机器学习复习

## 第二章 概念学习和一般到特殊序

1. 概念：每个概念可被看作对象或事件的集合，是一个大集合的子集。或是这个大集合中定义的布尔函数。如动物集animal，以及布尔函数isBird(animal)。

2. 概念学习：从有关某个布尔函数的输入输出训练样例中推断出该布尔函数。输入：布尔函数的真值表，输出：该布尔函数。

3. 假设：可能的布尔函数，在这里用各属性的合取式来进行表示。

文本, 信件

描述已自动生成

4. 假设、假设空间和实例之间的关系

文本, 信件

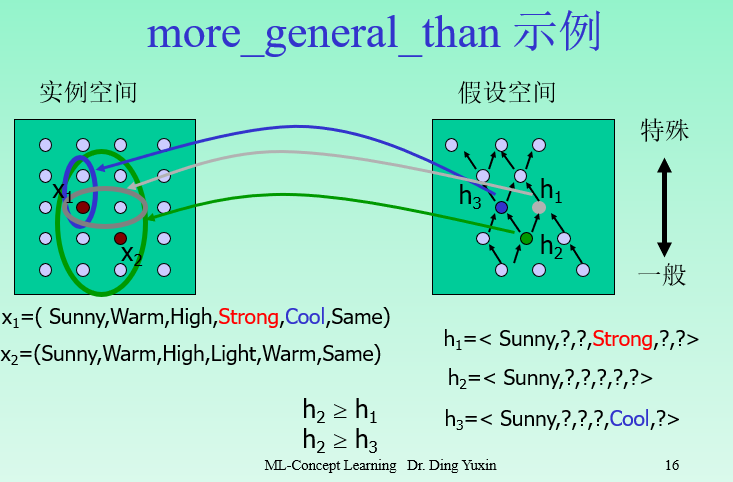
描述已自动生成

5. 归纳学习和归纳学习假设

* 归纳学习：从样例中获取普遍规律。只保证输出假设在训练数据上符合目标概念。
* 归纳学习假设：任一假设如果在足够大的训练样例集中很好地逼近目标函数，它也能在未见实例中很好地逼近目标函数。

6. 假设的偏序关系

* more\_general\_than\_or\_equal\_to 定义和为X上定义的布尔函数，称当且仅当。即当时，一定为1，可记作。



7. Find-S算法

从H中最特殊的假设开始，在该假设覆盖正例失败时将其一般化，忽略遇到的一切反例。为什么可以忽略遇到的一切反例？

* 假设目标概念c存在与假设空间H中，那么c对D一定是一致的。
* Find-S得到的假设h是对D中正例一致的最特殊假设，那么c比h更一般。
* 由更一般的定义，，于是h对所有反例也一致。

图示, 示意图

描述已自动生成

8. 变型空间

* 一致的定义
  + 一个假设h与训练样例集合D一致，当且仅当对D中每一个样例<x, c(x)>都有h(x)=c(x)。
* 满足的定义
  + 当且仅当样例x使得假设h为1，即h(x)=1，称x满足h
* 变型空间
  + **与训练样例一致的所有假设构成的集合**，包含了目标假设所有合理的变型。变型空间因此得名。

文本

描述已自动生成

* 变型空间更简洁的表示
  + 变型空间被表示为它的极大一般和极大特殊成员。这些成员形成了一般和特殊的边界，划分出变型空间。
  + S边界：特殊边界，是H中与D相一致的极大特殊成员的集合
  + G边界：一般边界，是H中与D相一致的极大一般成员的集合

文本, 信件

描述已自动生成

9. 候选消除算法

生成G边界和S边界的过程。

手机屏幕截图

描述已自动生成

文本, 信件

描述已自动生成

**遇到正例，从G和S中移去不一致的假设s，极小一般化S边界。**

**遇到反例，从G和S中移去不一致的假设s，极小特殊化G边界。**

注意：

* 极小一般化：
  + Input: <?, x, ?>, <?, y, ?>
  + Output: <?, ?, ?>
* 极小特殊化：
  + Input: <?, ?, ?>, <x1, y1, z1>
  + Output: <x2, ?, ?>, <?, y2, ?>, <?, ?, z2>
* 收敛条件：
  + 训练样例中没有错误
  + H中确实包含目标概念的正确假设

10. 归纳偏置

* **归纳学习需要某种形式的预先假定，称为归纳偏置**
* 无偏学习的无用性：无法对未见实例做分类。这是因为对一个未知实例x，在无偏的假设空间中，总有假设h1和h2，使得h1(x)=T,h2(x)=F，无偏学习无法确定h1和h2哪个是正确的。
* **机械式学习器的归纳偏置**：没有偏置
* **候选消除算法的归纳偏置**：假设空间中包含目标概念
* **Find-S算法的归纳偏置**：假设空间中包含目标概念，任何实例，除非它的逆实例可由其他知识逻辑推出，否则它为反例
* 算法有偏性越强，它的归纳能力越强，能分类更多未见实例

## 第三章 决策树学习

* ID3算法
  + ID3算法的归纳偏置
    - **短树比长树优先**
    - 高信息增益属性更靠近根节点的树优先
  + 解决ID3算法的过拟合
    - 错误率降低修剪：保证修剪后的树在验证集上性能不比原始树差
    - 规则后修剪：将树转换成等价的规则，删除规则前件，需要保证删除后可以提高估计精度
  + ID3 vs. 候选消除算法
    - ID3是优选偏置，即假定某些假设胜过其他假设
    - 候选消除算法是限定偏置，对待考虑假设的一种限定，这里是合取

## 第四章 神经网络学习

* 感知器的定义（线性阈值单元下）
* 感知器的表征能力
  + 单层能表征与、或
  + 双层能表示所有布尔函数
* 感知器的训练法则
  + 感知器法则

文本

中度可信度描述已自动生成

收敛条件：训练数据线性可分，学习率足够小

* + Delta法则
    - 使用梯度下降求得，能够克服线性可分的约束
    - 在E为MSE的情况下，，其中，为线性单元的输出

文本, 信件

描述已自动生成

* + - GD和SGD的差别
      * GD法则对所有训练样例求和后计算权值更新，SGD根据每一个单独样例的误差增量计算权值更新

文本

描述已自动生成

文本

描述已自动生成

注：这里的o是线性单元的输出，即

* + Delta法则和感知器法则的区别：
    - 感知器法则根据阈值单元的误差更新权重
    - Delta规则根据线性单元的误差来更新权重
* 前馈神经网络的表征能力
  + - 两层表示任意连续函数
    - 三层表示任意函数
* 反向传播算法
  + - 本质上就是对各个权值进行梯度更新。
    - 遵循Delta法则。以sigmoid函数为例，

文本, 信件

描述已自动生成

* BP算法过拟合问题
  + 损失函数中添加正则项，优化损失函数时也会降该权值变小
  + 使用交叉验证的方法
* 反向传播算法的归纳偏置
  + 梯度下降搜索
  + 权空间覆盖可表征函数空间的方式的相互作用性

## 第五章 贝叶斯学习

* 基础概率公式

文本

描述已自动生成

* 极大后验假设（MAP）和极大似然假设（ML）
  + MAP假设：使得p(h|D)最大的假设，即在观察到D的情况下最有可能的假设

文本, 信件

描述已自动生成

* + ML假设

假定每个假设都有相同的先验概率p(h)

则可进一步简化为：

文本, 信件

描述已自动生成

* 贝叶斯框架
  + 通过定义p(h)和p(D|h)，可以通过贝叶斯得到概率意义上的最优假设h。
  + P(h)和p(D|h)可以看作该算法表现最优时的隐含假定，也是学习器的归纳偏置。
* 贝叶斯最优分类器

文本, 信件

描述已自动生成

* 朴素贝叶斯

文本, 信件

描述已自动生成

文本, 信件

描述已自动生成

文本, 信件

描述已自动生成

## 第六章 基于实例的学习

* K-近邻（KNN）
  + 归纳偏置：的类别与它附近的实例的类别相似
  + 离散目标函数：投票法，k个相邻样本中最普遍的类别
  + 连续实值目标函数：用k个相邻样本的均值表示
  + 改进：加入权重w，w与距离成反比
  + 实现：KD树
* 局部加权回归
  + K近邻的推广。通过相邻样本和一些简单函数拟合出f。
  + 以线性拟合为例：
    - 定义近似函数
    - 定义误差函数

文本, 信件

描述已自动生成

* + - 定义训练规则（通常是梯度下降）
* 径向基函数（RBF）

文本, 信件

描述已自动生成

* 消极学习和积极学习
  + 消极学习：k近邻和局部加权回归
    - 消极学习器用查询点附近的多个局部逼近的组合表示目标函数
  + 积极学习：RBF
    - 积极学习器需要在训练时间完成全局逼近函数的学习

## 第七章 线性分类器

* 多类问题线性分类器
  + 划分问题
    - 一对其余
      * 表示属于i类，表示不属于i类
      * 若只有一个大于0，其余都小于等于0，认为属于大于0的那类
      * m类有m个判别函数，拒识区域较多

文本

描述已自动生成

* + - 一对一
      * 表示属于i类而不属于j类。
      * 若对所有j，都有，则认为属于i类
      * m类有m(m-1)/2个判别函数，拒识区域较少

文本

描述已自动生成

* + - 多对多（最常用）
      * 表示属于i类而不属于j类
      * 若对任意j，都有，则认为属于i类
      * m类有m个判别函数，没有拒识区域（拒识区域为线）

文本

中度可信度描述已自动生成

* 最近邻准则：实际上是线性分类器
* 感知器准则：适用于线性可分
* 最小平方误差准则：适用于线性不可分

## 第八章 无监督学习

* K-means

文本, 信件

描述已自动生成

* + 优势：简单，效率
  + 劣势：
    - 均值要能够被定义
    - K需要被指定
    - 算法对异常值敏感
    - 算法对初始种子敏感
    - 算法不适合发现那些形状不是椭球体的聚类
* 合并聚类
  + 伪代码

文本

描述已自动生成

## 需要准备的题型

1. 概念部分
2. 概念学习
   1. 从有关某个布尔函数的相关输入输出样例中推断出该布尔函数。
3. 变型空间
   1. 与数据集一致的假设所构成的集合
4. 归纳偏置
   1. 归纳学习中需要预先设立的假定
5. 熵
   1. 信息的混乱程度
6. 条件独立性
7. 消极学习与积极学习
   1. 消极学习在决定如何从训练数据D中泛化时考虑查询实例
   2. 积极学习在见到查询实例前就选取好了对目标函数的逼近
8. 利用线性判别函数解决多类决策问题
   1. 一对其余
   2. 一对一
   3. 多对多
9. 感知器定义
10. 特征选择
    1. 子集搜索
       1. 前向：增加特征
       2. 后向：删除特征
    2. 过滤式
    3. 包裹式
    4. 嵌入式
11. 算法部分
12. 候选消除算法
13. ID3决策树算法
14. BP算法
15. K均值聚类
16. 合并聚类
17. 最近邻和K近邻
18. 朴素贝叶斯算法

## PPT上的例题

文本

描述已自动生成

文本, 信件

描述已自动生成

文本, 信件

描述已自动生成

文本, 信件

描述已自动生成

文本

描述已自动生成