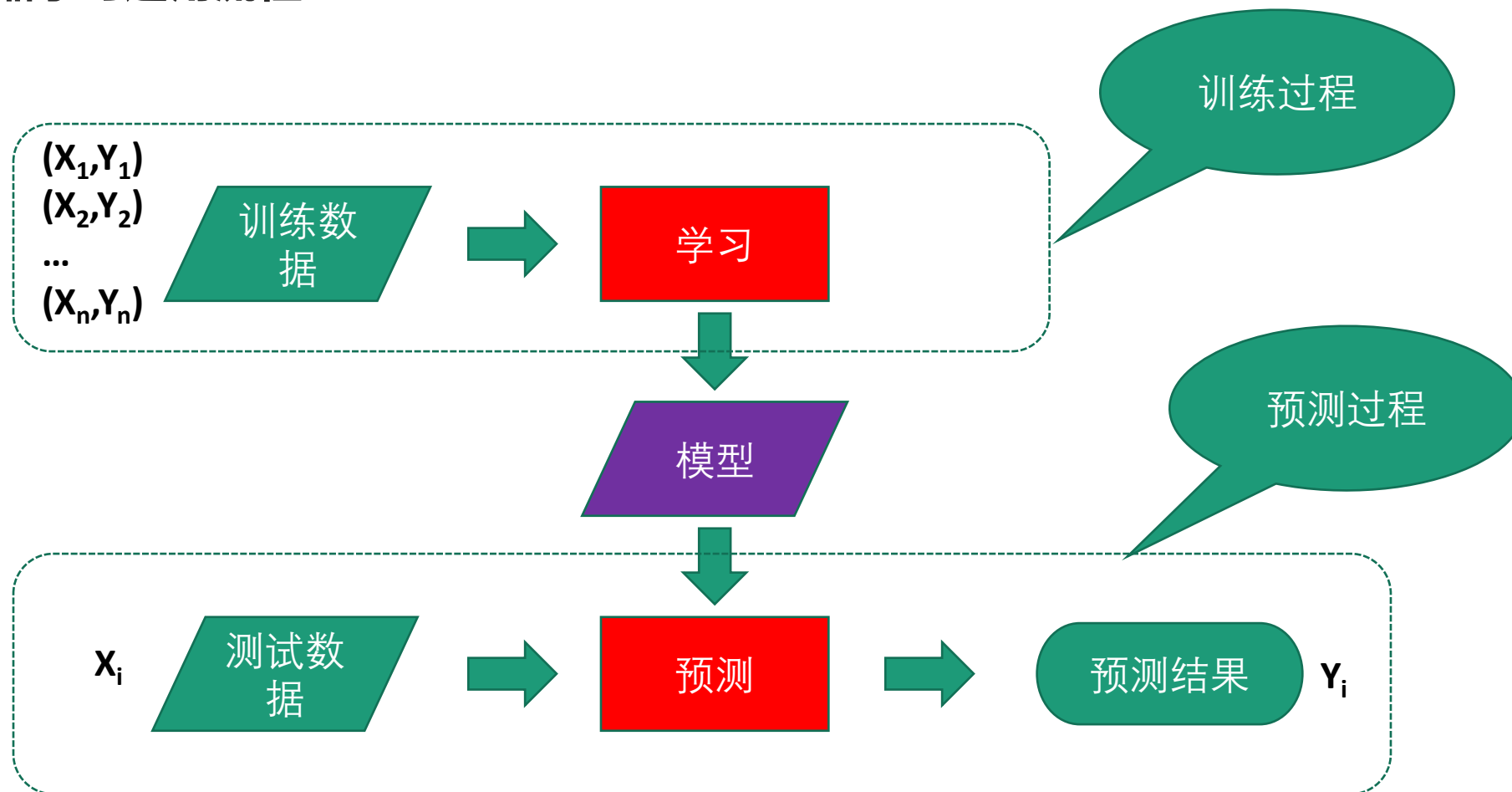
n	1	2	3	4	5
F(n)	6	34	102	228	?

• 430


$$f(n) = 3n^3 + 2n^2 + n$$

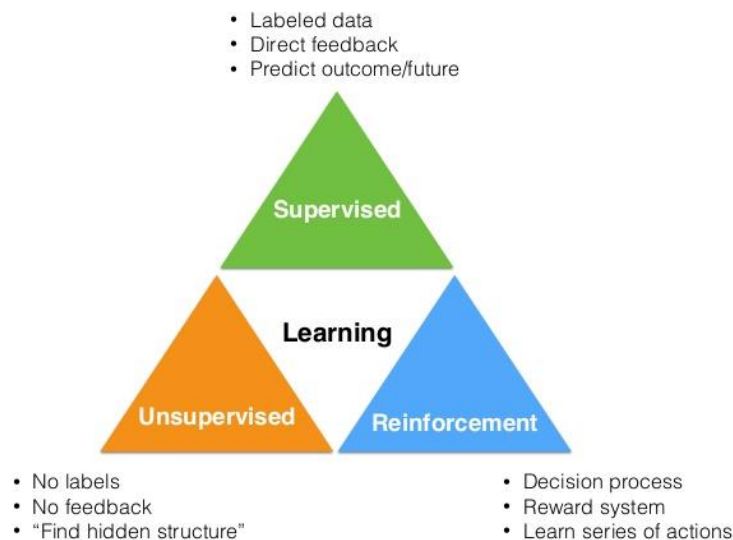
- 实际场景中的数据特征是高维的，非结构化的，或者隐式的。
- 实际任务是多样的，复杂的：图像分割、语音识别、文本实体抽取等。

机器学习通用流程



机器学习算法三大类别

- 监督学习
- 非监督学习
- 强化学习



• 监督学习 (Supervised Learning)

- 给定数据 $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$
- 对新的 X_i , 预测其 Y_i
- 分类, 回归

• 非监督学习 (Unsupervised Learning)

- 给定数据 X_1, X_2, \dots, X_n
- $f(X_i), f(X_i, X_j)$
- 降维, 聚类

核心差异在于数据**是否存在标签**

监督学习与非监督学习实例



已知标签：监督学习



未知标签或无标签：非监督学习

练习：下列机器学习任务应使用哪一类机器学习算法？

- 1. 根据学生GPA与科研经历判断是否会被大学录取**
- 2. 根据零售店的商品、位置、促销等信息预测其销售额**
- 3. 基于性格测试结果将性格相似的人聚在一起**
- 4. 根据某人的金融交易记录来决定是否对其提供贷款**
- 5. 数据特征维数太高，需要降维提升模型训练效率**
- 6. 利用卷积神经网络识别验证码图片**
- 7. 给学生作文评分**

监督学习经典模型与算法

- 线性回归
- 决策树
- 逻辑回归
- 朴素贝叶斯
- 支持向量机
- 神经网络 (MLP)

Kaggle常用模型与算法

- 数据挖掘: XGB/LGB/CAT ...
- CV: EfficientNet, ConvNeXt, U-Net, ViT, ...
- NLP: DeBERTa, RoBERTa ... / LLM: Llama, Mistral



Regression

What is the temperature going to be tomorrow?

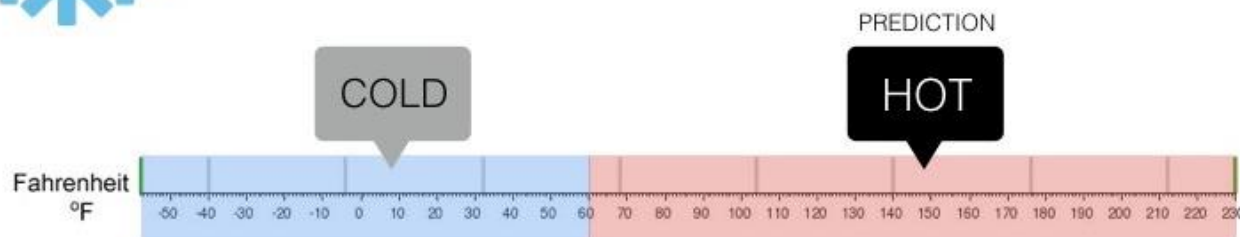


- MAE (Mean Absolute Error)
- MSE (Mean Square Error)
- ...



Classification

Will it be Cold or Hot tomorrow?



- 混淆矩阵
- 准确度、精度、召回
- F1-Score
- ...

回归问题的常用性能度量

- MAE(Mean Absolute Error)
平均绝对误差

$$MAE = \frac{1}{n} \sum_{i=1}^n |f_i - y_i|$$

- MSE(Mean Square Error)
均方误差

$$MSE = \frac{1}{n} \sum_{i=1}^n (f_i - y_i)^2$$

- RMSE(Root Mean Square Error) 均
方根误差

$$RMSE = \sqrt{MSE}$$

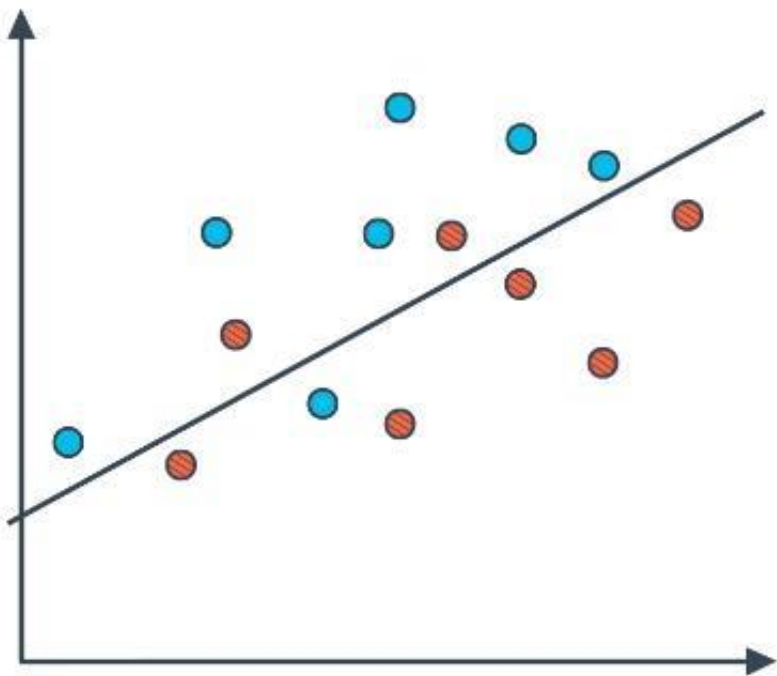
- R平方

$$r^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum (y_i - f_i)^2}{\sum (y_i - \bar{y})^2}$$

分类问题-模型评估参数：混淆矩阵

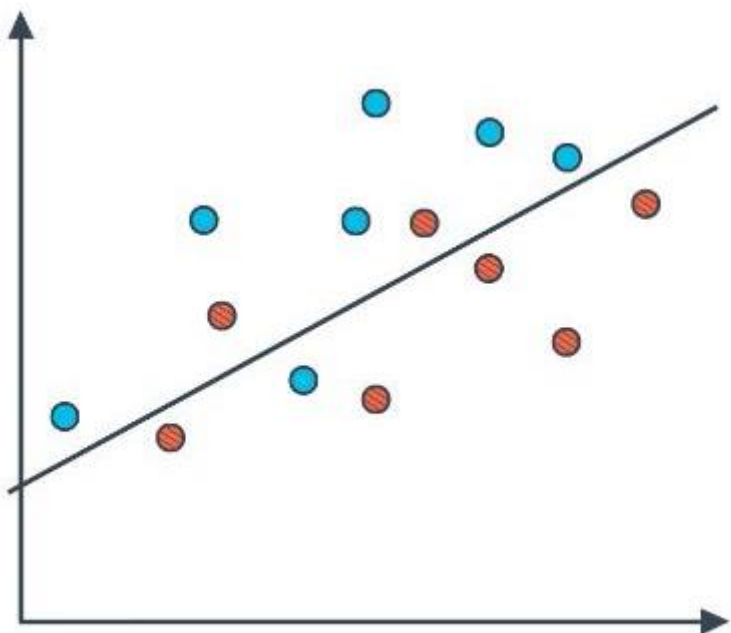
CONFUSION MATRIX

● Positive
● Negative



	Guessed Positive	Guessed Negative
Positive	True Positives	False Negatives
Negative	False Positives	True Negatives

分类问题-模型评估参数：准确度 (Accuracy)



$$Accuracy = \frac{6 + 5}{14} = 78.6\%$$

准确率的局限性

思考：如果我们的目的是检测一个人是否患有某种罕见病，那么准确率高的模型是否一定是有效的模型？

假设这种罕见病平均10000人中有1例

准确度的局限性



邮件数据集：



全部预测为正常邮件，准确度为 $Accuracy = \frac{284335}{284335+472} = 99.8\%$

该高准确度模型没有起到过滤垃圾邮件的目的


精度 (Precision)

Emails	Folder		
		SENT TO SPAM	SENT TO INBOX
	SPAM	100	170
	NOT SPAM	30 	700

被诊断为垃圾的邮件中，有多少是真正的垃圾邮件？

$$\text{Precision} = \frac{100}{100+30} = 76.9\%$$

召回率 (Recall)

		Folder	
E-mail		SENT TO SPAM	SENT TO INBOX
	SPAM	100	170
	NOT SPAM	30 	700

真正的垃圾邮件中，有多少被检测出来了？

$$\text{Recall} = \frac{100}{100 + 170} = 37.0\%$$

F1 Score



全部诊断为非垃圾邮件

精度 $\text{Precision} = \frac{0}{0} \rightarrow 100\%$

召回率 $\text{Recall} = \frac{0}{472} = 0\%$

全部诊断为垃圾邮件

精度 $\text{Precision} = \frac{472}{284335 + 472} = 0.17\%$

召回率 $\text{Recall} = \frac{472}{472} = 100\%$

调和平均

$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

F- β Score: 使模型更倾向于Precision或Recall

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

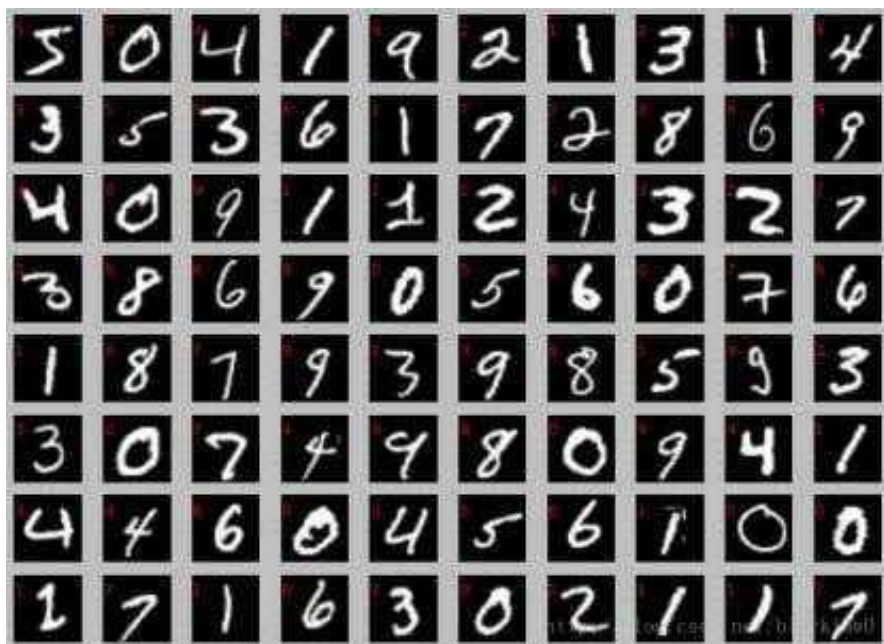
$$\text{F1 Score} = 2 \cdot \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

练习: $\beta = 0.5, 1, 2$

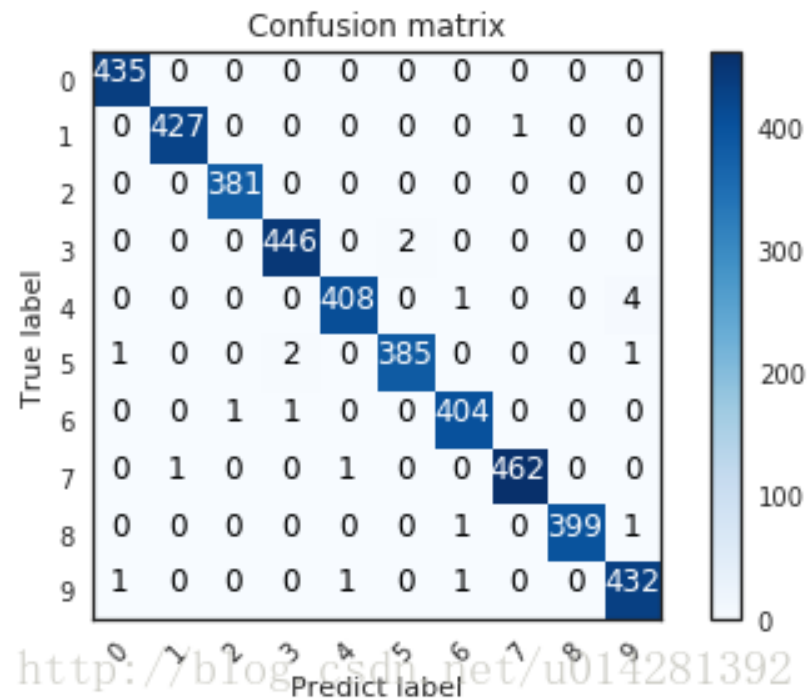
- | | |
|------------------------|-----|
| ● 检测宇宙飞船的故障零件 | 2 |
| ● 向潜在客户寄送免费样品 (样品有成本) | 0.5 |
| ● 向手机中发送关于用户可能喜欢的视频的通知 | 1 |

多分类问题：混淆矩阵


参考：<https://www.tinymind.cn/articles/591>



经典案例：MNIST手写数字识别



1. 计算模型的Accuracy
2. 如果选择“9”为模型分类的正例，计算其Precision 与 Recall

 Featured Prediction Competition

Human Protein Atlas Image Classification

Classify subcellular protein patterns in human cells

Human Protein Atlas · 2,160 teams · 4 years ago

\$37,000
Prize Money

[Overview](#) [Data](#) [Code](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Late Submission](#) [...](#)

Overview

Description

Evaluation

Prizes

Timeline

Special Prize Instructions


Submissions will be evaluated based on their **macro F1 score**.

Submission File

For each Id in the test set, you must predict a class for the Target variable as described in [the data page](#). Note that multiple labels can be predicted for each sample.

The file should contain a header and have the following format:

```
Id,Predicted
00008af0-bad0-11e8-b2b8-ac1f6b6435d0,0 1
0000a892-bacf-11e8-b2b8-ac1f6b6435d0,2 3
0006faa6-bac7-11e8-b2b7-ac1f6b6435d0,0
0008baca-bad7-11e8-b2b9-ac1f6b6435d0,0
000cce7e-bad4-11e8-b2b8-ac1f6b6435d0,0
00109f6a-bac8-11e8-b2b7-ac1f6b6435d0,1 28
...
```

 Research Code Competition

iMet Collection 2019 - FGVC6

Recognize artwork attributes from The Metropolitan Museum of Art

FGVC6 Fine-Grained Visual Categorization · 521 teams · 3 years ago

[Overview](#) [Data](#) [Code](#) [Discussion](#) [Leaderboard](#) [Rules](#) [Late Submission](#) [...](#)

Overview

Description

Evaluation

Timeline

CVPR 2019

Kernels Requirements

Submissions will be evaluated based on their **mean F2 score**. The F score, commonly used in information retrieval, measures accuracy using the precision p and recall r. Precision is the ratio of true positives (tp) to all predicted positives (tp + fp). Recall is the ratio of true positives to all actual positives (tp + fn). The F2 score is given by:

$$\frac{(1 + \beta^2)pr}{\beta^2 p + r} \text{ where } p = \frac{tp}{tp + fp}, r = \frac{tp}{tp + fn}, \beta = 2.$$

Note that the F2 score weights recall higher than precision. The mean F2 score is formed by averaging the individual F2 scores for each id in the test set.

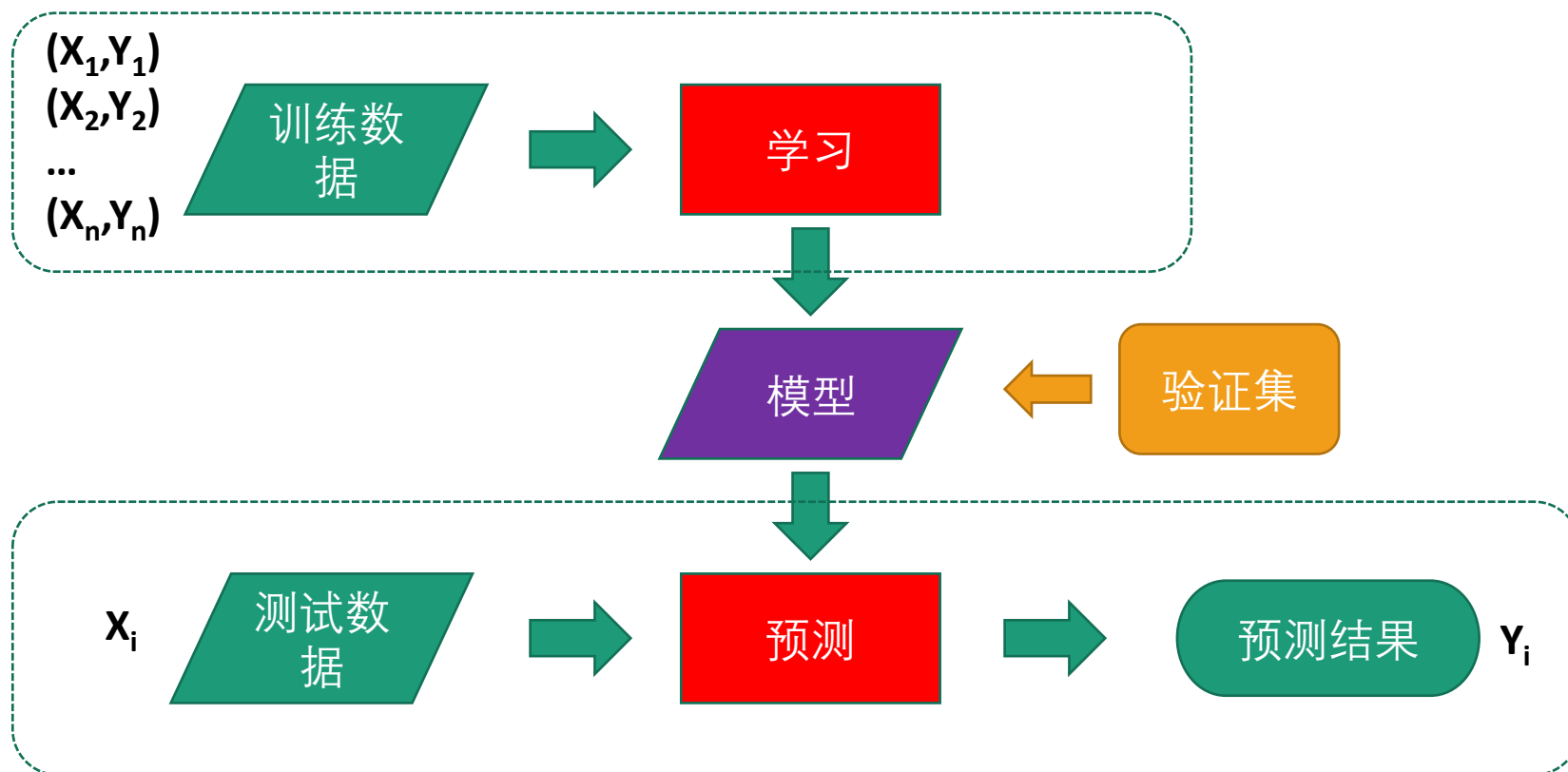
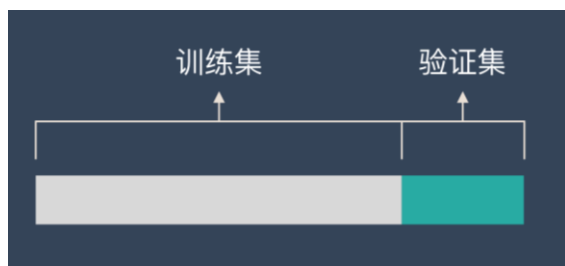
Submission File

For each image in the test set, predict a space-delimited list of tags which you believe are associated with the image. The file should contain a header and have the following format:

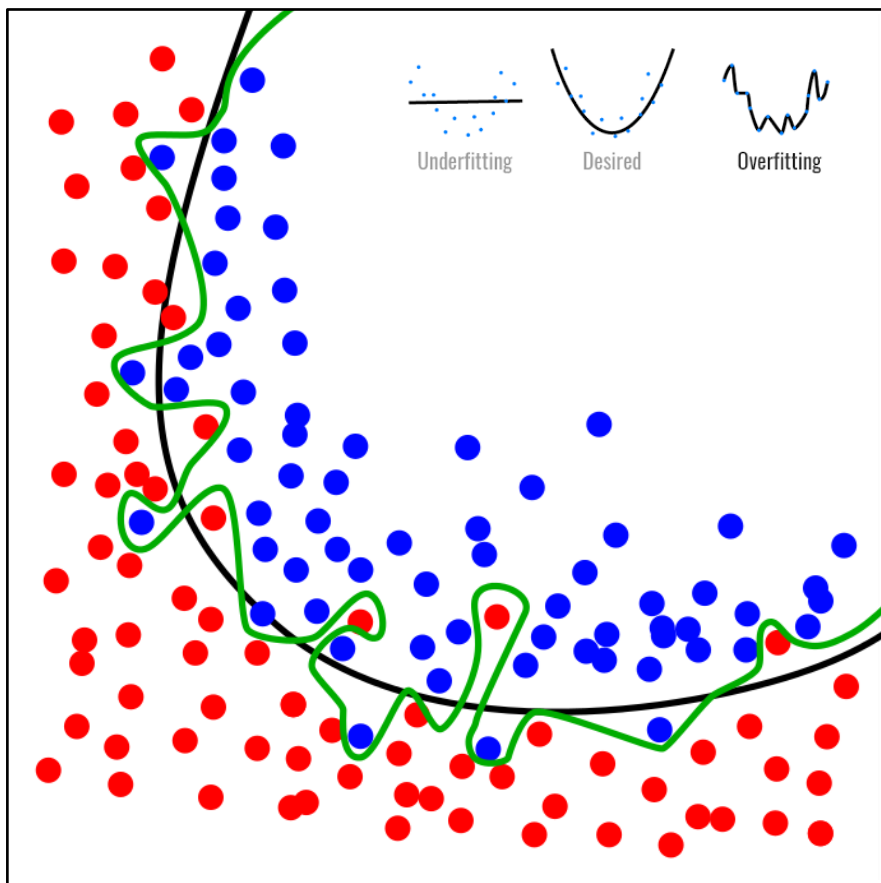
```
id,attribute_ids
10023b2cc4ed5f68,0 1 2
100fbe75ed8fd887,0 1 2
101b627524a04f19,0 1 2
etc...
```

机器模型数据集类别

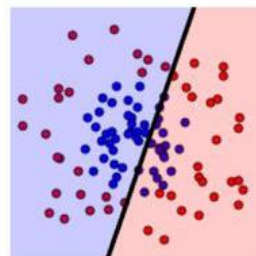
- 训练集 (Training Set)
- 验证集 (Validation Set)
- 测试集 (Testing Set)



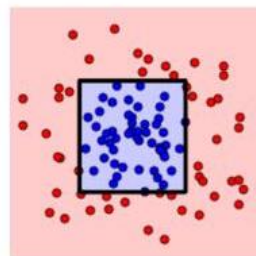
欠拟合与过拟合



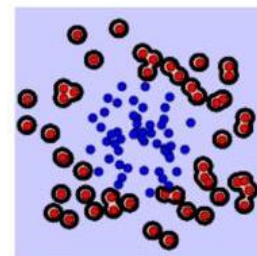
现在我们应该检测在实际模型中是否这样。当我们绘制每个模型的界限曲线时，结果如下所示：



Logistic Regression
(Underfitting)




Decision Tree
Just Right



Support Vector Machine
(Overfitting)

Kaggle: 全球最权威、认可度最高的数据科学竞赛平台

🕒 Active Competitions



Open Problems – Single-Cell Perturbations


Predict how small molecules change gene...

Featured

730 Teams

\$100,000

1mo to go



Stanford Ribonanza RNA Folding


Create a model that predicts the structure...

Research

450 Teams

\$100,000

1mo to go



Optiver - Trading at the Close


Predict US stocks closing movements

Featured · Code Competition

2209 Teams

\$100,000

2mo to go




NFL Big Data Bowl 2024

Help evaluate tackling tactics and strategy

Analytics

\$100,000

2mo to go



Linking Writing Processes to Writing Quality


Use typing behavior to predict essay quali...

Featured · Code Competition

722 Teams

\$55,000

2mo to go



AI Village Capture the Flag @ DEFCON31


Collect flags by evading, poisoning, steal...

Featured

1155 Teams

\$50,000

13d to go



Google - Fast or Slow? Predict AI Model Runtime

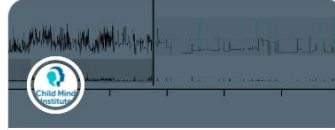
Predict how fast an AI model runs

Research

550 Teams

\$50,000

21d to go



Child Mind Institute - Detect Sleep States

Detect sleep onset and wake from wrist...

Featured · Code Competition

1133 Teams

\$50,000

1mo to go

竞赛类型（官方分类）：

- **Featured: Prize and Medal**
- **Research: Prize and Medal**
- **Playground: Kudos**
- **Knowledge: Nothing**
- **Analytics: Prize**

竞赛类型（内容分类）：

- **数据挖掘**：表格数据/时序数据
- **计算机视觉**：图像分类/目标检测/图像分割
- **自然语言处理**：文本评分/实体识别

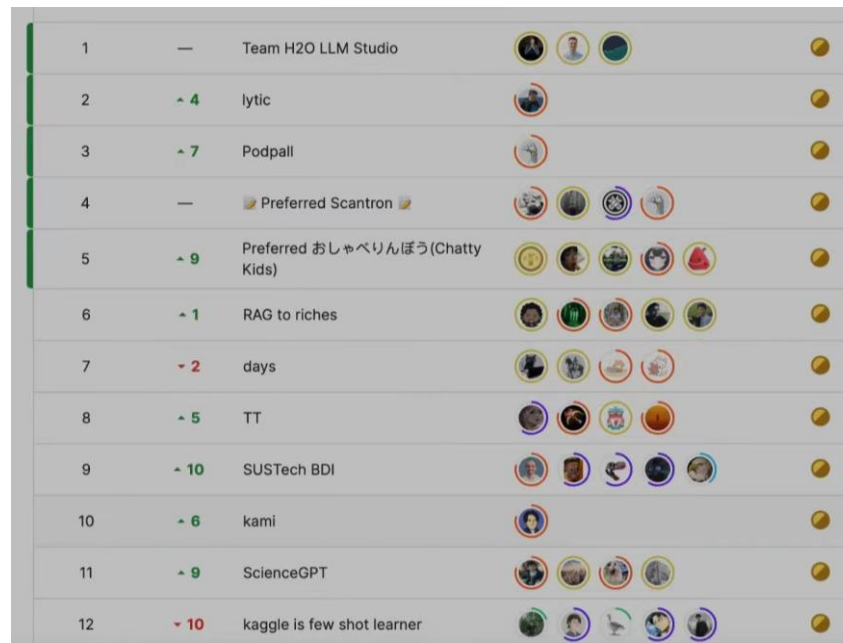
- **数据挖掘**：表格数据/时序数据
- **计算机视觉**：图像分类/目标检测/图像分割
- **自然语言处理**：文本评分/实体识别

数据挖掘比赛特点：

- 机器学习模型简单、硬件要求低；
- 预处理技巧与特征工程要求高；
- 奖牌运气成分比CV竞赛高（容易shake）

CV/NLP 比赛特点：

- **计算资源消耗大**：没有A100或者其他40G显存以上的 GPU一般拿不到银牌；
- 前沿知识要求高；
- shake的幅度一般较小



1	—	Team H2O LLM Studio		
2	~ 4	lytic		
3	~ 7	Podpall		
4	—	Preferred Scantron		
5	~ 9	Preferred おしゃべりんぼう (Chatty Kids)		
6	~ 1	RAG to riches		
7	~ 2	days		
8	~ 5	TT		
9	~ 10	SUSTech BDI		
10	~ 6	kami		
11	~ 9	ScienceGPT		
12	~ 10	kaggle is few shot learner		

Kaggle LLM 竞赛金牌

带的实习生拿了金牌，全参数微调了 70B 大模型，32 张 80GB A100 火力全开。🥳🥳🥳

如果说 6 年前自己拿块金牌很高兴的话，工作后时间不多 + 已取得 GM 称号就佛了，现在更高兴的是在我这无敌超强的实习小伙也能拿金牌了，而且是第一次参加 kaggle，一击必杀。🥳🥳🥳

奖牌机制



Competition Medals

Competition medals are awarded for top competition results. The number of medals awarded per competition varies depending on the size of the competition. Note that Community, Playground, and Getting Started competitions typically do not award medals.

	0-99 Teams	100-249 Teams	250-999 Teams	1000+ Teams
🥉 Bronze	Top 40%	Top 40%	Top 100	Top 10%
🥈 Silver	Top 20%	Top 20%	Top 50	Top 5%
🥇 Gold	Top 10%	Top 10	Top 10 + 0.2%*	Top 10 + 0.2%*

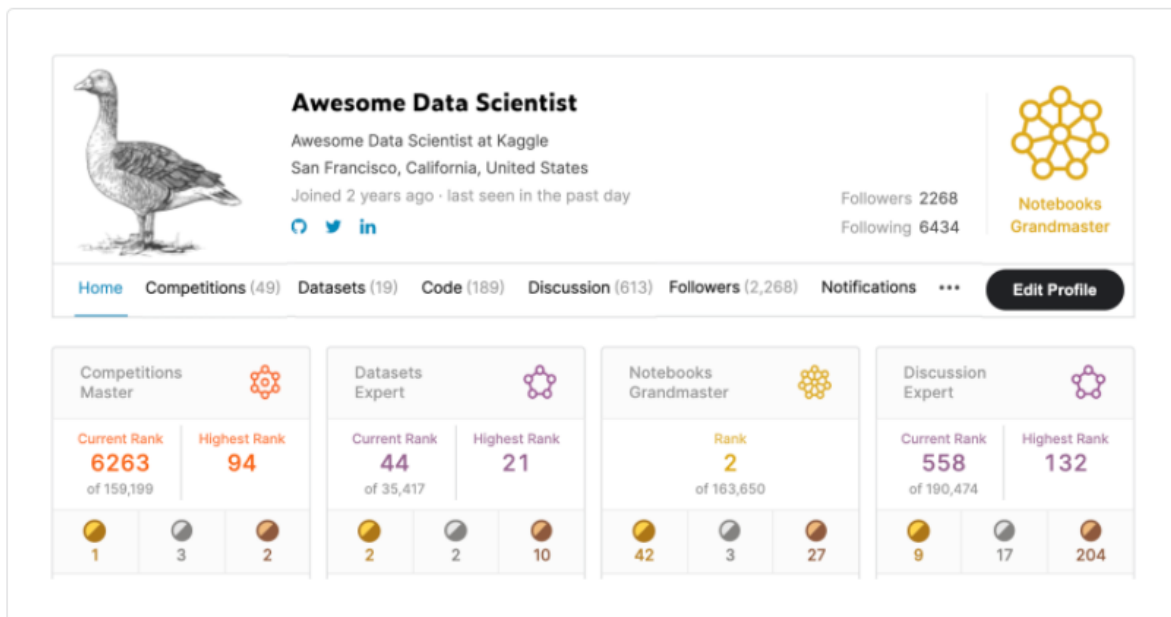
* (Top 10 + 0.2%) means that an extra gold medal will be awarded for every 500 additional teams in the competition. For example, a competition with 500 teams will award gold medals to the top 11 teams and a competition with 5000 teams will award gold medals to the top 20 teams.

身份系统

Performance Tiers

Within each category of expertise, there are five performance tiers that can be achieved in accordance with the quality and quantity of work you produce: **Novice**, **Contributor**, **Expert**, **Master**, and **Grandmaster**.

For example, you could be a Competitions Master, a Datasets Expert, a Notebooks Grandmaster, and a Discussion Expert:



The screenshot shows a Kaggle user profile for 'Awesome Data Scientist'. The profile includes a bio, location (San Francisco, California, United States), and social media links. Below the bio, there are four performance tier cards for different categories: Competitions, Datasets, Notebooks, and Discussion. Each card displays the user's current rank, highest rank, and a set of medals.

Category	Performance Tier	Current Rank	Highest Rank	Medals
Competitions	Master	6263 of 159,199	94	1 Gold, 3 Silver, 2 Bronze
Datasets	Expert	44 of 35,417	21	2 Gold, 2 Silver, 10 Bronze
Notebooks	Grandmaster	Rank 2 of 163,650		42 Gold, 3 Silver, 27 Bronze
Discussion	Expert	558 of 190,474	132	9 Gold, 17 Silver, 204 Bronze

- 两块奖牌：Expert
- 一金两银：Master
- 五金：Grandmaster（五金中有一块solo金牌）

CV/NLP 竞赛

- **PyTorch**
- **transformers**
- **timm**

```
MINGW64:/  
Think@DESKTOP-AQDP069 MINGW64 /  
$ pip install tensorflow==1.13.1
```

`pip install -i https://pypi.tuna.tsinghua.edu.cn/simple lightgbm`

Pytorch: <https://pytorch.org/>

数据挖掘竞赛

- **LightGBM**
- **XGBoost**
- **CatBoost**
- **PyTorch**

PyTorch Build	Stable (1.11.0)	Preview (Nightly)	LTS (1.8.2)	
Your OS	Linux	Mac	Windows	
Package	Conda	Pip	LibTorch	Source
Language	Python		C++ / Java	
Compute Platform	CUDA 10.2	CUDA 11.3	ROCm 4.5.2 (beta)	CPU
Run this Command:	<code>pip3 install torch torchvision torchaudio --extra-index-url https://download.pytorch.org/whl/cu113</code>			

1. 安装Anaconda3 (推荐**3.10版本**)
2. 配置环境变量
3. 安装PyTorch
4. 安装CPU版本LightGBM、XGBoost和CatBoost

遇到问题群里讨论并@助教和Mentor

<https://repo.anaconda.com/archive/>

Index of /

Filename	Size	Last Modified
.winzip/	-	
Anaconda3-2023.03-1-Windows-x86_64.exe	786.6M	2023-04-24 12:41:07
Anaconda3-2023.03-1-MacOSX-x86_64.sh	601.6M	2023-04-24 12:41:07
Anaconda3-2023.03-1-MacOSX-x86_64.pkg	600.1M	2023-04-24 12:41:06
Anaconda3-2023.03-1-MacOSX-arm64.sh	566.0M	2023-04-24 12:41:06
Anaconda3-2023.03-1-MacOSX-arm64.pkg	564.4M	2023-04-24 12:41:06
Anaconda3-2023.03-1-Linux-x86_64.sh	860.6M	2023-04-24 12:41:05
Anaconda3-2023.03-1-Linux-s390x.sh	361.2M	2023-04-24 12:41:05
Anaconda3-2023.03-1-Linux-ppc64le.sh	435.1M	2023-04-24 12:41:05
Anaconda3-2023.03-1-Linux-aarch64.sh	618.7M	2023-04-24 12:41:04
Anaconda3-2023.03-0-Windows-x86_64.exe	786.0M	2023-03-20 10:41:36
Anaconda3-2023.03-0-MacOSX-x86_64.sh	601.0M	2023-03-20 10:41:36
Anaconda3-2023.03-0-MacOSX-x86_64.pkg	599.7M	2023-03-20 10:41:36
Anaconda3-2023.03-0-MacOSX-arm64.sh	565.4M	2023-03-20 10:41:35
Anaconda3-2023.03-0-MacOSX-arm64.pkg	564.1M	2023-03-20 10:41:35