机器学习(Machine Learning)是人工智能（Artificial Intelligence）研究与应用的一个分支。

监督学习（Supervised Learning）：侧重对事物未知表现的预测，一般包括分类问题（Classification）和回归问题（Regression）。

分类问题：类别是离散的，也是预先知道数量的。例如根据人的体重，身高，三维来预测其性别。

回归问题：预测的目标是连续变量，例如根据房屋面积，地理位置等预测其价格。

无监督学习（Unsupervised Learning）：侧重对事物本身特性的分析，常用的技术有降维（Dimensionality Reduction）和聚类（Clustering）

数据降维：对事物的特性进行压缩和筛选，之前做的PCA主成分分析就是一种数据降维方法。

聚类：根据数据的相似性，将相似的数据划分为一个簇。不同于分类问题，大多数情况下我们并不预先知道簇的数量和具体含义。例如，大型点数对用户的信息和购买习惯进行聚类分析，从而针对特定人群投放特定广告。

特征：反映数据内在规律的信息。

特征向量：描述一个数据样本的规律信息。

统计学知识：均值，中值，方差

居中趋势测量：均值、中值、众数

众数：在统计分布上具有明显集中趋势点的数值，代表数据的一般水平（众数可以不存在或多于一个）。 修正定义：是一组数据中出现次数最多的数值，叫众数，有时众数在一组数中有好几个。

数据的离散性：四分位距法、异常值、标准偏差、贝塞尔修正。

不能将测试数据用于训练。

监督学习大体上可以分为：分类学习和回归预测。

监督学习的基本架构和流程：



线性分类

良性/恶性肿瘤分类

例：程序machine\_learning/classfication\_test1.py

import pandas as pd  
import numpy as np  
from sklearn.model\_selection import train\_test\_split  
from sklearn.preprocessing import StandardScaler  
from sklearn.linear\_model import LogisticRegression  
from sklearn.linear\_model import SGDClassifier  
from sklearn.metrics import classification\_report  
  
# 创建特征列表  
column\_names = ['Sample code number', 'Clump Thickness', 'Uniformity of Cell Size', 'Uniformity of Cell Shape', 'Marginal Adhesion', 'Single Epithelial Cell Size', 'Bare Nuclei', 'Bland Chromatin', 'Normal Nucleoli', 'Mitoses', 'Class']  
  
# 读取数据  
data = pd.read\_csv('https://archive.ics.uci.edu/ml/'  
 'machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data', names=column\_names)  
  
# 将缺省值?替换为NumPy的标准缺失值  
data = data.replace(to\_replace='?', value=np.nan)  
data = data.dropna(how='any')  
  
# 将数据保存成csv文件，不加行索引  
# data.to\_csv('total.csv', index=False)  
  
# print(data.shape)  
  
# 数据分割，多少训练，多少预测  
# data[column\_names[1:10]]：样本特征集  
# data[column\_names[10]]：样本标签集  
# X\_train：训练特征集  
# y\_train：训练标签集  
# X\_test：测试特征集  
# y\_test：测试标签集  
X\_train, X\_test, y\_train, y\_test = \  
 train\_test\_split(data[column\_names[1:10]], data[column\_names[10]], test\_size=0.25, random\_state=33)  
  
# 第1列为Excel列号+1  
# 第2列为'Class'列的值  
# print(y\_train)  
  
# 标准化数据，每个维度的特征数据方差为1，均值为0  
# 会将每个维度的数据标准为正负值  
ss = StandardScaler()  
  
# 将数据类型转换为float  
X\_train = X\_train.astype(float)  
X\_test = X\_test.astype(float)  
X\_train = ss.fit\_transform(X\_train)  
X\_test = ss.transform(X\_test)  
  
LR = LogisticRegression()  
SGDC = SGDClassifier()  
  
# 调用LogisticRegression来训练模型  
LR.fit(X\_train, y\_train)  
  
# LR预测  
lr\_y\_predict = LR.predict(X\_test)  
  
# 调用SGDClassifier来训练模型  
SGDC.fit(X\_train, y\_train)  
  
# SGDC预测  
sgdc\_y\_predict = SGDC.predict(X\_test)  
  
# 结果分析  
# LR  
print('Accuracy of LR Classifier: ', LR.score(X\_test, y\_test))  
print(classification\_report(y\_test, lr\_y\_predict, target\_names=['Benign', 'Malignant']))  
  
print('\n')  
  
# SGDC  
print('Accuracy of SGDC Classifier: ', SGDC.score(X\_test, y\_test))  
print(classification\_report(y\_test, sgdc\_y\_predict, target\_names=['Benign', 'Malignant']))

二分类的三个评价指标：

1. 准确性（Accuracy）：



对应到本例中，类别有两种：恶性/良性，记恶性为阳（Positive），良性为阴（Negative），则预测正确恶性为True Positive，预测错误恶性为False Positive，预测正确良性为True Negative，预测错误良性为False Negative。准确性为：



1. 精确率（Precision）：



对于本例：两个类别的精确率分别为：





1. 召回率(Recall)



其中：另1个分类预测错误样本即为应该被预测为本分类，却被识别为另1个分类的样本。

本例中：两个类别的召回率分别为：





另外还有一个调和指标F1 measure：



F-β得分：

F1为F-β在β=1的情况，β<1代表更偏向精度，β>1代表更偏向召回率。

对于多分类的情况，对于某一分类计算精确率和召回率时，将其他分类统一看成另一分类。

混淆矩阵：

|  |  |  |
| --- | --- | --- |
|  | 识别为阳 | 识别为阴 |
| 阳 | True positive | False Negative |
| 阴 | False positive | True Negative |



ROC曲线：

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predict Positive | Predict Negative |  |
| Positive | True Positive(TP) | False Negative(FN) | Actual Positive(TP + FN) |
| Negative | Fasle Positive(FP) | True Negative(TN) | Actual Negative(FP + TN) |
|  |  |  |  |

真正类率(True Positive Rate)：TPR = TP / (TP + FN)，也等于Positive的召回率。

负正类率(False Positive Rate)：FPR = FP / (FP + TN)。

真负类率(True Negative Rate)：TNR = TN / (FP + TN)，也等于Negative的召回率。

TPR又称为敏感度Sensitivity

TNR又称为特异度Specificity

FPR就是1-Specificity。

ROC曲线，横轴为FPR，竖轴为TPR。

ROC曲线的绘制：对于一个分类器，对于二分类问题，可以预测出一个样本为TP的概率，设置不同的阈值，就会得到不同的(FPR, TPR)。

<https://www.cnblogs.com/wuchuanying/p/6243987.html>

理想的分类器，FRP = 0，TPR = 1

AUC：ROC曲线下方的面积，理想的为1，随机情况（最坏的情形）为0.5，面积也可能小于0.5，即TPR的值一直小于FPR，识别正确的真阳性TP一直小于被误识别为阳性FN，但如果进行分类翻转，则会好于AUC为0.5的情形。

R2系数：

SST：总离差平方和，

SSE：残差平方和（误差平方和），



好的模型，R2的值应该接近于1，坏的模型，R2的值接近0.