## 1统计机器学习方法概论

统计机器学习：从数据出发，抽取出数据的特征，抽象出数据的模型，发现数据中的规律，然后用它来对数据进行分析和预测。

## 2.感知器

### 2.1感知器模型

损失函数：基于误分类使得损失函数最小化

学习方法：梯度下降法

感知器是神经网络和支持向量机的基础。



### 2.2感知机学习策略

定义学习策略，即定义一个损失函数并使其最小化

损失函数：误分类的总数（非连续可导，不易优化）

误分类点到超平面的距离（连续可导）

最小化过程推导：点到直线的距离表示为：

特征空间中任意点x到超平面S的距离为：

对于误分类点来说：

误分类点到分离超平面的距离为：

所有误分类点到分离超平面的距离为：

去除常数以后，可以得到损失函数为：

这个损失函数就是感知机学习的经验风险函数

感知器的学习问题（求出分离超平面） = = 损失函数的最小化问题（随机梯度下降方法）

### 2.3感知机学习方法

#### 2.3.1原始形式

算法描述：1 确定初始的w，b，以及学习率η。

2 把所有训练数据代入感知机模型

3 找出分类错误的数据，即的数据，任意选取一个错误数据来更新w、b，更新规则为：

4回到第2步，直至没有错误数据

收敛性证明

#### 2.3.2对偶形式



W、b可以被描述为：

算法描述：

## 3.KNN算法

### 3.1算法主要思想：

设训练数据集为 ，根据一定的距离度量规则，来寻找给定的数据的最相近的k个数据组成的集合，由集合来决定所属的类别。

即 ，

其中当时为1，否则为0

### 3.2主要问题

由上可见，KNN算法主要包含三个问题：决定距离度量的方式；决定聚类数目k；决定分类决策规则。

#### 3.2.1距离度量

距离度量反映了两个数据之间的相似程度。

距离：，其中l为空间维数

当p为1时，称为曼哈顿距离

当p为2时，称为欧氏距离

#### 3.2.2聚类数目k

当选取较小的k值时，近似误差较小，但估计误差较大，整体模型复杂，容易过拟合；

当选取较大的k值时，估计误差较小，但近似误差较大，整体模型简单，甚至当k为无穷时，相当于数据集的所有数据参与决策，模型过于简单。

一般选择一个较小的k值，采取交叉验证的方法迭代选取最优的k值。

#### 3.2.3分类决策规则

一般采取多数表决规则，即k个近邻数据中类别最多的为数据所属的类别。误分类率为：



这等价于经验风险最小化。

### 3.3 KNN算法实现问题

KNN算法实现起来的主要难度在于如何对数据进行快速的k近邻搜索。当数据集较小时，可以采用线性搜索，即计算每个数据的k近邻时，把所有的数据算一遍。当数据集较大时，就得考虑使用特殊的数据结构来减少计算距离的时间。这里选用kd树数据结构来实现，可以省去对大多数数据的搜索。

## 4 朴素贝叶斯方法

### 4.1基本方法

先验概率分布P(Y)

后验概率分布P(Y｜X)

条件概率分布P(X｜Y)

首先，基于特征条件独立假设，学习输入输出的联合概率分布P(X,Y)= P(X｜Y) P(Y)

然后，基于学习到的模型，利用贝叶斯定理求出使得后验概率P(Y｜X)最大的输出y

朴素贝叶斯法分类的基本公式——



后验概率最大化的含义为期望风险最小化

算法4.1 朴素贝叶斯法的基本流程

1）计算先验概率，条件概率

2）对于给定的样本x，计算后验概率

3）找到使得后验概率最大的类别

### 4.2朴素贝叶斯法的参数估计

#### 4.2.1极大似然估计

先验概率

条件概率

#### 4.2.2贝叶斯估计

先验概率

条件概率

## 8提升方法