HIT 2020 秋机器学习期末复习

姓名：秦海滨

学号：1180300523

2020 年 11 月 21 号

目录

1 机器学习课程的目的与研究内容

1.1 什么是机器学习？

是寻找一种对自然或人工现象可预测且可执行的机器理解方法。一般形式为通过对给定数据的分析建立模型，以良好的泛化能力处理训练数据之外的数据。

1.2 机器学习的应用以及机器学习的一般方法

机器学习常用于自然语言处理、语音识别、对象识别、机器人控制、文本挖掘等方面。

机器学习的一般泛型分为：监督学习、无监督学习以及强化学习。

1.3 决策树（判别式模型）

决策树（decision tree）是一种常见的机器学习方法，通过对训练数据的学习，建立一棵树型结构对数据进行分类判别。

基本流程为：基于一个可以最优化某项准则的属性来对样本集合进行划分，决策树的生成是一个递归的过程，一般有以下三种情况会导致递归返回（停止划分）：

1. 当前节点包含的样本属于同一类别，无需划分；
2. 当前属性集为空，或是所有样本在所有属性上取值相同，无法进行划分；
3. 当前节点包含的样本集合为空，不能划分。

在课程中以“信息增益”对决策树进行划分，信息增益衡量了划分节点前后数据“纯度”的提高程度。信息增益即划分前后数据集合的信息熵的差值。计算数据集的信息熵的公式如下：

$$H(D) = -\sum_{i=1}^{N} P(x=i) log_2 P(x=i)$$

在本章节还引入了自信息、条件熵、互信息、交叉熵等概念。

决策树通过剪枝方法克服过拟合，一般分为预剪枝和后剪枝。

需要掌握决策树对于连续属性和缺失属性的处理方式。对于连续属性，一般通过二分划分离散化后，当作离散属性处理；对于缺失属性，可以通过为每个样本加定权值的方式，计算未缺失样本的熵，再乘以权重作为整个样本的熵。

2 统计学习的建模工具

2.1 复习概率论主要概念以及方法并补充本课程的新概率知识

2.2 频率论和贝叶斯观点

频率主义学派（Frequentisi）认为模型参数虽然未知，但是客观存在的固定值，因此可以通过优化似然函数等准则来确定参数值；贝叶斯学派（Bayesian）认为参数是未观察到的随机变量，但其本身也有自己的分布，因此可以假设参数服从一个先验分布，然后基于观测到的数据来计算参数的后验分布。

2.3 最大似然法和最大后验法

参数估计的典型方法有代表频率主义学派观点的最大似然估计法（Maximum Likelihood Estimation, MLE）以及代表贝叶斯学派观点的最大后验估计法（Maximum A Posteriori Probability, MAP）。

MLE 将模型参数视作一个常量，要求我们选出一个能使当前数据集发生概率最大的参数 θ，即

$$\hat{\theta}_{MLE} = \arg\max_{\theta} P(D|\theta)$$

MAP 将模型参数视作一个有自己分布的变量，通过给定的数据以及先验，对参数进行估计：

$$\hat{\theta}_{MAP} = \arg\max_{\theta} P(\theta|D) = \arg\max_{\theta} P(D|\theta)P(\theta)$$

MLE 在数据集较小时容易发生过拟合，MAP 的估计结果与先验有着很大的关系。

3 回归分析与过拟合

3.1 线性回归

详见机器学习 Lab1 多项式拟合正弦曲线。

3.2 过拟合

过拟合（Overfitting）的定义为：学习器对于训练样本的学习“太好”，导致学习器将训练样本的一些自身特点当作了样本整体所具有的一般性质，导致学习器的泛化性能下降。在之后的每个模型中都会提到相关的如何在一定程度上克服过拟合问题。

与过拟合相对的有欠拟合（Underfitting）问题，指对训练样本的一般性质没有学习完全，无法对样本有效判别。

3.3 特征变换

在后续的核函数以及高维空间和低维空间之间的相互转换中都有设计到特征转换。

3.4 最小二乘和最大似然

在衡量模型预测结果与实际结果时通常会使用最小二乘法进行预估。详见Lab1。

4 贝叶斯判别

4.1 最优分类器

贝叶斯分类器是一类分类算法的总称，贝叶斯定理是这类算法的核心，因此统称为贝叶斯分类。贝叶斯决策论通过相关概率已知的情况下利用误判损失来选择最优的类别分类。为了最小化总体风险，只需要最小化每个样本风险，最小化分类错误的贝叶斯分类器为：

$$h^*(x) = \arg\min_{c\in\mathcal{Y}} R(c|x)$$

贝叶斯判断准则：对每个样本，都选择能使后验概率最大的类标记。

4.2 线性判别

线性判别的原理是将高纬度空间中的数据投影到低维度空间后能进行简单的划分。投影后的数据需最大程度满足，类间差距大，类内差距小，以该原则为基础，构造优化函数。

4.3 生成式模型和判别式模型

生成式模型：由数据学习联合概率分布 $P(X,Y)$，然后由 $P(Y|X) = \frac{P(X,Y)}{P(X)}$ 作为预测的模型。该方法表示了给定输入与产生输出之间的生成关系。典型的有朴素贝叶斯、k-Means 聚类算法和混合高斯模型（GMM）。

判别式模型：由数据直接学习决策函数 $Y = f(X)$ 或条件概率分布 $P(Y|X)$ 作为预测模型，即判别模型。判别方法关系的是对于给定的输入，应该预测怎样的输出。典型的有决策树、线性回归 (Linear Regression)、K 近邻 (KNN)、逻辑回归 (Logistic Regression) 和支持向量机 (SVM)。

4.4 KNN

k 近邻（k-Nearest Neighbor）学习是一种监督、无参数、生成式模型。算法步骤如下：

1. 计算测试数据与各个训练数据之间的距离；

2. 按照距离的递增关系进行排序；

3. 选取距离最小的 K 个点；

4. 确定前 K 个点所在类别的出现频率；

5. 返回前 K 个点中出现频率最高的类别作为测试数据的预测分类。

KNN 算法是一种典型的懒惰学习（lazy learning），即对输入的数据集不做处理，训练开销为 0，当收到测试数据时再进行处理。

KNN 虽然简单，但其泛化错误率不超过贝叶斯最优分类器的两倍。

5 朴素贝叶斯以及逻辑回归

5.1 条件独立与朴素贝叶斯

基于贝叶斯公式来估计后验概率的主要困难在于类条件概率 $P(x|c)$ 是所有属性上的联合概率，难以从有限样本集上直接估计。为了解决这个问题，简化运算，朴素贝叶斯（Naive Bayes）分类器采用了属性条件独立性假设，即假设对于一直的类别，所有属性之间相互独立。换言之，假设每个属性独立地对分类结果发生影响。朴素贝叶斯的判断准则为：

$$h_{nb}(x) = \arg\min_{c\in\mathcal{Y}} P(c)\prod_{i=1}^{d} P(x_i|c)$$

公式中的两个概率都可以通过统计样本数据获取。

5.2 高斯朴素贝叶斯

即假设每个维度上的分布为高斯分布的情况。对连续的值可以考虑概率密度函数，即 $p(x_i|c)\ \mathcal{N}(\mu_{c,i}, \sigma^2_{c,i})$

5.3 逻辑回归

如何直接利用之前学习过的线性回归问题解决分类问题呢。只需要找到一个单调可微函数作为激活函数，将回归得出的数值映射到离散的数值上即可。以二分类问题为例，我们选取 Sigmoid 函数作为激活函数。

逻辑回归又称为对数几率回归，若将 y 视作样本 x 为正例的概率，1-y 视作样本 x 为反例的概率，则可将 $\frac{y}{1-y}$ 称为几率，反映了 x 作为正例的相对可能性，对其取对数后称为对数几率。逻辑回归是一个线性模型。

6 SVM 与核方法

支持向量机（Support Vector Machine, SVM）是一类按监督学习（supervised learning）方式对数据进行二元分类的广义线性分类器（generalized linear classifier），其决策边界是对学习样本求解的最大边距超平面（maximum-margin hyperplane）。

SVM 使用铰链损失函数（hinge loss）计算经验风险（empirical risk）并在求解系统中加入了正则化项以优化结构风险（structural risk），是一个具有稀疏性和稳健性的分类器。SVM 可以通过核方法（kernel method）进行非线性分类，是常见的核学习（kernel learning）方法之一。

6.1 最大间隔和过拟合

SVM 试图找出一个最优的划分平面以对样本空间进行划分，划分超平面可通过 $w^T x + b = 0$ 来描述。在样本空间中距离超平面最近的几个训练样本点被称为支持向量，两个异类支持向量到超平面的距离之和称为间隔（margin）。支持向量机的原理即为最大化间隔。

6.2 原始对偶求解

SVM 中存在强对偶性。min max = max min。

6.3 内积与核函数

在 SVM 的最终优化目标中，存在 $x_i^T x_j$ 的内积形式，由此引出了向量内积的一个含义，即表明了两个向量之间的相似度。

由于在低维空间中数据可能线性不可分，需要我们进行升维到高维空间，再求其内积。但升维后再求内积的过程较为繁琐，我们可以使用核函数直接模拟升维后的内积结果，而不需要进行显式的升维处理。

6.4 线性可分与特征变换

由 Cover 定理，低维空间线性不可分的数据在高维空间中线性可分。

7 无监督学习

7.1 层次聚类与相似度函数

层次聚类假设类别之间存在层次结构，将样本聚到层次化的类中。层次聚类有聚合（自下而上）和分裂（自上而下）两种方法。判断两个类之间的距离的方式就是相似度函数，也可以称为距离函数。常见的有闵可夫斯基距离、马氏距离、相关系数、夹角余弦等等。

7.2 K-Means 聚类

一种基于样本集合划分的聚类算法。通过迭代的方式进行求解，每次迭代有以下两个步骤：

1. 确定 k 个类的中心，将样本逐个指派到与其最近的中心的类中；
2. 更新每个类的样本的均值，作为新的类的中心。

7.3 GMM 及 EM 算法

将样本空间视作多个高斯分布的加权叠加。其中引入了一个隐变量 z_i 表示样本属于哪一个高斯分布。

7.4 PCA

在最大可分性和最近重构性的原则上对高维空间样本进行降维。