方骏-2020年3月8日-读书报告

1. 自己提出的问题的理解：
2. 提出的问题1：结构风险最小化中的复杂度项J(f)是怎么确定的？

讨论后的理解：这个复杂度项表示的就是模型的复杂度，模型越复杂，复杂度项应该是越高的，但是这个复杂度项的选择要根据模型而定，能够代表模型的复杂度，通常是通过求模型函数的泛函来作为复杂度项。

1. 别人提出的问题的理解：
2. 问题2：1.7中，生成方法和判别方法分别适用于什么情况？

自己的理解：生成方法学习联合概率密度分布P(X,Y)，所以就可以从统计的角度表示数据的分布情况，能够反映同类数据本身的相似度。但它不关心到底划分各类的那个分类边界在哪。生成方法可以还原出联合概率分布P(Y|X)，而判别方法不能。生成方法的学习收敛速度更快，即当样本容量增加的时候，学到的模型可以更快的收敛于真实模型，当存在隐变量时，仍可以用生成方法学习。此时判别方法就不能用。判别方法直接学习的是决策函数Y=f(X)或者条件概率分布P(Y|X)。不能反映训练数据本身的特性。但它寻找不同类别之间的最优分类面，反映的是异类数据之间的差异。直接面对预测，往往学习的准确率更高。由于直接学习P(Y|X)或P(X)，可以对数据进行各种程度上的抽象、定义特征并使用特征，因此可以简化学习问题。生成学习方法更实用于那些需要知道具体模型并可以类比在同类数据中的情况，但是判别方法更关注于简单的分类问题，如果只是要分类，就没必要进行复杂的建模。

1. 问题3： 能否简单解释一下联合分布具体指什么？

自己的理解：联合分布就是指X和Y两个变量都一起满足一个特别的概率分布，当X和Y为什么值时，都可以知道此时的概率。比如，X=2且Y=3时的概率就可以通过联合分布直接求出来，这样通过联合分布就可以求出条件概率，可以进行比较准确的数据预测，当X为多少时，Y为多少的概率就可以全部求出，并选择概率大的即可。

1. 问题4： 奥卡姆剃刀原理中，很好的解释已知数据评判标准是什么？

自己的理解：就是说，能够对数据的分析做的好，且模型复杂度小的模型往往是最好的。当有很多解释数据效果，即泛化误差很小的模型时，选择复杂度最小的模型是最小的。

1. 读书计划

1、本周完成的内容章节：《统计机器学习》第一章

2、下周计划：《统计机器学习》第二、三章

四、读书摘要及理解

1、统计学习是以数据为研究对象的学科，目的是对数据进行预测和分析，统计学习以方法为中心，统计学习方法构建模型并应用模型进行预测与分析，统计学习以方法为中心，统计学习方法构建模型并应用模型进行预测与分析，统计学习是概率论、统计学、信息论、计算理论、最优化理论及计算机科学等多个领域的交叉学科，并且在发展中逐步形成独自的理论体系与方法论。

2、统计学习包括监督学习、非监督学习、半监督学习及强化学习。监督学习的任务是学习一个模型，使模型能够对任意给定的输入，对其相应的输出做出一个好的预测。在监督学习中，将输入与输出所有可能取值的集合分别称为输入空间与输出空间。输入与输出空间可以是有限元素的集合，也可以是整个欧氏空间。

每个具体的输入是一个实例，通常由特征向量表示。所有特征向量存在的空间称为特征空间。特征空间的每一维对应于一个特征。模型实际上就是定义在特征空间上的。监督学习从训练数据集合中学习模型，对测试数据进行预测，训练数据和测试数据都是由输入与输出对组成。同时监督学习中假设输入与输出的随机变量X和Y遵循联合概率分布P(X, Y)。

模型属于由输入空间到输出空间的映射的集合，这个集合是假设空间。监督学习的模型可以是概率模型或非概率模型。

3、统计学习三要素：统计学习方法都是由模型、策略和算法构成的，即统计学习方法由三要素构成，可以简单地表示为：

方法=模型+策略+算法

3.1、模型。统计学习首要考虑的问题就是学习什么样的模型。在监督学习过程中，模型就是所要学习的条件概率分布或决策函数。模型的假设空间包含所有可能的条件概率分布或决策函数。假设空间中的模型一般有无穷多个。假设空间定义为决策函数的集合：



假设空间定义为条件概率的集合：



3.2、策略。统计学习的目标在于从假设空间中选取最优模型。损失函数和风险函数可以度量模型的一次预测的好坏和平均意义下模型预测的好坏。

损失函数是度量预测错误的程度，是f(X)和Y的非负实值函数，常用的损失函数有0-1损失函数、平方损失函数、绝对损失函数、对数损失函数等。损失函数越小，模型越好。

经验风险最小的模型一般认为是最优的模型，但是样本容量要足够大时，经验风险最小化能保证有很好的学习效果，但是样本容量很小时，会出现过拟合，此时就是用结构风险最小化，加上了正则化项，控制模型的复杂度，防止过拟合。

3.3、算法。算法就是指学习模型的具体计算方法，从假设空间中选择了最优模型后，考虑用什么样的计算方法求解最优模型，可以借用之前的优化算法，也需要自己开发独自的最优化算法。

4、模型的评估也是重要的。当损失函数给定时，基于损失函数的模型的训练误差和模型的测试误差就自然成为学习方法评估的标准。训练误差是模型关于训练数据集和平均损失：



测试误差是模型关于测试数据集的平均损失：



训练误差的大小，对判断给定的问题是不是一个容易学习的问题是有意义的，但本质上不重要。测试误差反映了学习方法对未知的测试数据集的预测能力。测试误差小的方法具有更好的预测能力，是更有效的方法。

当模型的复杂度增大时，训练误差会逐渐减小并趋向于0，而测试误差会先减小，达到最小值后又增大。当选择的模型复杂度过大时，过拟合现象就会发生。适当的选择模型的复杂度，让测试误差达到最小，这是我们的目的。

5、模型选择的典型方法是正则化，就是结构风险最小化策略的实现，在经验风险上加一个正则化项或罚项。正则化项一般是模型复杂度的单调递增函数，模型越复杂，正则化值就越大，目的就是让此达到最小，能够很好的解释数据且十分简单才是最好的模型。

交叉验证在数据挖掘书中学过，这里可以参考之前的笔记。

6、泛化能力是指由该方法学习到的模型对位置数据的预测能力，是学习方法本质上重要的性质。现实中采用最多的方法就是通过测试误差来评价学习方法的泛化能力。测试数据集是有限的，所以需要给出理论上的泛化能力分析。泛化误差就是：



反映了学习方法的泛化能力。如果一种方法学习的模型比另一种方法学习的模型具有更小的泛化误差，那么这种方法就更有效。研究泛化误差的上界就是界定泛化能力的，但是这通常根据模型的不同而不同。

7、监督学习方法又可以分为生成方法和判别方法，生成方法是由数据学习联合概率分布P(X,Y)，然后求出条件概率分布P(Y|X)作为预测的模型，即生成模型：



判别方法由数据直接学习决策函数f(X)或者条件概率分布P(Y|X)作为预测的模型，即判别模型。判别方法关心的是对给定的输入X，应该预测什么样的输出Y。

8、 分类问题和标注问题在数据挖掘中学习过具体的算法和评价标准，这里就不再赘述。

9、回归问题是用于预测输入变量和输出变量之间的关系，特别是当输入变量的值发生变化时，输出变量的值随之发生的变化。回归模型正是表示从输入变量到输出变量之间映射的函数。回归学习等价于函数拟合，常用的损失函数是平方损失函数。