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代表更偏向召回率，β>1代表更偏向精度。

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。

支持向量机：寻找可能最佳的线性分类器。