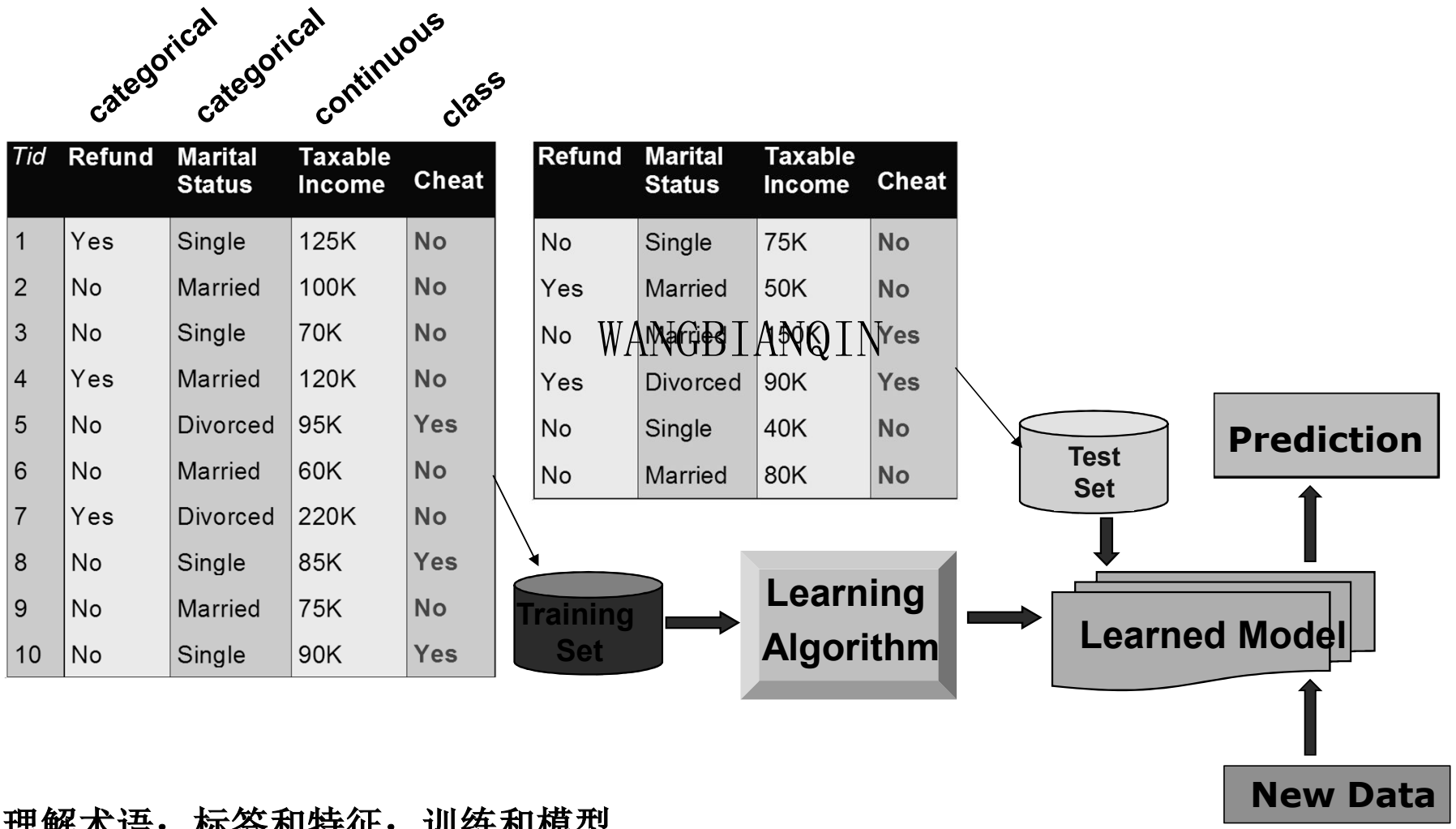
第1讲 机器学习概述

- 机器学习简介
- 机器学习流程
- 机器学习方法
- 实践：Python

WANGBIANQIN

机器学习-分类示例



理解术语：标签和特征；训练和模型

机器学习？

- 人工智能(Artificial Intelligence, AI)：研究使计算机来模拟人的某些思维过程和智能行为的科学。例如，学习、推理、思考、规划等。人工智能包括计算智能、感知智能和认知智能等层次，目前人工智能还介于前二者之间。

- 机器学习(Machine Learning, ML)：人工智能的一个分支，它是实现人工智能的一个核心技术，即以机器学习为手段解决人工智能中的问题

机器学习？

- 机器学习所研究的主要内容是关于在计算机上从数据中产生“模型（model）”的算法，即“学习算法（learning algorithm）”
- 学习算法基于经验数据（标签数据）产生模型，然后模型基于新数据给出判断

源自周志华《机器学习》

机器学习定义

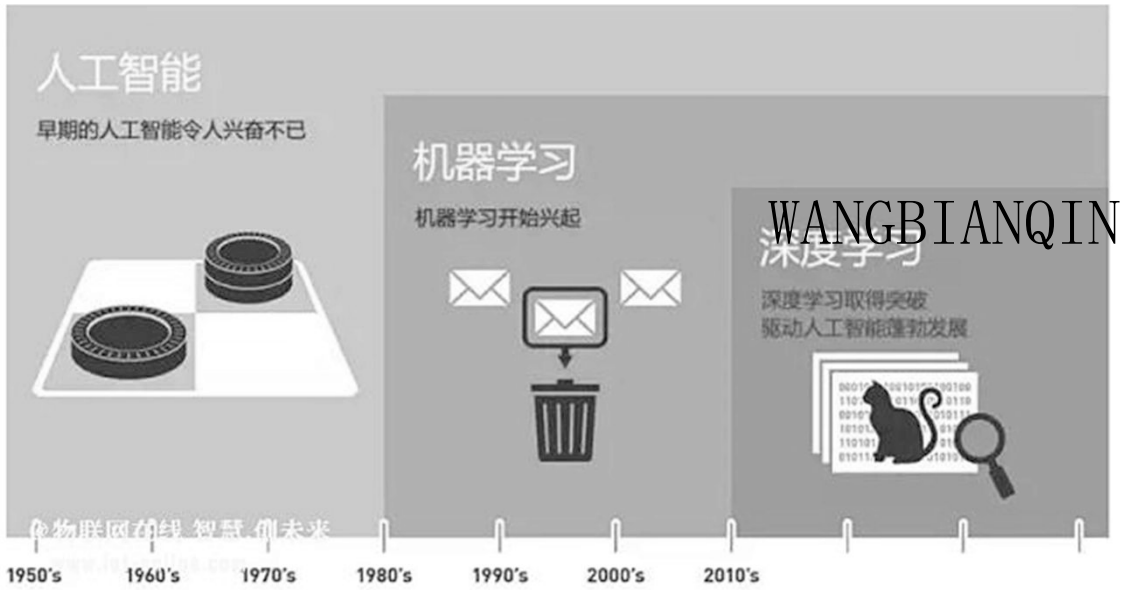
- Tom Mitchell在他的《 Machine Learning》一书的序言开场白中又给出了一个更为广泛引用的定义：“机器学习这门学科所关注的问题是计算机程序如何随着经验积累自动提高性能。”
- 形式化描述：“对于某类任务T和性能度量P，如果一个计算机程序在T上以P衡量的性能随着经验E而自我完善，则称这个计算机程序在从经验E学习。”
- 例如，一个邮箱过滤程序接受到一封邮件，如何判断此邮件是否是垃圾邮件？

  任务T： 判断邮件是不是垃圾邮件

  经验E： 观察到之前标记过是不是垃圾邮件的邮件

  性能P： 正确分类垃圾邮件与非垃圾邮件的数量

AI、ML、DL之间关系

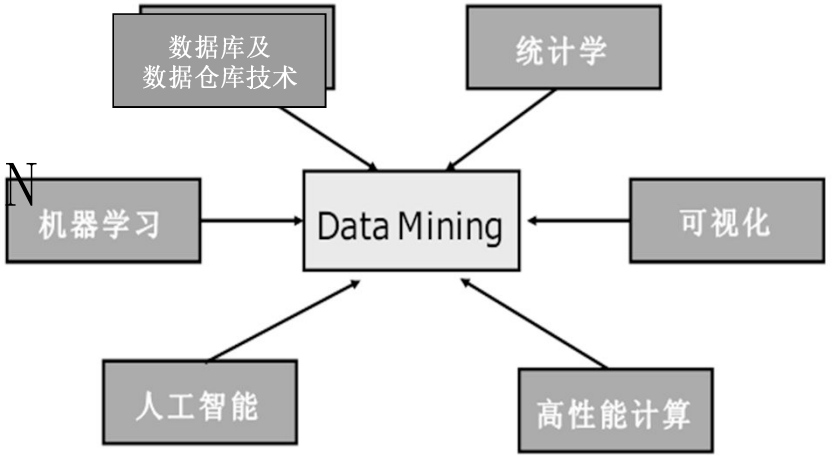


机器学习+

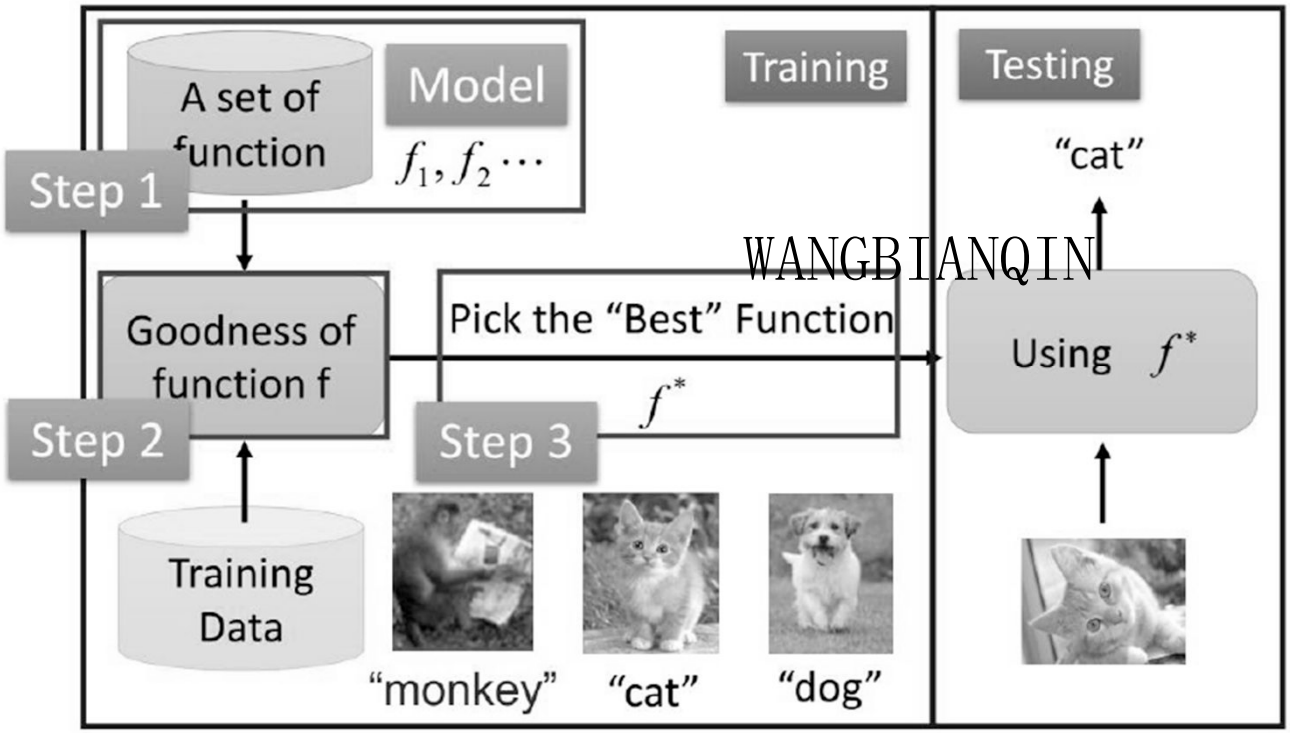

模式识别
计算机视觉
WANGBIANQIN
数据挖掘
语音识别
机器学习
统计学习
自然语言处理
Google
Translate

机器学习与数据挖掘

- 数据挖掘是一门实用学科，研究实用算法（大多是机器学习算法），利用各个领域产出的数据来解决各个领域相关问题
- 机器学习偏理论，注重研究算法有效性；数据挖掘偏应用，则以应用需求为导向，研究具体问题本身



机器学习框架


Training
Testing
A set of function
Model
$f_1, f_2 \cdots$
Step 1
"cat"
WANGBIANQIN
Goodness of function f
Pick the "Best" Function
Using f^*
f^*
Step 2
Step 3
Training Data
"monkey"
"cat"
"dog"

机器学习流程

≈ 寻找一个函数

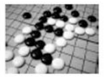
- 语音识别

  f() = “你好吗？”

- 图像识别

  f() = “猫”

- 围棋对战

  f() = “5-5” (下一步)

- 对话系统（如Siri）

  f(“你好！” (用户发问)) = “您好！” (系统回应)

- **例如，典型线性回归问题**

  包含一组数据点(x, y)：

  (0, 0), (1, 20), (2, 60), (3, 68), (4, 77), (5, 110)

  假设x与y之间服从线性分布，即函数

  $$f(x) = y = ax + b$$

  $f(x)$是假设模型

- **如何衡量模型模型预测的好坏？**

机器学习流程

- 损失函数（loss function）或代价函数（cost function）：度量模型预测的值和真实值之间的误差。

  例如，对于典型的线性回归问题，常用最小二乘法构造损失函数：

WANGBIANQIN

$$L(a,b)=\sum_{i=1}^{n}\left(y_i-(ax_i+b)\right)^2$$

- 如何使损失函数最小？

机器学习流程

- 优化求参：优化问题只是一个单纯的函数值最小化问题，靠的是如何定义要优化的函数，从优化的角度，损失函数也称目标函数(objective function)
- 通过目标函数优化求参：WANGBIANQIN

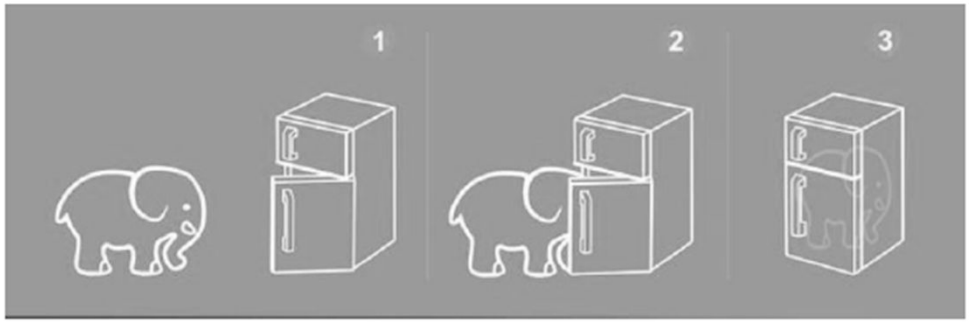
$$L(a,b)=\sum_{i=1}^{n}\left(y_i-\left(ax_i+b\right)\right)^2$$

机器学习流程

基本步骤

WANGBIANQIN

机器学习就是这么简单…



机器学习方法

- 有监督学习(supervised learning)：学习带有标签的数据，即训练数据
- 无监督学习(unsupervised learning)：学习未被标记的数据
- 半监督学习(semi-supervised learning)：数据集中包含有标签和未被标记的数据
- 强化学习(Reinforcement Learning)：强调如何基于环境而行动，以取得最大化的预期利益。例如，AlphaGo背后的机器学习方法

WANGBIANQIN

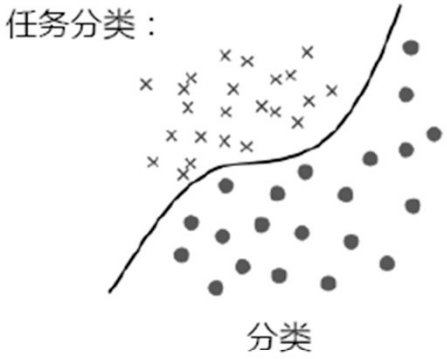
有监督算法

- 基于距离：利用相似性或者距离度量
  例如，*k*NN
- 统计算法：利用统计信息
  例如，朴素贝叶斯、逻辑回归、SVM
- 决策树类：利用树型结构
  例如，ID3，C4.5\C5.0，CART
- 组合方法：聚集多个分类器提高分类精度
  例如，随机森林、AdaBoot、提升树
- 神经网路/深度学习：利用自身的结构，例如，各种神经网络模型，DNN, CNN, RNN

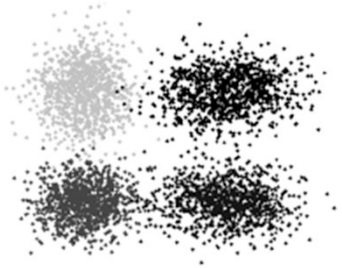
无监督聚类算法

基于划分算法 (partitioning methods)	• **典型算法：K-means、K-medoids（k中心）、CLARANS算法**
基于层次算法 (hierarchical methods)	• **代表算法：最短距离法、BIRCH、CURE、CHAMELEON算法** WANGBIANQIN
基于密度方法 (density-based methods)	• **代表算法：DBSCAN、OPTICS、DENCLUE算法**
基于网格方法 (grid-based methods)	• **代表算法：STING、CLIQUE、WAVE-CLUSTER算法**
基于模型方法 (model-based methods)	• **基于统计方案（COBWEB、CLASSIT等）和神经网络（SOM等）方案**

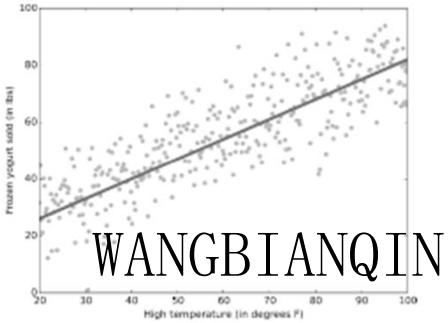
机器学习任务



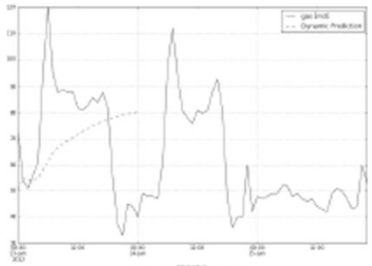
分类



聚类



回归



时序分析

- 监督学习，训练样本包含对应的“标签”，如识别问题
  - 分类问题，样本标签属于两类或多类（离散）
  - 回归问题，样本标签包括一个或多个连续变量（连续）

无监督学习：主要包括聚类，数据变换

分类Classification

- 目标：将数据映射到定义好的群组或类，类的个数是确定的、预先定义好的。
- 原理：从数据中选出已经分好类的训练集，在该训练集上建立分类模型，对未分类数据进行分类。
- 例子：
  - 信用卡申请者，分类为低、中、高风险
  - 分配客户到预先定义的客户分片

分类过程

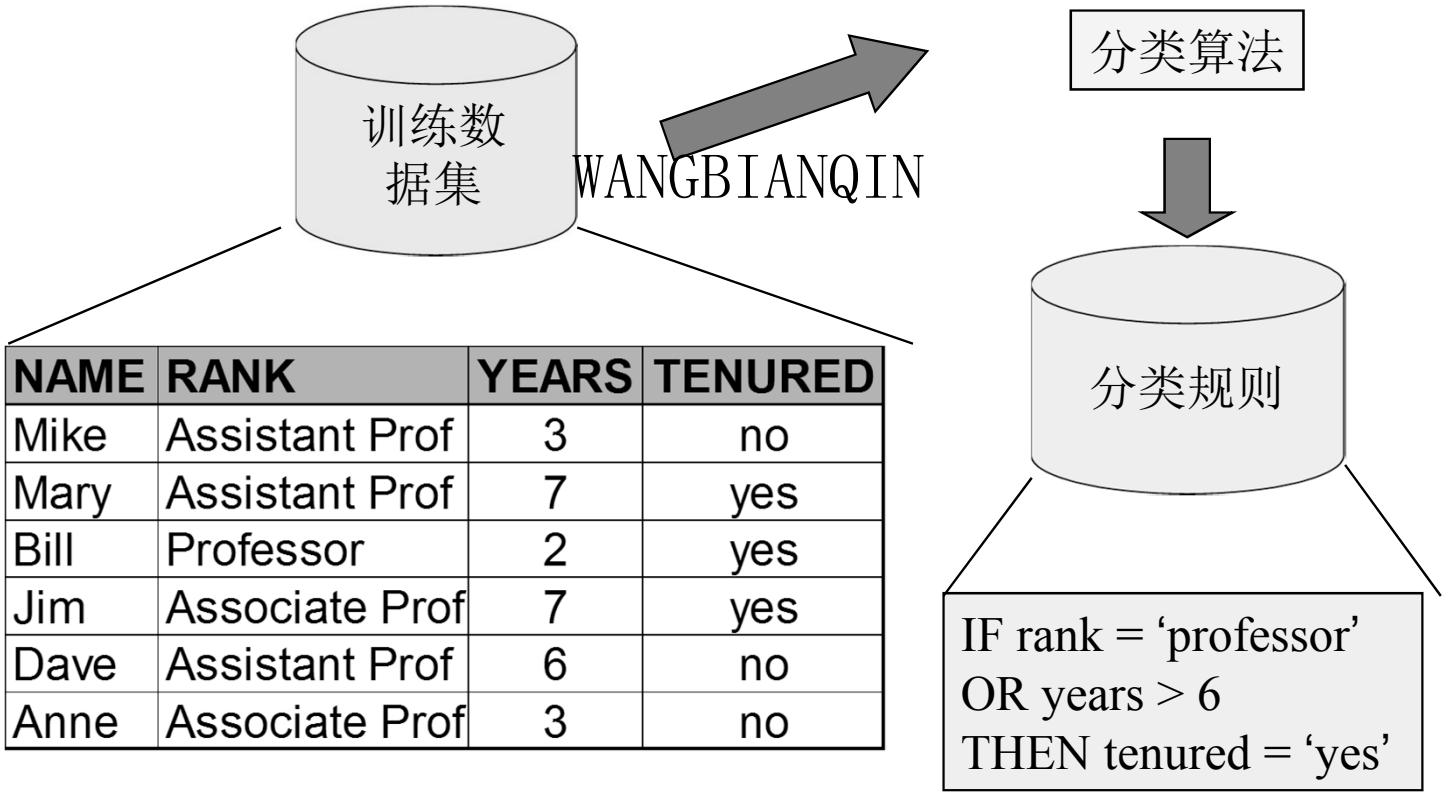
- 两个步骤：
  - 模型构建（归纳）

    通过对训练集合的归纳，建立分类模型
  - 预测应用（推论）

    根据建立的分类模型，对测试集合进行测试

学习过程

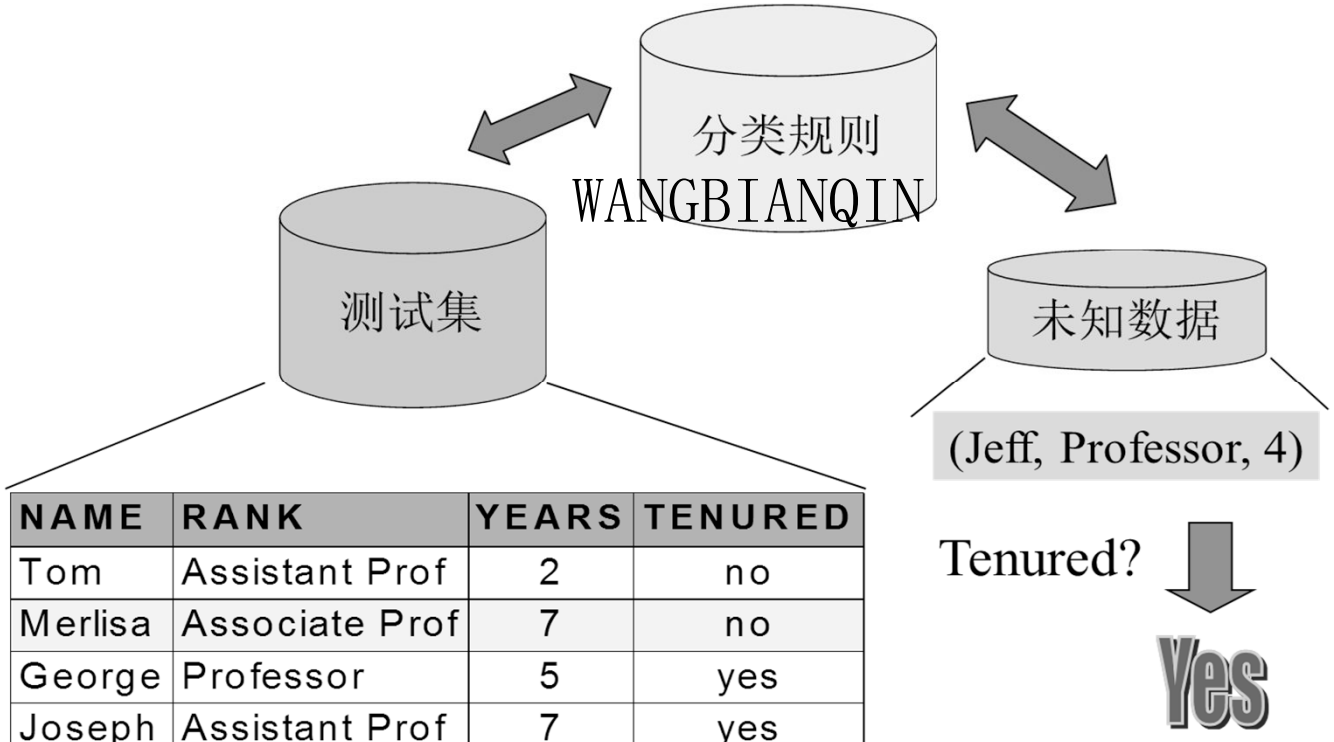
- 通过对训练集的学习，建立分类模型或分类器



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

预测过程

- 利用模型对未知数据对象进行分类



NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

回归Regression

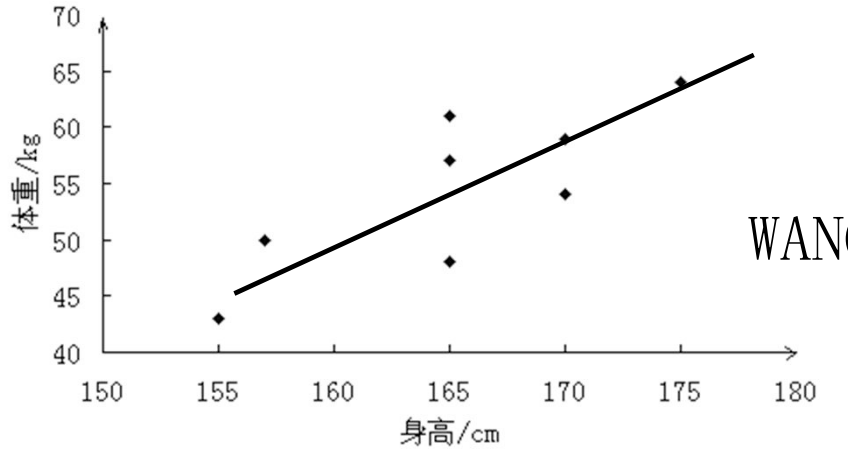
- 分类与回归（预测）区别：
  - 目标属性值是离散值：分类
  - 目标属性值是连续值：回归（预测）
- 目标：将数据项映射到一个实值预测变量

- 例子：

  根据购买模式，预测一个家庭的孩子数

  根据购买模式，预测一个家庭的收入
- 与分类区别？

回归示例



回归方程：

$$\hat{y} = 0.849x - 85.172$$

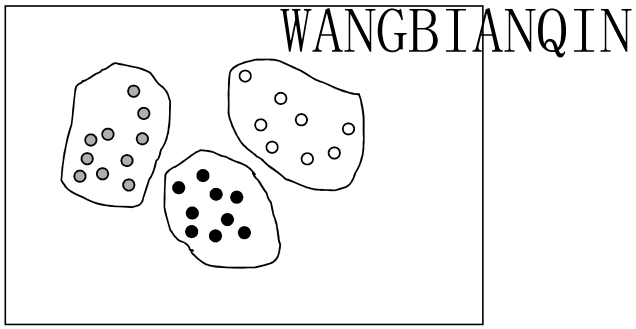
身高172cm女大学生体重

$\hat{y} = 0.849 \times 172 - 85.712 = 60.316(kg)$

结果准确吗？如果不准，你能解析一下原因吗？

聚类Clustering

- 目标：使相似的数据自然聚在同一个簇
- 与分类区别：聚类不依赖于预先定义好的类，不需要训练集
- 原理：簇内对象具有最大相似性，簇间对象具有最小相似性



- 例子：

  一些特定症状的聚类可能预示了一个特定的疾病

  不同的兴趣爱好，暗示不同的客户人群

关联分析Association analysis

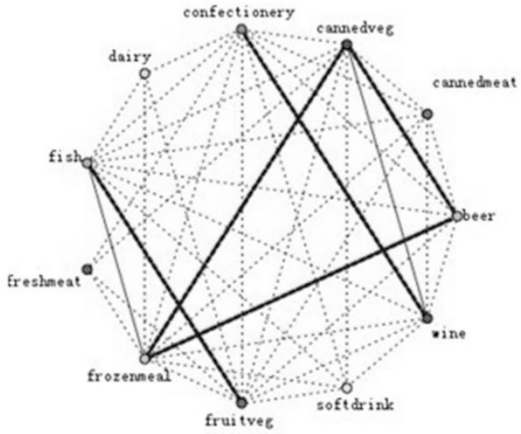
- 目标：发现关联规则
  - 两种技术：关联规则和序列模式
  - 关联规则：寻找在同一个事件中出现的不同项的相关性
  - 序列模式：寻找的是事件之间时间上的相关性
- 例子：

  客户在购买A的同时，经常会购买B，

  A => B (关联规则)

  客户在购买A后，隔段时间，会购B，

  [{A}, {B}] (序列模式)
- 经典算法：Apriori、 AprioriAll、FPtree



经典数据挖掘算法

排名	挖掘主题	算法	得票数	发表时间	作者
1	分类	C4.5	61	1993	Quinlan, J.R
2	聚类	K-Means	60	1967	MacQueen, J.B
3	统计学习	SVM	58	1995	Vapnik, V.N
4	关联分析	Apriori	52	1994	Rakesh Agrawal
5	统计学习	EM	48	2000	McLachlan, G
6	链接挖掘	PageRank	46	1998	Brin, S.
7	集装与推进	AdaBoost	45	1997	Freund, Y.
8	分类	kNN	45	1996	Hastie, T
9	分类	Naïve Bayes	45	2001	Hand, D.J
10	分类	CART	34	1984	L.Breiman

国际权威的学术组织the IEEE International Conference on Data Mining (ICDM) 2006年12月评选出

课程涉及算法

- **Supervised learning: kNN, Native Bases, Linear model(logistic Regression, softmax Regression), SVM, CART, Random forests, AdaBoost, GBDT, DNN, CNN, RNN**

WANGBIANQIN

- **Unsupervised learning : Apriori, kMeans, DBSCAN, PCA**

构建ML模型步骤

- 收集数据
- 准备输入数据
- 分析输入数据　　WANGBIANQIN
- 训练算法
- 测试算法
- 使用算法

实验环境

- 实验环境：Python
- Python：通用，脚本，开源，垮平台，多模型
- Python库：Numpy、Scipy、Matplotlib、pandas、Scikit-Learn
- Scikit-learn：内置热门机器学习算法
- TensorFlow：一个较比低级库，能创建自定义机器学习算法

Python版本选择

- Python由荷兰人Guido van Rossum于1989年发明
- 存在两个版本：Python2.x，Python3.x，二者之间存在一定差异，其中Python2的支持和维护将于2020年1月1日期停止
- 请学习Python3.x
- 思考：如何快速判断一个Python代码是Python3.x

Python环境配置

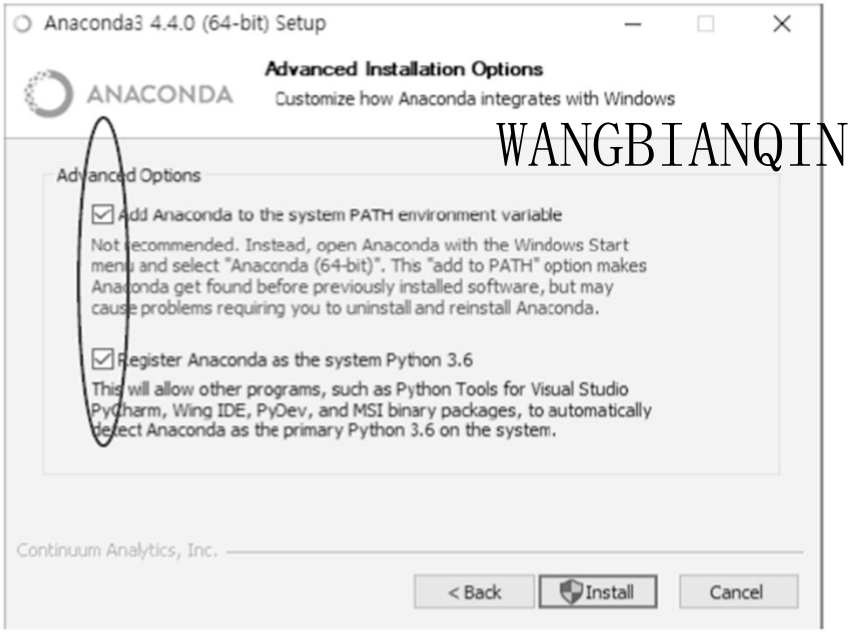
- Anaconda：专注数据分析的Python发行版，预集数据科学和机器学习中常用的软件包，包含Jupyter notebook、Numpy、Scipy、Matplotlib、 Scikit-learn等多个科学包及其依赖项

  https://www.anaconda.com/download/

安装Anaconda

安装Anaconda

确认勾选将Python添加到系统环境变量



实验环境搭建

- 安装Anaconda，下载OS相应版本
  https://www.anaconda.com/download/
  双击安装文件，全部默认选项
  WANGBIANQIN
- 安装第三方库，例如，
  安装Tensorflow（确保path中有相应的环境变量，如C:\Users\Anaconda3\Scripts）
  在命令窗口cmd中，执行 pip install tensorflow

  必要时请先运行**python -m pip install --upgrade pip**

Python IDE

- Spyder和Jupyter notebook数据处理
- Pycharm: Web应用开发

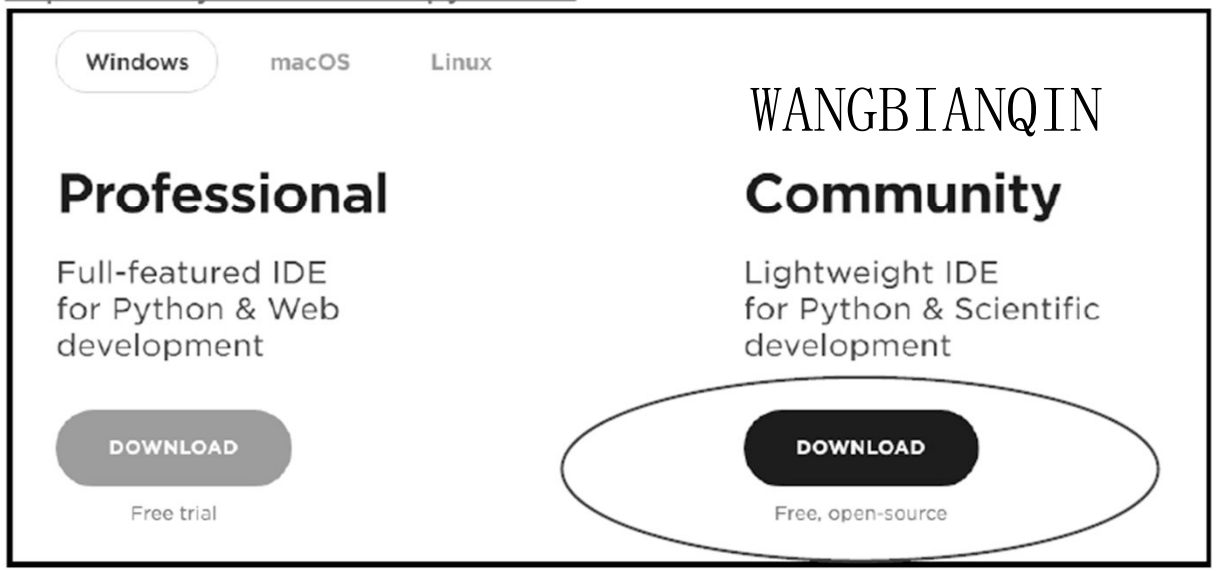
- Spyder：完全免费，适合熟悉matlab开发者
  https://github.com/spyder-ide/spyder
- Pycharm社区版，可满足不涉及Web开发，适合大多数开发者
  https://www.jetbrains.com/pycharm/download/#section=windows

Pycharm安装

PyCharm配置

- 下载

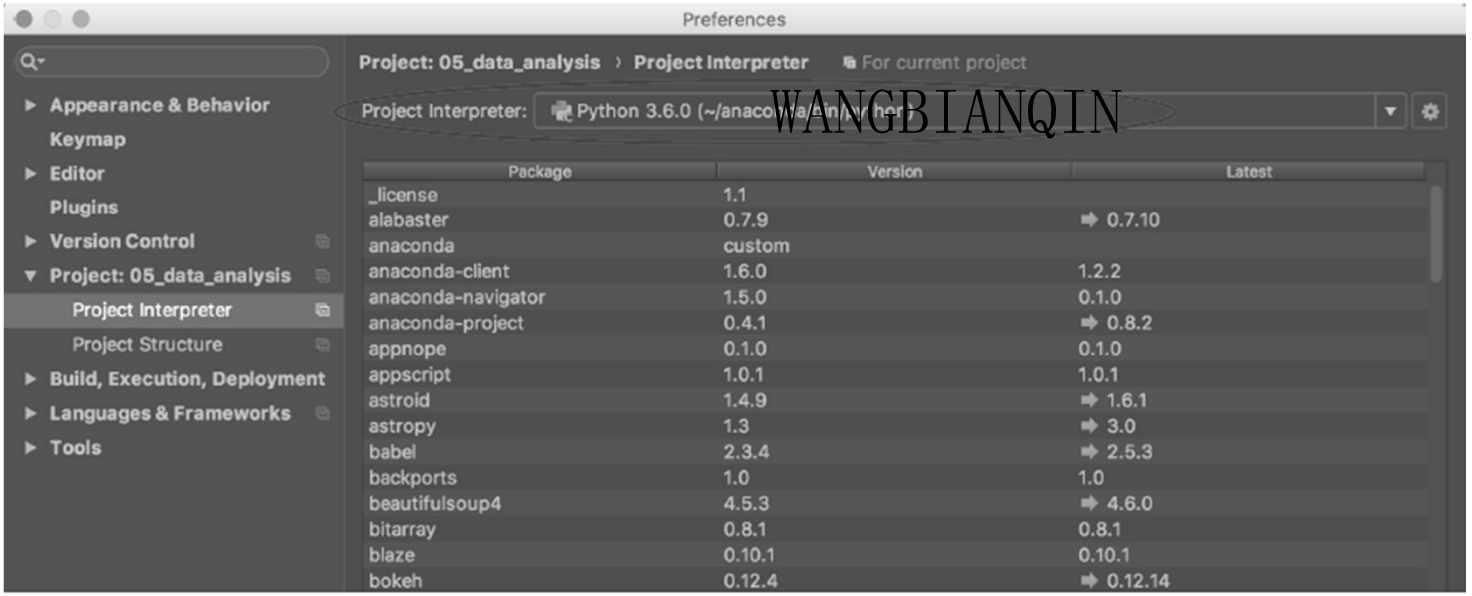
https://www.jetbrains.com/pycharm/



Pycharm安装

PyCharm配置

- 新建项目，选择Python的安装路径



Anaconda组件

- Anaconda Navigtor：管理工具包和环境的图形用户界面
- Jupyter notebook：基于web的交互式计算环境，便于展示数据分析的过程
- Anaconda Prompt：终端，可以使用命令行来管理包和环境。
- Qtconsole：一个可执行 IPython 的仿终端图形界面程序，相比 Python Shell 界面，qtconsole 可以直接显示代码生成的图形，实现多行代码输入执行，以及内置许多有用的功能和函数
- Spyder：使用Python语言、跨平台的、科学运算集成开发环境

Python包管理

Python包管理

- 安装：pip install xxx, conda install xxx
- 卸载：pip uninstall xxx, conda uninstall xxx
- 升级：pip install –upgrade xxx, conda update xxx
- 详细用法： https://pip.pypa.io/en/stable/reference/

#pip show pip
#python -m pip install --upgrade pip

1. conda update conda
2. conda update anaconda

Python虚拟环境

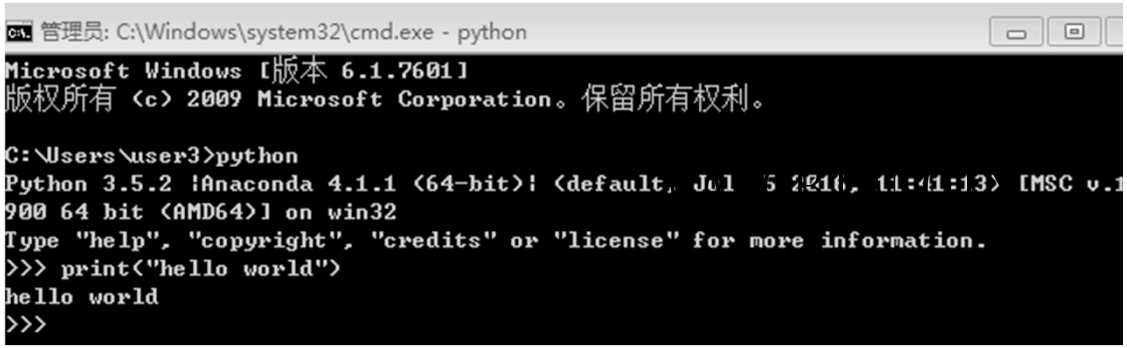
- Virtualenv: https://virtualenv.pypa.io/en/stable/userguide/
- conda 虚拟环境: https://conda.io/docs/using/envs.html

多版本Python管理

- conda管理: https://conda.io/docs/py2or3.html

Hello World

- 交互模式



- 脚本文件 (.py)运行

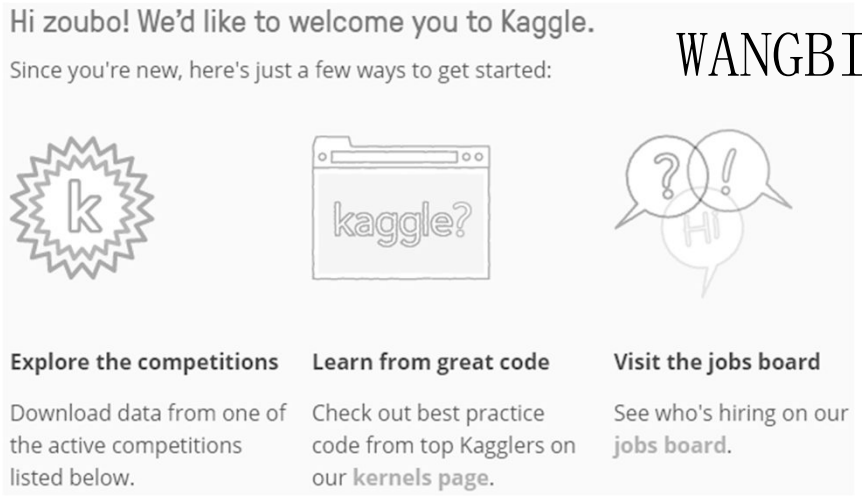
  创建python脚本文件，在文件中输入代码:

  print(“hello world”)

  运行编写好的文件

Kaggle竞赛平台

- 网址：https://www.kaggle.com/
- 注册新账号后的导航界面：



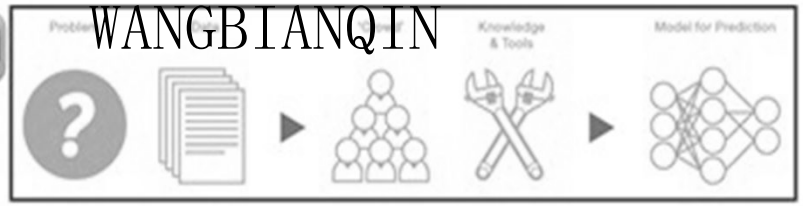
- 企业或者研究者可以将数据、问题描述、期望的指标发布到Kaggle上，以竞赛的形式向广大的数据科学家征集解决方案
- Kaggle上的参赛者将下载数据，分析数据，然后运用机器学习、数据挖掘等知识，建立算法模型，解决问题得出结果，最后将结果提交，如果提交的结果符合指标要求并且在参赛者中排名第一，将获得比赛丰厚的奖金

Twitter用户个性分析案例

- 假设一个互联网广告公司收集了大量关于用户广告点击行为数据，想从这些数据中发现用户点击的规律、模式，以此来优化广告投放、提高用户点击转化率。

Kaggle, Data, Leaderboard

File Name	Formats
Personality_traits_trainingset	.csv(2.27mb)
Personality_traits_testset	.csv(1.49mb)
Sample_code	.R(1.27kb)
Random_forest	.csv(121.7kb)



#	1	2	3	4	5	6
Team	Audaque	NaiveBayes	binghsu	Makagan	LogisticRegression	RandomForest
Loss	0.60698	0.89247	1.27450	2.13716	3.21066	3.90846

Figure 1 Kaggle 的工作方式。

重要会议和期刊

人工智能
会议
IJCAI
AAAI
期刊
Artificial Intelligence
Journal of Artificial Intelligence Research

数据挖掘
会议
KDD
ICDM
期刊
ACM Transactions on Knowledge Discovery from Data
Data Mining and Knowledge Discovery

计算机视觉与模式识别
会议
CVPR
期刊
IEEE Transactions on Pattern Analysis and Machine Intelligence

机器学习
会议
国际机器学习会议(ICML)
国际神经信息处理系统会议(NIPS)
国际学习理论会议(COLT)
欧洲机器学习会议(ECML)
亚洲机器学习会议(ACML)
期刊
Journal of Machine Learning Research
Macine Learning

神经网络
期刊
Neural Computation
IEEE Transaction on Neural Networks and Learning Systems
国内活动:每两年一届的中国机器学习大会CCML，每年举行的机器学习及其应用研讨会MLA