统计机器学习复习提纲

一、综述

1. 统计学vs机器学习

研究方法

- 统计学：研究**形式化和推导**
- 机器学习：容忍一些**新方法**

维度差异

- 统计学：研究**低维空间**的统计推导
- 机器学习：研究**高维空间**的预测问题

领域差异

- 统计学：生存分析、空间分析、多重检验、极大极小理论、反卷积、半参数推理
- 机器学习：在线学习、监督学习、非监督学习、强化学习....

2. 统计学习的特点（不是统计学）

- 以**计算机和网络**为平台，构建在计算机和网络上的
- 以**数据**为研究对象，是数据驱动学科
- 目的是对数据进行**预测、分析**
- 以**方法**为中心，统计学习方法通过构建模型和应用模型对数据进行预测和分析
- 统计学习是概率论、统计学、计算理论、信息论、最优化理论以及计算机科学等多个学科构建的交叉学科，已经逐渐形成独立的理论体系

3. 统计学习的学习对象

- 数据，包括计算机及网络上的各种数字、文字、图像、视频、音频数据以及他们的组合
- 数据的**基本假设**就是同类数据具有一定的统计规律性

4. 学习目的

- 对数据（特别是未知数据）进行预测与分析
- 对数据的预测可以是计算机更加智能，或者说某些性能得到强化

5. 分类（大致的）

- 监督学习
- 非监督学习
- 强化学习

6. 监督学习

定义

- 是从**标记数据**中学习预测模型的机器学习问题，学习输入到输出的统计规律

基本假设：联合概率分布

- 对数据的基本假设：同类数据具有一定的统计规律
- 输入X和输出Y遵循联合概率分布$P(X,Y)$，对学习系统是未知的
- 训练数据和测试数据**独立同分布**于$P(X,Y)$

目的

- 学习一个从**输入到输出**的映射

模型的集合

- 就是假设空间：满足上述假设的映射都行

分类

- 概率模型：学条件概率分布$P(Y|X)$
- 非概率模型：直接学决策函数$Y = f(X)$

一般化流程！！！！！（2023期末简答第一题：什么是机器学习，以监督学习为例，简述一般化流程）

- 得到有限的**训练数据**集合
- 确定**假设空间**
- 确定模型选取**策略**
- 找到合适的最优模型求解方法，即学习的**算法**
- 通过学习算法选取**最优模型**
- 利用学习到的最优模型对未知数据进行**预测分析**

7. 无监督学习

定义

- 从**无标注**的数据中学习预测模型的机器学习问题，本质是学习数据中的规律和潜在结构

8. 强化学习

定义

- 指通过**智能系统**和**环境**的**连续互动中**学习最优化行动策略的机器学习问题。

目标

- 不是短期激励最大化，而是使得**长期累积激励最大化**，强化学习过程是让机器不断试错，达到学习最优策略的目的

本质

- 学习**最优序贯策略**

9. 半监督学习

定义

- 利用**标注数据**和**未标注数据**学习预测模型的机器学习问题

目的

- 利用**未标注数据**的信息，**辅助标注数据**进行监督学习，以较低的标注成本达到更好的效果

是大数据时代的发展趋势

10. 主动学习

定义

- 模型**不断给出实例**给教师进行标注，然后利用标注数据学习预测模型的机器学习问题

目的

- 找出对问题最有用的数据给教师进行标注，以较少的标注代价得到较好的结果

11. 统计学习方法=模型+策略+算法

模型：给定数据集和任务，如何选择模型，即假设空间

- 概率模型：条件概率分布函数
- 非概率模型：决策函数

策略：什么模型才是好的，即如何评价一个假设

- 损失函数：**一次**预测的好坏

- 风险函数：**平均意义下**的模型预测好坏
- 经验风险函数：因为风险函数需要求期望，但是样本概率分布函数我们实际上并不知道，只能根据**大数定理**，对多次预测的损失函数取平均来近似风险函数
- 结构风险函数：考虑其他因素的影响，如模型复杂度，此时加入**正则化**等操作

算法：如何以最快的搜索速度，找到最优的假设

- 最小二乘法：仅针对**线性**模型
- 梯度下降、上升法（批梯度、增量梯度）：**任何**模型

12. 模型评估与模型选择

- 训练误差
- 测试误差
- ***过拟合！！！！！！！！（2023期末简答第三题：什么是过拟合现象，导致过拟合的原因，缓解过拟合的方法）***
- 正则化
- ***交叉验证！！！！！！！（2023期末简答第五题：作用及操作说明）***
  - 简单交叉验证
  - k折交叉验证
  - 留一交叉验证
- 泛化能力：该方法学习到的模型对未知数据的预测能力
- 泛化误差：学习到的模型对未知数据预测的误差期望
- 泛化误差上界

二、感知机

1. 损失函数选取

- 误分类样本数，即示性函数之和对w、b不可导
- 样本到$y=wx+b$的距离可导

2. 计算！！！！！！！！（2023期末计算第一题）

三、决策树

1. 熵H(D)、条件熵H(D|A)、信息增益g(D, A)、信息增益比gR(D, A)

- ID3(信息增益), C4.5(信息增益比)

2. 计算！！！！！！！（2023期末计算第三题，只会纯按计算器算的时间很久，建议学计算器的变量功能）

3. 实际应用注意事项

- 数据清理
- 数据转化
  - 数据归一化
  - 数据归类：比如连续型数据通过定义区间归类
  - 类别限制：一个特征的取值**不超过7个**（最好不超过5个）
- 相关性分析
  - 对于问题无关的属性：删除
  - 对于取值超过7种而且不能归纳的：删除

4. gini系数、剪枝、CART算法（要求貌似不高）

- 树的损失函数、评估剪枝前后整体损失下降程度的指标$g(t)$
  - 树的损失函数$C_\alpha\left(T_t\right)=C\left(T_t\right)+\alpha|T|$
  - 公式中的$C(T_t)$是节点t对应子树的分类结果的“不纯度”，$C(t)$则是剪枝后只剩下该节点后的分类结果的“不纯度”；$|T|$是剪枝前节点t对应子树的节点个数（不包含该根节点）
    - “不纯度”越低，模型分类结果越好；而一般的剪枝操作会带来“不纯度”的上升，损失会增大

        - 而$|T|$是考虑了树的复杂度对损失函数的权衡，是惩罚项（正则化的一种），剪枝之后会导致复杂度下降，损失函数会减小
    - 实际过程中很难直接取到合适的权重，去判断“不纯度”与树的复杂度对我们结果的影响程度
        - 不能是单纯设$\alpha=1$，剪枝后损失降低，就说我们的结果会更好，这是没有逻辑的，所以设计了一个很巧妙的办法进行剪枝
    - 剪枝算法内函数及步骤含义
        - 损失函数意义
            - 剪枝前$C_\alpha\left(T_t\right)=C\left(T_t\right)+\alpha|T|$
            - 剪枝后$C_\alpha(t)=C(t)+\alpha$
            - 当$\alpha=0$时，刚刚说了剪枝后“不纯度”会上升，所以$C_0\left(T_t\right) \leqslant C_0(t)$
            - 而当α很大时，模型复杂度对损失函数影响极大，会有$C_\alpha\left(T_t\right)>C_\alpha(t)$
        - $g(t)$意义
            - 损失函数是α的连续函数，中间肯定有一点是两者相等的，这一点我们记作$g(t)=\frac{C(t)-C\left(T_t\right)}{|T|-1}$，是个正数
            - 我们知道$g(t)$是每个节点剪枝前后损失函数不变的α临界值，**如果$g(t)$很低，即给复杂度权重很低的情况下，剪枝能让损失减少，（$C(T_t)$没多大，反而$|T|$相对很大）,证明这个节点很冗余，应该剪掉**
        - 产生$\{T_0, T_1 \ldots T_k\}$后再进行交叉验证
            - 每次选择当前树T_i中$g_i(t)$最小的节点进行剪枝，记$\alpha_i=\min\limits_t g_i(t)$并把剪枝后的树$T_{i+1}$用来下一次迭代剪枝
                - 剪枝算法原论文里有证明：$\alpha_i \leqslant \alpha_{i+1}$
                - 这样会得到k棵α在区间$[\alpha_i, \alpha_{i+1})$的局部最优子树，因为每棵树都是上一棵树基础上剪去了“最冗余”的子树得到的，即**从T_0每次都修剪掉最冗余的子树得到后续新的子树**
            - 但是**我们并不知道修剪掉“冗余子树”对结果（分类/预测）前后的实际影响，也不知道修剪多少“冗余子树”对结果最好**，所以需要采用交叉验证对上述局部最优子树进行选取，得到$\alpha \in [\alpha_1, \alpha_k)$最优树

四、k近邻算法

1. *k的含义（2023期末简答第二题：knn和kmeans的异同，knn、kd树、kmeans的k是什么意思）*

2. k的选择

- k越小代表模型越复杂，要找到最相似的少数样本点投票，$k=1$时只有一个样本点；k很大的时候模型会找很多个样本点，$k=$样本数，预测结果恒为样本中较多的类，比较简单；
- k的选择一般为奇数，避免偶数导致的“平票”的情况
- 选择最优的k：交叉验证，选验证集平均准确率最高
- 确认了k之后的经验风险最小原则（误分类率最小），等价于多数表决原则

3. kd树构建

- 选中位数的时候，如果总数是偶数，则选中间两个数的后一个数作为中位数（不取平均，这样可以保留一个样本在当前节点）

4. kd树搜索（要求不高，k>2的kd树搜索不作要求）

五、SVM

1. 什么样的超平面冗错性最好

- 样本点距离超平面最近距离最大的超平面

2. 支持向量、正负平面距离（即SVM要最大化的函数）

3. 函数间隔$y_i(w * x_i+b)$

- 函数间隔是从样本到超平面距离演化来的，$\frac{w * x i+b}{||w||}$是点到超平面的距离
- 用距离d刻画Margin间距：$\frac{w * x_i+b}{||w||} \geqslant d\,(\leqslant -d)$，来刻画点在超平面上方还是下方，并且距离$d$尽可能取大，越大则正负超平面距离($2d$)越大，这是SVM的目标“最大间距”的来源
- 而我们任何时候都可以通过对w, b放缩保证$||w|| d=1$
    - $\frac{w x_i+b}{||w||} \geqslant d\,(\leqslant -d)$转化为$w x_i+b-||w|| d>=0$
    - 放缩$\frac{w}{||w|| d} x_i+\frac{b}{||w|| d}-1>=0$
    - 令$w^{\prime}=\frac{w}{||w|| d}$，$b^{\prime}=\frac{b}{||w|| d}$，则式子变为$w^{\prime} x_i+b^{\prime}>1$
    - 最后求出来的超平面方程$w^{\prime} x+b^{\prime}=0$和原本的$w x+b=0$是等价的

- 所以上述放缩相当于直接令$||w||d=1$，后续只需要控制$wx_i+b\geqslant 1(\leqslant -1)$，即$y_i(wx_i+b)\geqslant 1$，此时前面这个数值$y_i(wx_i+b)$即称为“函数间隔”。
- 放缩完之后再看回我们要求的间距$d=\frac{1}{||w||}$，正负平面距离为$2d=\frac{2}{||w||}$，就是教材上写的要最大化的目标函数（s.t. $y_i(wx_i+b)\geqslant 1$，即保证所有类都分对了）

4. 凸集

- 连接集合中的两个点，线段包含在集合中，则称这个集合为凸集，否则为非凸集

5. 对偶问题、KKT条件、求解对偶形式求解

6. 硬间隔、软间隔

- 硬间隔的支持向量机w^*，b^*均唯一
- 软间隔的w^*唯一，b^*不唯一（即惩罚项可以调节b^*，分离超平面可以容许上下平移）

7. 核函数！！！！（2023期末简答第三题：什么是核函数，作用是什么，常见的核函数有什么）

定义！！！

- 存在原空间到特征空间的映射$\phi(x)$，定义一个函数$K(x,z)=\phi(x)\cdot\phi(z)$(内积)，则称$K(x,z)$为核函数，$\phi(x)$为对应的映射函数
- 在应用过程中一般只显式定义$K(x,z)$而不去定义$\phi(x)$来求内积
- $\phi(x)$：是输入空间R^n到特征空间H的映射，特征空间一般是高维的，一个核函数可以由不同$\phi(x)$定义

常见类别！！！！

- 正定核
- 高斯核：特征很少，但是样本数量不多也不少
- 多项式核
- 线性核：特征很多，与样本数量差不多（如果特征数量很少，样本数量很多，需要额外添加特征来用线性核）
- sigmoid核

六、朴素贝叶斯

1. 贝叶斯网络

- 概率图模型、有向无环图

2. 贝叶斯定理

- 后验概率$P(A_i|B)=P(Ai)\frac{P(B|A_i)}{\sum_{i=1}^{m}P(B|A_i)P(A_i)}$=先验概率*调整因子
  - $P(A_i|B)$为后验概率
  - $P(Ai)$为先验概率
  - $\frac{P(B|A_i)}{\sum_{i=1}^{m}P(B|A_i)P(A_i)}$为调整因子
- 但事实上P(Ai)是未知的，即分类任务时，标签的真实概率分布是不知道的，贝叶斯算法应用时常将其假设为正态分布、beta分布或者泊松分布，这并没有特别的依据，所以很多人不承认这个算法，但是效果很好、计算方便，所以广为流传

3. 朴素贝叶斯

- 基于贝叶斯定理和特征条件独立假设提出的算法
- ***特征条件独立假设，即条件概率相互独立，是“朴素”的来源！！！！！！！（2022期末填空考到）***
- 后验概率最大化 ⇔ 期望风险最小化
- 先验概率按照极大似然估计结果，即频率

4. 拉普拉斯平滑计算！！！！（2023期末计算题第二题考了没有拉普拉斯平滑的正常贝叶斯）

- 需要注意的是，调整因子的计算和先验概率的计算都要加上拉普拉斯平滑

七、逻辑回归

1. 假设

- 假设X是连续变量而且服从逻辑分布
    - 概率分布函数$F(x)=\frac{1}{1+exp(-(x-\mu)/\gamma)}$
    - 概率密度函数$f(x)=\frac{exp(\frac{-(x-\mu)}{\gamma})}{\gamma\left[1+exp(\frac{-(x-\mu)}{\gamma})\right]}$
    - 密度函数关于$(\mu,\frac{1}{2})$中心对称
- sigmoid函数
    - 上述$\mu=0$，$\gamma=1$，即$f(x)=\frac{1}{1+exp(-x)}$
    - $f'(x)=f(x)(1-f(x))$

2. 事件发生的几率(odds)

- 逻辑回归的事件事件发生的对数几率是$log_2[\frac{P(1|x)}{1-P(1|x)}]=wx+b=w'x'$

3. 逻辑回归模型

- 似然函数、$w'=(w_1,w_2...w_n,b)$的极大似然估计

4. 最大熵模型

原理

- 我们认为，在满足一切约束条件的情况后，剩下的模型熵最大的是最好的

熵的计算:$H(P)=\sum_x P(x)log_2P(x)$

- 有$0\leqslant H(P)\leqslant log_2|X|$，$|X|$为$X$的取值数量
    - 当每个x取值概率相等时，右侧等号成立
    - 即原理中，满足一切约束条件后，剩余的取值尽量相等的模型是最好的

5. 推导（没什么要求）

八、聚类

1. 相似度

- 闵可夫斯基距离
- 余弦相似度
- 马哈拉诺比斯距离
- 相关系数

2. 相似度影响因素

- 不同指标的量纲
- 特征之间的相似性
- 样本分布

3. 软聚类、硬聚类

4. 类或簇的定义（四种定义）

- 类的特征
- 类均值
- 类直径
- 样本散布矩阵和样本协方差矩阵
    - 样本散布矩阵$A_G=\sum_{i=1}^{n_G}(x_i-\bar{x})(x_i-\bar{x})^T$
    - 样本协方差矩阵$S_G=\frac{AG}{m-1}$，m是x的维数

5. 类的连接

- 最短距离
- 最长距离
- 中心距离
- 平均距离

6. 层次聚类

- 属于硬聚类
- **聚合聚类（自下而上聚类）**
  - 三个要素
    - 距离或相似度
      - 闵可夫斯基距离
      - 马哈拉诺比斯距离
      - 余弦相似度
      - 相关系数
    - 合并规则
      - 类间距离最小
      - 类间距离可以是：最短距离、最长距离、中心距离、平均距离
    - 停止规则
      - 个数达到阈值k
      - 类直径超过阈值
  - 时间复杂度$O(n^3m)$
- 分裂聚类（自上而下聚类）应用较少

7. kmeans聚类

- ***算法步骤！！！！！！！！！！！(2023期末计算第四题)***
- 总体特点
  - 基于划分的聚类方法
  - 类别数k需要预先指定
  - 以欧氏距离表示距离，以中心或样本均值表示类别
  - 以样本与其所属的类别中心的距离和作为优化对象
  - 得到的类别是平坦的、非层次化的
  - 算法是迭代计算、不能保证得到全局最优解
- 收敛性
  - k均值属于启发式算法，不能保证收敛到全局最优解
  - k均值中心的移动不会太大，因为每次样本归属于距离最近的类
- 初始类的选择
  - 类中心初始化不同，最后聚类的结果不同
  - 可以先用层次聚类把样本聚成k个类，用k个类的中心作为初始化类中心
- 类别数k的选择
  - k需要预先给定，实际应用中最优的k是不知道的
  - 可以尝试不同的k，比较聚类结果质量推测最优的k
  - 聚类结果质量可以用类直径
  - 一般的，类别数变小，类直径变大
- 选择方法
  - 拐点法
    - 簇内平方和的拐点就是最佳分类点
  - 临界平均直径
    - 类别数增大到一定程度，平均直径会不变，这个值就是临界平均直径，对应的k就是最佳类别数
- 优缺点
  - 优点
    - 当k变大时，k均值计算会比层次聚类快
    - 与层次聚类相比，kmeans聚类结果更加紧凑，尤其是球状簇
    - 大数据集合效率更高
    - 当结果簇是紧密的时候，簇与簇之间分隔比较开
  - 缺点
    - 没有指明初始化方式，常用方法只有随机选k个样本作为初始类中心

- 初始化不同，会导致多种次优结果，解决方法是多尝试几种不同的初始中心
- 会出现距离类中心mj最近的样本集合为空的情况，因此mj得不到更新
- 不适合非凸面的簇，且对噪声和离群点十分敏感，因为少量这类点对均值影响也是很大的

提纲整理：21数学卢锦鹏

（下面问题好奇的话可以直接问老师或者讨论解决）

笔者复习时的问题积累：

1. 为什么样本量少的情况下训练效果不理想?

2. 对数似然损失函数$loss_{log} = -log_2 P(Y|X)$为什么要取对数?

3. CART选最优子树的时候，用平方误差和gini系数选最优子树的意义？只能用这两个？（交叉熵和方差？）

4. CART剪枝的时候，T0剪去了第一个子树变T1，之后的剪枝是在T1基础上继续剪?T1上的节点（特别是父节点）会不会再出现$g(t)$比上一个剪枝的α_1还小的情况?

5. $\max\limits_{\alpha \geqslant 0, \beta} \min\limits_{x} f(x) \leqslant \min\limits_{x} \max\limits_{\alpha \geqslant 0, \beta} f(x)$?

6. b^*为什么唯一（证明对偶问题不同$\alpha_j^* > 0$得出的b^*都是一样的吗）？

7. 证明$0 \leqslant H(P) \leqslant log_2 |X|$?

8. kmeans中除了l2范数，其余相似度的中心计算怎么算（有解析方法吗，l1范数是中位数）？

9. 什么时候会出现“可能发生距离簇中心m_j最近的样本集为空的情况，因此，m_j将得不到更新”?

10. kd树搜索，怎么从树上面直接判断所谓的“另一个子树范围和超球面相交”