机器学习导论 习题二

学号, 姓名, 邮箱

2023 年 4 月 13 日

作业提交注意事项

1. 请在 LaTeX 模板中第一页填写个人的学号、姓名、邮箱;
2. 本次作业需提交作答后的该 pdf 文件, **请将其打包为 .zip 文件上传**. 注意命名规则, 两个文件均命名为"学号_姓名"+". 后缀"(例如 211300001_张三"+".pdf"、".zip");
3. 若多次提交作业, 则在命名 .zip 文件时加上版本号, 例如 211300001_ 张三 _v1.zip"(批改时以版本号最高的文件为准);
4. 本次作业提交截止时间为 **4 月 19 日 23:59:59**. 未按照要求提交作业, 提交作业格式不正确, **作业命名不规范**, 将会被扣除部分作业分数; 除特殊原因 (如因病缓交, 需出示医院假条) 逾期未交作业, 本次作业记 0 分; **如发现抄袭, 抄袭和被抄袭双方成绩全部取消**;
5. 本次作业提交地址为 here, 请大家预留时间提前上交, 以防在临近截止日期时, 因网络等原因无法按时提交作业.

1 [20pts] Linear Discriminant Analysis

线性判别分析 (Linear Discriminant Analysis, 简称 LDA) 是一种经典的线性学习方法. 请仔细阅读《机器学习》第三章 3.4 节, 并回答如下问题.

(1) **[10pts]** (二分类) 假设有两类数据, 其中正类服从高斯分布 $P=\mathcal{N}(\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_1)$, 负类服从高斯分布 $Q=\mathcal{N}(\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_2)$. 对于任一样本 $\boldsymbol{x}$, 若分类器 h 满足:

$$h(\boldsymbol{x})=\begin{cases}0 & P(\boldsymbol{x})\leq Q(\boldsymbol{x}),\\ 1 & P(\boldsymbol{x})>Q(\boldsymbol{x}),\end{cases}$$

则认为 h 实现了最优分类. 假设 $\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2$ 均已知, 请证明当 $\boldsymbol{\Sigma}_1=\boldsymbol{\Sigma}_2=\boldsymbol{\Sigma}$ 时, 通过 LDA 得到的分类器可实现最优分类. (提示: 找到满足最优分类性质的分类平面)

(2) **[10pts]** (多分类) 将 LDA 推广至多分类任务时, 可采用教材中式 (3.44) 作为优化目标. 通过求解式 (3.44), 可得到投影矩阵 $\mathbf{W}\in\mathbb{R}^{d\times d'}$, 其中 d 为数据原有的属性数. 假设当前任务共有 N 个类别, 请证明 $d'\leq N-1$. (提示: 对于任意 n 阶方阵, 其非零特征值个数小于等于其秩大小)

Solution. 此处用于写解答 (中英文均可)

(1) 考虑分类器 h 的分类平面 $P(\boldsymbol{x})=Q(\boldsymbol{x})$, 可以得到:

$$(\boldsymbol{x}-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)=(\boldsymbol{x}-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2),$$

化简后可得到如下分类平面:

$$(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}^{-1}\left(2\boldsymbol{x}-(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\right)=0.$$

另外, 对于已知 $\boldsymbol{\mu}_1,\boldsymbol{\mu}_2,\boldsymbol{\Sigma}_1,\boldsymbol{\Sigma}_2$ 的二分类任务, LDA 方法对应的分类平面为:

$$2\boldsymbol{w}^\top\boldsymbol{x}=\boldsymbol{w}^\top(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2),$$

代入 $\boldsymbol{w}=\boldsymbol{S}_w^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)=2\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$, 可以得到:

$$(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)^\top\boldsymbol{\Sigma}^{-1}\left(2\boldsymbol{x}-(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)\right)=0.$$

两者分类平面相同, 因此通过 LDA 得到的分类器实现了最优分类.

(2) 由于 $\mathrm{rank}(\mathbf{S}_w^{-1}\mathbf{S}_b)\leq\mathrm{rank}(\mathbf{S}_b)$, 现只需证明 $\mathrm{rank}(\mathbf{S}_b)\leq N-1$. 对 $\mathbf{S}_b$ 进行分解:

$$\begin{aligned}\mathbf{S}_b&=\sum_{i=1}^{N}m_i\left(\boldsymbol{\mu}_i-\boldsymbol{\mu}\right)\left(\boldsymbol{\mu}_i-\boldsymbol{\mu}\right)^\top\\&=\underbrace{\left(\sqrt{m_1}\left(\boldsymbol{\mu}_i-\boldsymbol{\mu}\right),\sqrt{m_2}\left(\boldsymbol{\mu}_2-\boldsymbol{\mu}\right),\ldots,\sqrt{m_N}\left(\boldsymbol{\mu}_N-\boldsymbol{\mu}\right)\right)}_{\mathbf{H}}\begin{pmatrix}\sqrt{m_1}\left(\boldsymbol{\mu}_i-\boldsymbol{\mu}\right)\\\sqrt{m_2}\left(\boldsymbol{\mu}_2-\boldsymbol{\mu}\right)\\\vdots\\\sqrt{m_N}\left(\boldsymbol{\mu}_N-\boldsymbol{\mu}\right)\end{pmatrix},\end{aligned}$$

即 $\mathbf{S}_b=\mathbf{H}\mathbf{H}^\top$, $\mathrm{rank}(\mathbf{S}_b)=\mathrm{rank}(\mathbf{H})$. 由于 $\mathbf{H}$ 中所有列可通过线性组合得到零向量, 因此 $\mathrm{rank}(\mathbf{S}_w^{-1}\mathbf{S}_b)\leq\mathrm{rank}(\mathbf{H})\leq N-1$, 即 $d'\leq N-1$.

2 [20pts] Multi-Class Learning

现实场景中我们经常会遇到多分类任务, 处理思路主要分为两种: 一是利用一些基本策略 (OvO,OvR,MvM), 将多分类任务拆分为若干个二分类任务; 二是直接求解, 将常见的二分类学习器推广为多分类学习器. 请仔细阅读《机器学习》第三章 3.5 节, 并回答如下问题.

(1) **[5pts]** 考虑如下多分类学习问题: 样本数量为 n, 类别数量为 K, 每个类别的样本数量一致. 假设一个二分类算法对于大小为 m 的数据训练的时间复杂度为 $\mathcal{O}(m^\alpha)$, 试分别计算该算法在 OvO、OvR 策略下训练的总体时间复杂度.

(2) **[5pts]** 当我们使用 MvM 处理多分类问题时, 正、反类的构造需要有特殊的设计, 一种最常用的技术是"纠错输出码"(ECOC). 考虑 ECOC 中的编码矩阵为"三元码"的形式, 即在正、反类之外加入了"停用类". 请通过构造具体的编码矩阵, 说明 OvO、OvR 均为此 ECOC 的特例.

(3) **[10pts]** 对数几率回归 (logistic regression) 是一种常用的二分类模型, 简称对率回归. 现如今问题由二分类推广至多分类, 其中共有 K 个类别即 $y \in \{1, 2, \cdots, K\}$. 基于使用线性模型拟合对数几率这一思路, 请将对数几率回归算法拓展至多分类任务, 给出该多分类对率回归模型的"对数似然", 并给出该"对数似然"的梯度.

提示 1: 考虑如下 $K-1$ 个对数几率, 分别用 $K-1$ 组线性模型进行预测,

$$\ln \frac{p(y=1 \mid \mathbf{x})}{p(y=K \mid \mathbf{x})}, \ln \frac{p(y=2 \mid \mathbf{x})}{p(y=K \mid \mathbf{x})}, \cdots, \ln \frac{p(y=K-1 \mid \mathbf{x})}{p(y=K \mid \mathbf{x})}$$

提示 2: 定义指示函数 $\mathbb{I}(\cdot)$ 使得答案简洁,

$$\mathbb{I}(y=j) = \begin{cases} 0 & \text{若 } y \text{ 不等于 } j \\ 1 & \text{若 } y \text{ 等于 } j \end{cases}$$

Solution. 此处用于写解答 (中英文均可)

(1) 由于 K 个类别样本数量相同, 故每个类别均有 $\frac{n}{K}$ 个训练样本.

对于 OvO 策略, 总体训练时间复杂度为

$$\frac{K(K-1)}{2} \cdot \mathcal{O}\left(\left(\frac{2n}{K}\right)^\alpha\right) = \mathcal{O}\left(K^2 \left(\frac{n}{K}\right)^\alpha\right) = \mathcal{O}(K^{2-\alpha} n^\alpha);$$

对于 OvR 策略, 总体训练时间复杂度为

$$K\mathcal{O}\left(n^\alpha\right) = \mathcal{O}\left(K n^\alpha\right).$$

(2) 在使用三元码的 ECOC 编码矩阵中, (i, j) 元素取值为"+1"、"-1"分别表示学习器 f_i 将第 j 类样本作为正、反例; 取值为"0"表示学习器 f_i 训练时不使用第 j 类样本. 该矩阵的列数对应于多分类方法中训练分类器的个数.

OvO 策略每次将一个类别作为正类、一个类别作为负类, 故对应的 ECOC 编码矩阵共有 $c = \frac{K(K-1)}{2}$ 列. 在该矩阵中, 每一列都对应一组类别标记 $(l, l'), l \neq l'$, 第 l 行取

值为 $+1$、第 l' 行取值为 -1, 其余所有行的取值均为 0. 编码矩阵的各个列对应的类别标记对应互不相同.

OvR 策略每次将一个类别作为正类, 其余所有类别作为负类, 故对应的 ECOC 编码矩阵共有 $c=K$ 列. 该矩阵除对角线元素取值为 $+1$ 外, 其余元素的取值均为 -1.

(3) 设 $\hat{\mathbf{x}}=(\mathbf{x};1)\in\mathbb{R}^{d+1}$, 根据提示, 将多分类问题的 $K-1$ 个对数几率表示为如下形式:

$$
\begin{aligned}
\ln\frac{P(y=1\mid\mathbf{x})}{P(y=K\mid\mathbf{x})}&=\mathbf{w}_1^\top\mathbf{x}+b_1=\boldsymbol{\beta}_1^\top\hat{\mathbf{x}}\\
\ln\frac{P(y=2\mid\mathbf{x})}{P(y=K\mid\mathbf{x})}&=\mathbf{w}_2^\top\mathbf{x}+b_2=\boldsymbol{\beta}_2^\top\hat{\mathbf{x}}\\
&\cdots\\
\ln\frac{P(y=K-1\mid\mathbf{x})}{P(y=K\mid\mathbf{x})}&=\mathbf{w}_{K-1}^\top\mathbf{x}+b_{K-1}=\boldsymbol{\beta}_{K-1}^\top\hat{\mathbf{x}}.
\end{aligned}
$$

其中 $\boldsymbol{\beta}_n=(\mathbf{w}_n;b_n)$ 将偏置项 b_n 与分类器权重合并. 由于 $\sum_{i=1}^K P(y=i\mid\mathbf{x})=1$, 可得

$$
p(y=k\mid\mathbf{x})=\begin{cases}\dfrac{e^{\boldsymbol{\beta}_k^T\hat{\mathbf{x}}}}{1+\sum_{i=1}^{K-1}e^{\boldsymbol{\beta}_i^T\hat{\mathbf{x}}}}, \text{ if } k\le K-1\\ \dfrac{1}{1+\sum_{i=1}^{K-1}e^{\boldsymbol{\beta}_i^T\hat{\mathbf{x}}}}, \text{ if } k=K.\end{cases}
$$

由此可以得到似然函数

$$
\prod_{i=1}^n P(y_i\mid\hat{\mathbf{x}}_i;\boldsymbol{\beta})=\prod_{i=1}^n\prod_{j=1}^K\mathbb{I}(y_i=j)P(y_i=j\mid\hat{\mathbf{x}}_i;\boldsymbol{\beta}),
$$

以及对率回归模型的对数似然

$$
\begin{aligned}
\ell(\boldsymbol{\beta})&=\sum_{i=1}^n\sum_{j=1}^K\mathbb{I}(y_i=j)\ln P(y_i=j\mid\hat{\mathbf{x}}_i;\boldsymbol{\beta})\\
&=\sum_{i=1}^n\sum_{j=1}^{K-1}\mathbb{I}(y_i=j)\boldsymbol{\beta}_j^T\hat{\mathbf{x}}_i-\sum_{i=1}^n\ln\left(1+\sum_{i=1}^{K-1}e^{\boldsymbol{\beta}_i^T\hat{\mathbf{x}}_i}\right),
\end{aligned}
$$

从而计算出对数似然关于第 n 类分类器 $\boldsymbol{\beta}_n$ 的梯度如下,

$$
\begin{aligned}
\frac{\partial\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}_n}&=\sum_{i=1}^n\mathbb{I}(y_i=n)\hat{\mathbf{x}}_i-\sum_{i=1}^n\frac{e^{\boldsymbol{\beta}_n^T\hat{\mathbf{x}}_i}\cdot\hat{\mathbf{x}}_i}{1+\sum_{i=1}^{K-1}e^{\boldsymbol{\beta}_i^T\hat{\mathbf{x}}_i}}\\
&=\sum_{i=1}^n\left(\mathbb{I}(y_i=n)-P(y_i=n\mid\hat{\mathbf{x}}_i)\right)\hat{\mathbf{x}}_i,\quad n=1,2,\cdots,K-1.
\end{aligned}
$$

3 [20pts] Decision Tree Analysis

决策树在实际应用中的性能虽然不及深度神经网络等复杂模型, 但其可以作为弱学习器, 在强大的集成算法如 XGBoost 中发挥重要的作用. 假设分类问题中标记空间 $\mathcal{Y}$ 的大小为 $|\mathcal{Y}|$, 训练集 D 中第 k 类样本所占比例为 $p_k(k = 1, 2, \cdots, |\mathcal{Y}|)$, 请仔细阅读《机器学习》第四章, 并回答如下问题.

(1) **[5pts]** 给定离散随机变量 X 和 Y, 随机变量的信息熵 $H(X)$ 定义如下:

$$H(X) = -\sum_x P(x)\log_2 P(x),$$

衡量了 X 的不确定性. 条件熵 (conditional entropy)$H(Y|X)$ 定义如下:

$$H(Y \mid X) = \sum_x P(x)H(Y \mid X = x) = -\sum_x P(x)\sum_y P(y \mid x)\log_2 P(y \mid x),$$

衡量了 Y 中不依赖 X 的信息量;X 和 Y 的互信息 (mutual information) 定义如下:

$$I(X;Y) = \sum_{x,y} P(x,y)\log_2 \frac{P(x,y)}{P(x)P(y)}.$$

请证明 $I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) \geq 0$, 给出等号成立的条件, 并用一句话描述互信息的含义. (提示: 使用 Jensen 不等式)

(2) **[5pts]** 在 ID3 决策树的生成过程中, 使用信息增益 (information gain) 为划分指标以生成新的结点. 试证明或给出反例: 在 ID3 决策树中, 根结点处划分的信息增益不小于其他结点处划分的信息增益.

(3) **[5pts]** 设离散属性 a 有 V 种可能的取值 $\{a^1, \cdots, a^V\}$, 请使用《机器学习》4.2.1 节相关符号证明:

$$\mathrm{Gain}(D,a) = \mathrm{Ent}(D) - \sum_{v=1}^{V}\frac{|D^v|}{|D|}\mathrm{Ent}(D^v) \geq 0$$

即信息增益是非负的. (提示: 将信息增益表示为互信息的形式, 你需要定义表示分类标记的随机变量, 以及表示属性 a 取值的随机变量)

(4) **[5pts]** 除教材中介绍的信息熵、基尼指数 (gini index) 外, 也可以使用误分类错误率 (misclassification error)

$$1 - \max_k p_k$$

作为衡量集合纯度的指标. 请从决策树生成过程的角度给出这一指标的合理性, 并结合二分类问题 ($|\mathcal{Y}| = 2$) 下三种纯度指标的表达式, 分析各衡量标准的特点.

Solution. 此处用于写解答 (中英文均可)

(1) 根据随机变量信息熵、条件熵以及互信息的定义, 容易证明

$$\begin{aligned}
I(X;Y) &= \sum_{x,y} P(x,y)\log_2 \frac{P(x,y)}{P(x)P(y)} \\
&= -\sum_{x,y} P(x,y)\log_2 P(x) + \sum_{x,y} P(x,y)\log_2 P(x \mid y) \\
&= H(X) - H(X \mid Y).
\end{aligned}$$

同理可以验证 $I(X;Y) = H(Y) - H(Y \mid X)$. 下面证明互信息的非负性,

$$\begin{aligned}
I(X;Y) &= -\sum_{x,y} P(x,y)\log_2 \frac{P(x)P(y)}{P(x,y)} \\
&\geq -\log_2 \sum_{x,y} P(x,y)\frac{P(x)P(y)}{P(x,y)} \\
&= -\log_2 \sum_{x,y} P(x)P(y) = -\log_2 1 = 0.
\end{aligned}$$

其中不等号利用了 $\log_2(x)$ 函数的凹性以及 Jensen 不等式.

互信息取值为 0, 当且仅当满足 Jensen 不等式的取等条件, 即对所有的 $P(x,y) > 0$, $\frac{P(x)P(y)}{P(x,y)} = c$ 是一个常数. 这意味着随机变量 X 与随机变量 Y 相互独立. 由此可以看出, 互信息代表着随机变量之间的公共信息 (共有的信息量), 当两个随机变量独立时, 共有信息量为 0, 互信息为 0.

(2) 该结论错误, 给出一种反例即可. 例如考虑包含两个布尔属性 X, Y 的空间, 目标函数为亦或函数 $f(x,y) = x$ XOR y. 生成相应的决策树如下:

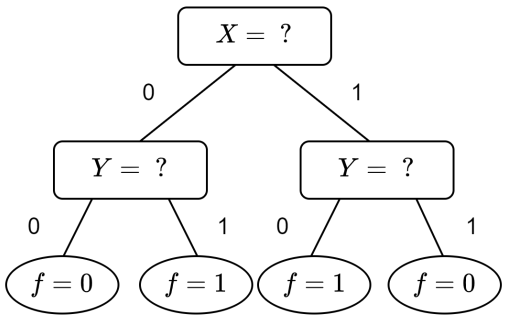


图 1: XOR 决策树

其根节点处划分的信息增益为 0, 而任意一个子结点处的划分的信息增益均为 1.

(3) 定义随机变量 L 表示分类的标记, 其分布与数据集 D 中类别的经验分布一致. 定义随机变量 A 为划分属性 a 的取值, D^v 是属性 a 上取值为 a^v 的样本集合, 随机变量 A 的分布为数据集 D 中属性 a 取值的经验分布 (各个 D^v 的占比). 基于目前的定义, 我们有如下观察

– 随机变量 L 的信息熵即为样本集合 D 的信息熵, $H(L) = \mathrm{Ent}(D)$;

– 信息增益中的权重 $\frac{|D^v|}{|D|}$ 可以看作 A 取值为 a^v 的概率;

– 给定 $A=a^v$ 条件下随机变量 L 的信息熵, 即为样本子集 D^v 的信息熵, $H(L\,|\,A=a^v)=\mathrm{Ent}(D^v)$.

因此可将信息增益表达为互信息的形式, 并得到其非负性:

$$\begin{aligned}\mathrm{Gain}(D,a)&=\mathrm{Ent}(D)-\sum_{v=1}^{V}\frac{|D^v|}{|D|}\mathrm{Ent}(D^v)\\&=H(L)-\sum_{v=1}^{V}P(A=a^v)H(L\mid A=a^v)\\&=H(L)-H(L\mid A)=I(L;A)\geq 0.\end{aligned}$$

(4) 决策树生成过程中, 叶结点中使用包含样例最多的类别作为其预测结果, 故误分类错误率从经验误差的角度衡量了样本子集的纯度. 假设正类所占比例为 p, 可得二分类问题下信息熵、基尼指数以及误分类错误率的形式分别为

$$-p\log_2 p-(1-p)\log_2(1-p),\quad 2p(1-p),\quad 1-\max(p,1-p).$$

各指标关于 p 的变化趋势如图3所示 (信息熵按 0.5 的比例缩放). 可以看出, 信息熵和基尼指数的曲线更加平滑并且可导, 便于直接针对指标进行数值优化, 而误分类错误率是分段线性函数, 在 $p=0.5$ 处不可导. 误分类错误率对结点概率 p 变化的敏感程度也不如信息熵和基尼指数. 与此同时, 使用误分类错误率为集合纯度指标、“最小训练误差”作为决策树划分选择, 容易导致决策树过拟合.(言之有理即可)

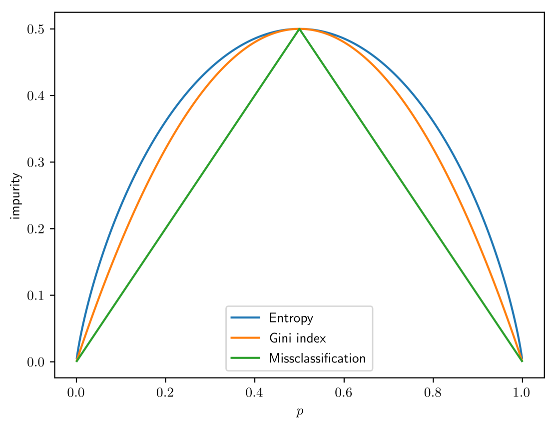


图 2: 数据纯度衡量指标对比

4 [20pts] Training a Decision Tree

剪枝 (pruning) 是决策树学习算法对抗“过拟合”的主要手段. 考虑下面的训练集: 共计 8 个训练样本, 每个训练样本有三个特征属性 X, Y, Z 和标签信息. 详细信息如表1所示.

表 1: 训练集信息

编号	X	Y	Z	f	编号	X	Y	Z	f
1	1	1	0	1	5	0	0	0	0
2	1	1	1	1	6	1	0	1	0
3	0	0	1	0	7	1	1	0	1
4	0	1	0	0	8	0	1	1	1

(1) **[5pts]** 请通过训练集中的数据训练决策树, 要求使用“信息增益”(information gain) 作为划分准则.(需说明详细计算过程)

(2) **[10pts]** 进一步考虑如表2所示的验证集, 对上一问得到的决策树基于这一验证集进行预剪枝、后剪枝. 生成叶子结点时, 若样例最多的类别不唯一, 可任选其中一类. 请画出所有可能的剪枝结果.(需说明详细计算过程)

表 2: 验证集信息

编号	X	Y	Z	f
9	1	1	1	1
10	1	0	1	0
11	1	0	1	1
12	0	1	0	0
13	0	1	1	1
14	1	0	0	0

(3) **[5pts]** 请给出预剪枝决策树和后剪枝决策树分别在训练集、验证集上的准确率. 结合本题的结果, 讨论预剪枝与后剪枝在欠拟合风险、泛化能力以及训练时间开销层面各自的特点.

Solution. 此处用于写解答 (中英文均可)

(1) 记现有数据集为 D, 属性集合 $A = \{X, Y, Z\}$. 第一次划分时, 先计算出根结点信息熵

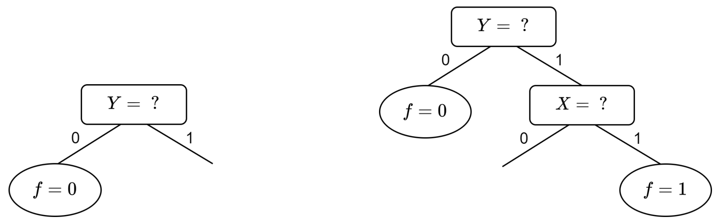
$$\mathrm{Ent}(D) = -\sum_{f=0}^{1} p_f \log_2 p_f = 1.$$

对于属性 X, 信息增益为: $\mathrm{Gain}(D, X) = \mathrm{Ent}(D) - \sum_{x=0}^{1} \frac{|D^x|}{|D|}\mathrm{Ent}(D^x) = \frac{3}{4}\log_2 3 - 1;$

对于属性 Y, 信息增益为: $\mathrm{Gain}(D, Y) = \mathrm{Ent}(D) - \sum_{y=0}^{1} \frac{|D^y|}{|D|}\mathrm{Ent}(D^y) = 2 - \frac{5}{8}\log_2 5;$

对于属性 Z, 信息增益为: $\mathrm{Gain}(D, Z) = \mathrm{Ent}(D) - \sum_{z=0}^{1} \frac{|D^z|}{|D|}\mathrm{Ent}(D^z) = \frac{3}{2} - \frac{5}{8}\log_2 5.$

因此第一次应对属性 Y 进行划分. 进而, 在 $Y=0$ 的分支, 不难发现该分支上的样本类别相同, 因此 $Y=0$ 分支不必再进行划分, 得到图3(a).



(a) 第一次划分 (b) 第二次划分

图 3: 决策树的前两次划分

接下来对 $Y=1$ 的分支进行划分, 此时 $Y=1$ 的分支对应的数据为表1中标号为 $\{1,2,4,7,8\}$ 的样本, 对数据集进行更新, 使得 D 只包含这些样本. 对属性集合进行更新 $A \leftarrow A \setminus \{Y\}$.

对于属性 X, 信息增益: $\mathrm{Gain}(D,X) = \mathrm{Ent}(D) - \sum_{x=0}^{1} \frac{|D^x|}{|D|}\mathrm{Ent}(D^x) = \log_2 5 - 2$;

对于属性 Z, 信息增益: $\mathrm{Gain}(D,Z) = \mathrm{Ent}(D) - \sum_{z=0}^{1} \frac{|D^z|}{|D|}\mathrm{Ent}(D^z) = \log_2 5 - \frac{3}{5}\log_2 3 - \frac{6}{5}$;

因此第二次应对属性 X 进行划分. 进而, 在 $X=1$ 的分支, 不难发现该分支上的样本类别相同, 因此 $X=1$ 分支不必再进行划分, 得到图3(b). 此时仅剩下属性 Z, 并且数据集在左分支中并没有得到相同的标记. 因此, 左分支以属性 Z 进行划分即可, 得到最终如图4所示的决策树.

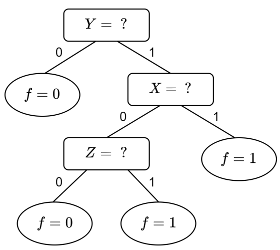


图 4: 完整决策树

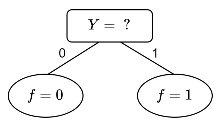


图 5: 预剪枝决策树

(2) 我们先讨论预剪枝. 在根结点处, 若不进行划分, 则该结点将被标记为训练样例数最多的类别. 任选其中一类, 得到验证集精度均为 50%. 在使用属性 Y 划分之后, 各子结点 (从左至右) 分别包含编号为 $\{3,5,6\}$、$\{1,2,4,7,8\}$ 的训练样例, 将分别被标记为叶结点 "$f=0$"、"$f=1$". 此时验证集中编号为 {9,10,13,14} 的样例被分类正确, 验证集精度为 $\frac{4}{6} \times 100\% = 66.7\%$, 因此用属性 Y 进行划分得以确定.

接下来我们分析是否需要用属性 X 进行进一步划分. 划分后各子结点 (从左至右) 分别包含编号为 {4,8}、{1,2,7} 的训练样例, 后者标记为叶结点 "$f=1$", 前者任选一类作为标记. 可以发现, 此时的验证集精度均为 66.7%, 没有因为该划分得到提升, 于是预剪枝策略禁止第二次使用属性 X 的划分. 最终的预剪枝决策树如图5所示.

然后针对完整的决策树如图4, 讨论后剪枝. 该决策树的验证集精度为 $\frac{5}{6} \times 100\% = 83.3\%$. 首先考虑4中的 "$Z=?$" 结点, 若将其领衔的分支剪除, 则替换后的叶结点包含编号为 $\{4,8\}$ 的训练样本. 任选正类或负类作为该叶结点的标记, 此时决策树的验证集精度均会降低至 66.7%. 于是后剪枝策略会决定保留, 最终的后剪枝决策树如图6所示, 与完整决策树一致.

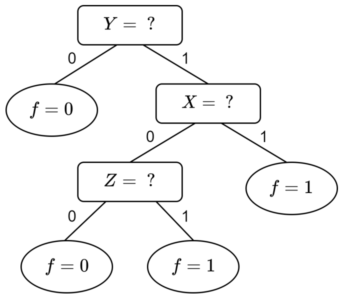


图 6: 后剪枝决策树

(3) 预剪枝决策树在训练集、验证集上的准确率分别为 $\frac{7}{8}, \frac{2}{3}$, 后剪枝决策树在训练集、验证集上的准确率分别为 $1, \frac{5}{6}$. 通过本题的计算, 可以看到后剪枝决策树的过拟合风险更低、泛化性能更强. 与此同时, 后剪枝操作需要在生成完整决策树之后才能进行, 故后剪枝决策树的训练时间开销比预剪枝决策树要大.(言之有理即可)

5 [20pts] Kernel Function

核函数是 SVM 中常用的工具, 其在机器学习中有着广泛的应用与研究. 请自行阅读学习《机器学习》第 6.3 节, 并回答如下问题.

(1) **[5pts]** 试判断 $\kappa(\mathbf{x},\mathbf{z}) = (\langle \mathbf{x},\mathbf{z}\rangle - 1)^2$ 是否为核函数, 并给出证明或反例.

(2) **[5pts]** 试证明: 对于半正定矩阵 $\mathbf{A}$, 总存在半正定矩阵 $\mathbf{C}$, 成立 $\mathbf{A} = \mathbf{C}^\top\mathbf{C}$

(3) **[5pts]** 试证明: 若 κ_1 和 κ_2 为核函数, 则两者的直积

$$\kappa_1 \otimes \kappa_2(\mathbf{x},\mathbf{z}) = \kappa_1(\mathbf{x},\mathbf{z})\kappa_2(\mathbf{x},\mathbf{z})$$

也是核函数;

(4) **[5pts]** 试证明 $\kappa(\mathbf{x},\mathbf{z}) = \langle \mathbf{x},\mathbf{z}\rangle^p$ 对 $\forall p \in \mathbb{Z}^+_+(p < \infty)$ 均为核函数.

Solution. 此处用于写解答 (中英文均可)

(1) 该函数不能作为核函数, 给出反例如下.

考虑一维变量, 并取数据集 $D = \{\mathbf{x}_1, \mathbf{x}_2\}$（$\mathbf{x}_1 = 2$,$\mathbf{x}_2 = -2$）. 此时对应的核矩阵（kernel matrix）为:

$$\mathbf{K} = \begin{bmatrix} 9 & 25 \\ 25 & 9 \end{bmatrix}$$

其行列式为 $|\mathbf{K}| = 9 \times 9 - 25 \times 25 = -544 < 0$, 说明其存在负特征值, 不为半正定矩阵. 故 $\kappa(\mathbf{x},\mathbf{z}) = (\langle \mathbf{x},\mathbf{z}\rangle - 1)^2$ 不是核函数.

(2) 对实对称矩阵 $\mathbf{A}$ 可以进行谱分解如下:

$$\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$$

其中 $\mathbf{Q}$ 为正交矩阵, 因此成立:

$$\mathbf{Q}^\top\mathbf{Q} = \mathbf{I}$$

同时, 因为 $\mathbf{A}$ 为半正定矩阵, 所以其特征值均非负. 故成立:

$$\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top = \mathbf{Q}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{Q}^\top\mathbf{Q}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{Q}^\top$$

因此, 取半正定矩阵 $\mathbf{C} = \mathbf{Q}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{Q}^\top$, 成立 $\mathbf{A} = \mathbf{C}^\top\mathbf{C}$.

(3) 考虑到核函数与核矩阵的充要关系, 本题等价于证明: 若矩阵 $\mathbf{A} = \{a_{ij}\}_{m\times m}, \mathbf{B} = \{b_{ij}\}_{m\times m}$ 均为半正定矩阵, 则矩阵 $\mathbf{H} = \{a_{ij}b_{ij}\}_{m\times m}$ 也为半正定矩阵. 下证明该结论.

首先由第 (2) 问结论, 因为 $\mathbf{A}$ 半正定, 可设 $\mathbf{A} = \mathbf{C}^\top\mathbf{C}$, 即 $a_{ij} = \sum_{k=1}^m c_{ik}c_{jk}$.

任取 $\mathbf{x} \in \mathcal{R}^m$, 成立:

$$\begin{aligned}
\mathbf{x}^\top\mathbf{H}\mathbf{x} &= \sum_{i,j} a_{ij}b_{ij}x_ix_j \\
&= \sum_{i,j}\left(\sum_{k=1}^m c_{ik}c_{jk}\right) b_{ij}x_ix_j \\
&= \sum_{k=1}^m \left[\sum_{i,j} b_{ij}\left(c_{ik}x_i\right)\left(c_{jk}x_j\right)\right]
\end{aligned}$$

同时, 因为 $\mathbf{B}$ 也为半正定矩阵, 因此对于任意 k, 成立:

$$\sum_{i,j} b_{ij}\left(c_{ik}x_i\right)\left(c_{jk}x_j\right) \geq 0$$

故对任意 $\mathbf{x} \in \mathcal{R}^m$, 成立 $\mathbf{x}^T\mathbf{H}\mathbf{x} \geq 0$, 即 $\mathbf{H}$ 也为半正定矩阵, 证毕.

(4) 取 $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m\}\,(\mathbf{x}_i \in \mathcal{R}^n)$, 并设矩阵 $\mathbf{K} = \{K_{ij}\}_{m\times m} = \{\langle \mathbf{x}_i, \mathbf{x}_j\rangle^p\}_{m\times m}, \forall p \in \mathbb{Z}_+^+$. 则原命题的充要条件, 即证明矩阵 $\mathbf{K}$ 半正定. 下用数学归纳法证明矩阵 $\mathbf{K}$ 半正定.

(a) 当 $p = 1$ 时:

$$\mathbf{K}_1 = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_m^\top \end{bmatrix} \cdot \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_m \end{bmatrix} = \mathbf{X}^\top \mathbf{X}$$

其中 $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_m \end{bmatrix}$.
任取向量 $\mathbf{y}$, 成立 $\mathbf{y}^T\mathbf{K}_1\mathbf{y} = (\mathbf{X}\mathbf{y})^\top(\mathbf{X}\mathbf{y}) \geq 0$, 因此 $\mathbf{K}_1$ 半正定.

(b) 假设 $p = k$ 时 $\mathbf{K}_k$ 半正定仍然成立, 即 $\mathbf{K}_k = \left\{\langle \mathbf{x}_i, \mathbf{x}_j\rangle^k\right\}_{m\times m}$ 半正定.

(c) 当 $p = k+1$ 时, 成立:

$$\mathbf{K}_{k+1} = \left\{\langle \mathbf{x}_i, \mathbf{x}_j\rangle^{k+1}\right\}_{m\times m} = \left\{\langle \mathbf{x}_i, \mathbf{x}_j\rangle^k \langle \mathbf{x}_i, \mathbf{x}_j\rangle\right\}_{m\times m}$$

由初始条件 (a) 及归纳假设 (b) 可知 $\{\langle \mathbf{x}_i, \mathbf{x}_j\rangle\}_{m\times m}$,$\left\{\langle \mathbf{x}_i, \mathbf{x}_j\rangle^k\right\}_{m\times m}$ 均半正定. 故由第 (3) 问结论, 知 $\mathbf{K}_{k+1}$ 半正定.

综上所述, $\mathbf{K}$ 为半正定矩阵得证, 进而 $\kappa(\mathbf{x}, \mathbf{z}) = \langle \mathbf{x}, \mathbf{z}\rangle^p$ 是核函数得证.