机器学习实战

机器学习通常有以下几个步骤：

1. 收集数据
2. 准备输入数据
3. 分析输入数据
4. 训练算法（可使用交叉验证）
5. 测试算法（一般是监督学习算法需要此步骤）
6. 使用算法

本书算法代码使用的是Python语言，本书使用了Python库NumPy/Beautiful Soup/Mrjob/Vote Smart/Python-Twitter。

## 第一章 K-近邻算法

### 1 算法原理及实现

#### 1.1 原理

在训练集中选取离输入的数据点最近的k个邻居，根据k个邻居中出现最多的类别（最大表决规则），作为该数据点的类别。

#### 1.2 实现过程

1. 计算所有点与当前的距离

根据距离值进行排序

1. 选取与当前点距离最短的K个点
2. 确定K个点所在的类别
3. K个点中频率最高的类别可当作当前点的类别

### 2 算法优缺点

K值的选取对kNN学习模型有很大的影响。若K值过小，预测结果会对噪音样本点显得异常敏感。

优点：精度高、对异常值不敏感、无数据输入假定。

缺点：计算复杂度度高，空间复杂度高。

适用数据范围：数值型与标称型。

### 3 算法应用

1. 文本分类--使用K-近邻算法改进约会网站的匹配效果

### 识别手写数字

## 第二章 决策树

### ID3算法原理及实现

#### 1.1 原理

决策树是一个树结构（可以是二叉树或者非二叉树）。其每个非叶节点表示一个特征属性上的测试，每个分支代表这个特征属性分类结果。通过训练数据可以构造出一个决策树，测试数据从根节点输入，从叶子节点输出，该子节点的分类就是测试点的分类。

#### 1.2 信息增益

信息熵表示的是不确定度，均匀分布时，不确定度最大，此时熵就最大。在划分数据集之前之后信息发生的变化称为信息增益。信息增益可以衡量某个特征对分类结果的影响大小。集合信息的度量方式称为香农熵。

熵定义为信息的期望值，在明晰这个概念之前，我们必须知道信息的定义。如果待分类的事物可能划分在多个分类之中，则符号的信息定义为



其中是选择该分类的概率。

为了计算熵，我们需要计算所有类别所有可能值包含的信息期望值，通过下面的公式得到：，其中n是分类的数目。

#### 1.3 算法实现

1、计算数据香农熵

2、划分数据并选择最好的数据集划分方式

3、直到所有数据属于同一类，划分结束

4、创建树并保存

### 2 算法优缺点

优点：决策过程非常直观，容易被人理解；计算复杂度不高，且决策树可存储；对中间值的确定不敏感，可以处理不相关特征数据。

缺点：可能会产生过度匹配的问题。

### 3 算法应用

1. 使用决策树预测隐形眼镜类型

## 第三章 朴素贝叶斯

### 1 算法原理及实现

#### 1.1 原理

条件概率定义：设A,B是两个事件，且P(A)>0称P(B|A)=P(AB)/P(A)为在条件A下发生的条件事件B发生的条件概率。

贝叶斯公式

 

从而得到X属于各类的概率，最高的就是X的类别。

#### 1.2 算法实现

1. 首先需要构建词向量
2. 在多分类中，对于某一门类，用一个向量PVect记录所有关于本类的词向量的向量加，另一个标量PNum记录所有词的总量,则Pvect/Pnum就是概率P(A|B)。
3. 测试数据，针对数据词向量中的词计算出就每个类别的概率。
4. 比较概率，得出结果

### 算法优缺点

优点:可识别多类别问题，在数据较少的情况仍有效。

缺点：对输入数据的准备方式较为敏感。

### 3 算法应用

1. 文档分类
2. 垃圾邮件检测

垃圾邮件一般是宁愿多一点垃圾邮箱，也不想误删非垃圾邮件。

## 第四章 Logistic回归

### 1算法原理及实现

#### 1.1 算法原理

Sigmoid函数 

### 2算法优缺点

### 3 算法应用