监督学习-课程导学

ML11

礼欣
www.python123.org

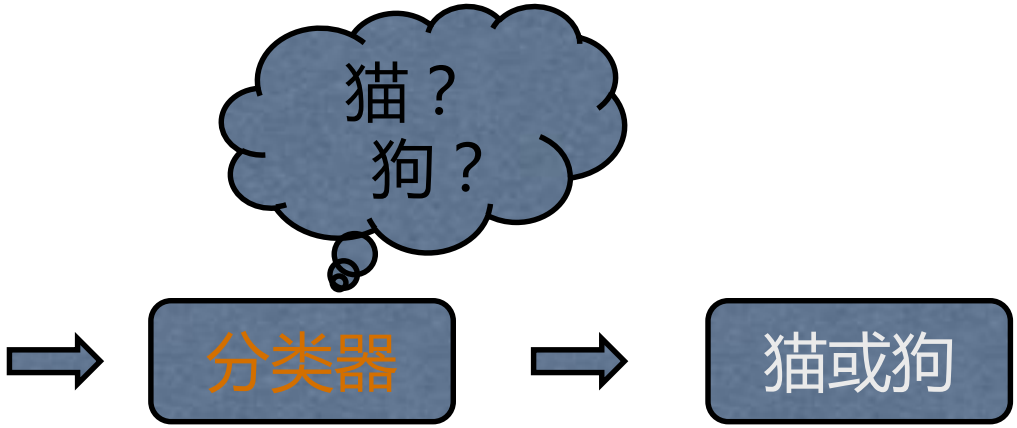
监督学习的目标

利用一组带有标签的数据，学习从输入到输出的映射，然后将这种映射关系应用到未知数据上，达到分类或回归的目的。

分类：当输出是离散的，学习任务为分类任务。

回归：当输出是连续的，学习任务为回归任务。

分类任务



分类学习

输入：一组有标签的训练数据(也称观察和评估)，标签表明了这些数据（观察）的所署类别。

输出：分类模型根据这些训练数据，训练自己的模型参数，学习出一个适合这组数据的分类器，当有新数据（非训练数据）需要进行类别判断，就可以将这组新数据作为输入送给学好的分类器进行判断。

分类学习-评价

- 训练集(training set):顾名思义用来训练模型的已标注数据，用来建立模型，发现规律。
- 测试集(testing set):也是已标注数据，通常做法是将标注隐藏，输送给训练好的模型，通过结果与真实标注进行对比，评估模型的学习能力。

训练集/测试集的划分方法：根据已有标注数据，随机选出一部分数据（70%）数据作为训练数据，余下的作为测试数据，此外还有交叉验证法，自助法用来评估分类模型。

分类学习-评价标准

精确率：精确率是针对我们预测结果而言的，（以二分类为例）它表示的是预测为正的样本中有多少是真正的正样本。那么预测为正就有两种可能了，一种就是把正类预测为正类(TP)，另一种就是把负类预测为正类(FP)，也就是

$$P = \frac{TP}{TP + FP}$$

分类学习-评价标准

召回率：是针对我们原来的样本而言的，它表示的是样本中的正例有多少被预测正确了。那也有两种可能，一种是把原来的正类预测成正类(TP)，另一种就是把原来的正类预测为负类(FN)，也就是

$$R = \frac{TP}{TP + FN}$$

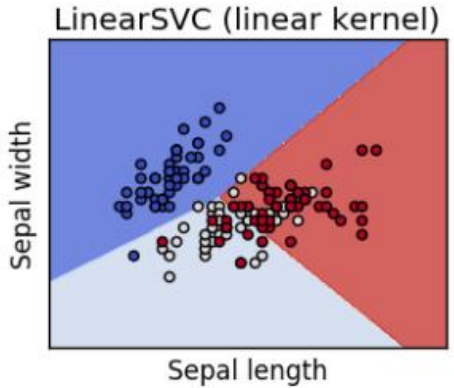
分类学习-评价标准

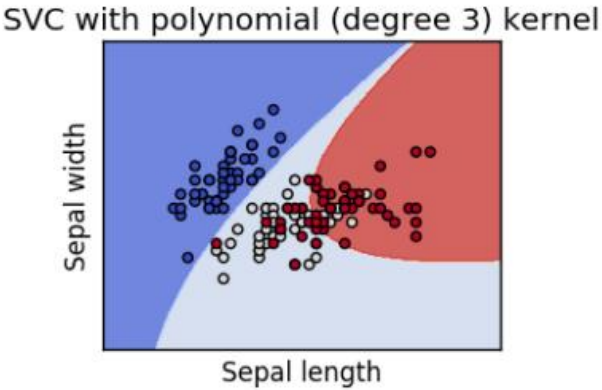
假设我们手上有60个正样本，40个负样本，我们要找出所有的正样本，分类算法查找出50个，其中只有40个是真正的正样本，TP：将正类预测为正类数 40；FN：将正类预测为负类数 20；FP：将负类预测为正类数 10；TN：将负类预测为负类数 30

准确率（accuracy）=预测对的/所有 = (TP+TN)/(TP+FN+FP+TN) = 70%
精确率（precision）=?
召回率（recall）=?

Sklearn vs. 分类

与聚类算法被统一封装在sklearn.cluster模块不同，sklearn库中的分类算法并未被统一封装在一个子模块中，因此对分类算法的import方式各有不同。

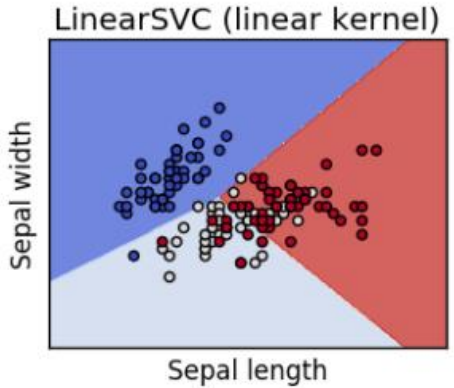


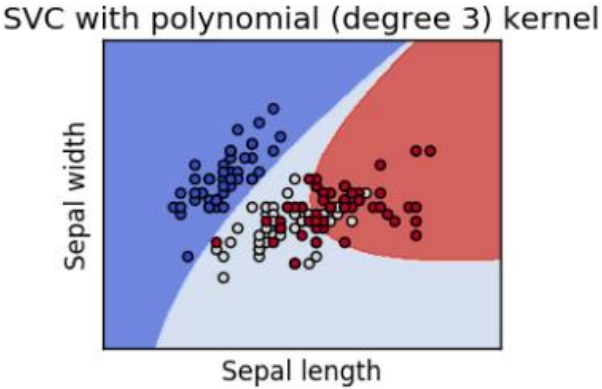


Sklearn vs. 分类

Sklearn提供的分类函数包括：

- k近邻（knn）
- 朴素贝叶斯（naivebayes），
- 支持向量机（svm），
- 决策树 （decision tree）
- 神经网络模型（Neural networks）等
- 这其中有线性分类器，也有非线性分类器。





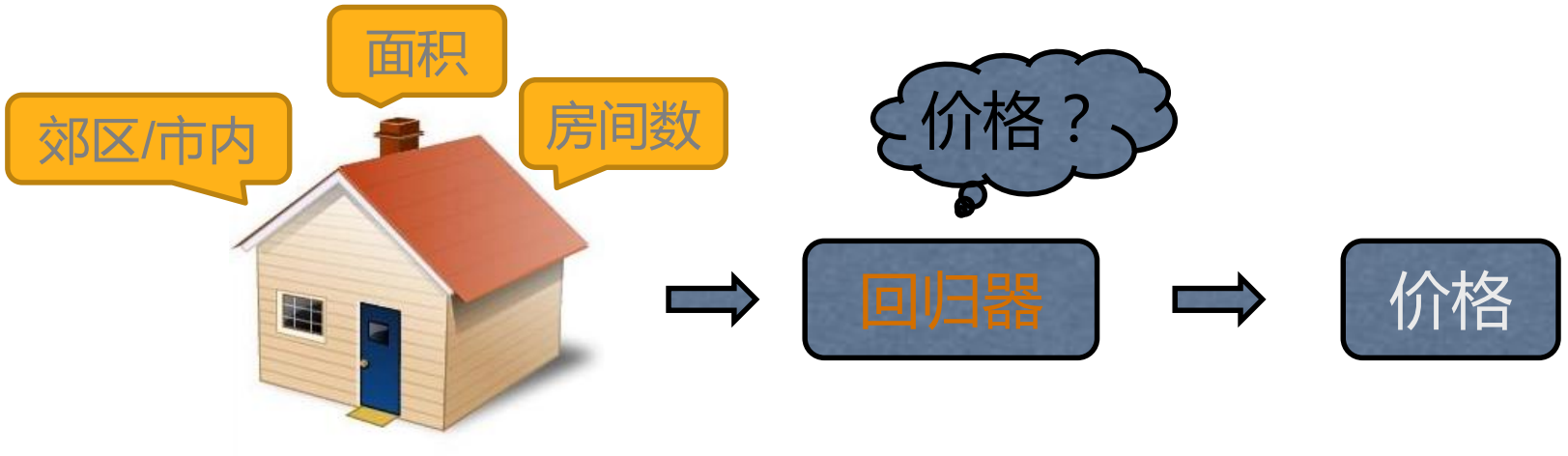
分类算法的应用

- 金融：贷款是否批准进行评估
- 医疗诊断：判断一个肿瘤是恶性还是良性
- 欺诈检测：判断一笔银行的交易是否涉嫌欺诈
- 网页分类：判断网页的所属类别，财经或者是娱乐？

回归分析

回归：统计学分析数据的方法，目的在于了解两个或多个变数间是否相关、研究其相关方向与强度，并建立数学模型以便观察特定变数来预测研究者感兴趣的变数。回归分析可以帮助人们了解在自变量变化时因变量的变化量。一般来说，通过回归分析我们可以由给出的自变量估计因变量的条件期望。

回归任务



Sklearn vs. 回归

Sklearn提供的回归函数主要被封装在两个子模块中，分别是sklearn.linear_model和sklearn.preprocessing。sklearn.linear_modlel封装的是一些线性函数，线性回归函数包括有：

- 普通线性回归函数（ LinearRegression ）
- 岭回归（Ridge）
- Lasso（Lasso）

非线性回归函数，如多项式回归（PolynomialFeatures）则通过sklearn.preprocessing子模块进行调用

回归应用

回归方法适合对一些带有时序信息的数据进行预测或者趋势拟合，常用在金融及其他涉及时间序列分析的领域：

- 股票趋势预测
- 交通流量预测