# 机器学习导论 习题二

211300024, 石睿, 211300024@smail.nju.edu.cn

2023 年 4 月 8 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) 解:

由题目表述，需要证明在 $\sum_{\mathbf{1}}=\sum_{\mathbf{2}}=\sum$ 的时候, 证明

$Condition1: h(x)=0\Leftrightarrow P(x)\leq Q(x)$

$Condition2: h(x)=1\Leftrightarrow P(x)\geq Q(x)$

以下证明 Condition1，Condition2 同理的呢！

$Condition1:$

$\because$ 二元正态分布的概率密度函数$f(x)=\dfrac{1}{(2\pi)^{\frac{n}{2}}\cdot|\sum_{\boldsymbol{i}}|^{\frac{1}{2}}}\cdot e^{\left(-\frac{1}{2}\cdot(\mathbf{x}-\boldsymbol{\mu}_i)^T\sum_{\boldsymbol{i}}^{\mathbf{-1}}(\mathbf{x}-\boldsymbol{\mu}_i)\right)}$ (i=1、2)

$\because P(x)\leq Q(x)$

$\therefore (\mathbf{x}-\boldsymbol{\mu}_1)^T\sum^{\mathbf{-1}}(\mathbf{x}-\boldsymbol{\mu}_1)\geq(\mathbf{x}-\boldsymbol{\mu}_2)^T\sum^{\mathbf{-1}}(\mathbf{x}-\boldsymbol{\mu}_2)$

$\therefore 2\mathbf{x}^T\sum^{\mathbf{-1}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\leq\boldsymbol{\mu}_1^T\sum^{\mathbf{-1}}\boldsymbol{\mu}_1-\boldsymbol{\mu}_2^T\sum^{\mathbf{-1}}\boldsymbol{\mu}_2 \qquad (*)$

因为 h(x) 是 LDA 训练的最优分类器，其训练的结果是得到一个单位方向向量 w。

每一个样本 x 在 w 上的投影离哪一类样本中心的投影近就归为哪一类

$\because h(x)=0\Leftrightarrow P(x)\leq Q(x)\Leftrightarrow$ x 属于 Q 类

$\therefore \left|\boldsymbol{w}^T\mathbf{x}-\boldsymbol{w}^T\boldsymbol{\mu}_2\right|\leq\left|\boldsymbol{w}^T\mathbf{x}-\boldsymbol{w}^T\boldsymbol{\mu}_1\right| \qquad (**)$

因为 LDA 的最优解$\boldsymbol{w}=\boldsymbol{S}_w^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)=(2\sum)^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)=\frac{1}{2}\sum^{\mathbf{-1}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) \qquad (***)$

综上，依据条件 *、**、***，本题可以证明以下等价结论

$$\left|\boldsymbol{w}^T\mathbf{x}-\boldsymbol{w}^T\boldsymbol{\mu}_2\right| \le \left|\boldsymbol{w}^T\mathbf{x}-\boldsymbol{w}^T\boldsymbol{\mu}_1\right| \Leftrightarrow 2\mathbf{x}^T\sum\nolimits^{\mathbf{-1}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) \le \boldsymbol{\mu}_1^T\sum\nolimits^{\mathbf{-1}}\boldsymbol{\mu}_1-\boldsymbol{\mu}_2^T\sum\nolimits^{\mathbf{-1}}\boldsymbol{\mu}_2 \qquad (\Delta)$$

$\because \boldsymbol{w}=\dfrac{1}{2}\sum\nolimits^{\mathbf{-1}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$ 且协方差矩阵是半正定的

$\therefore \boldsymbol{w}^T(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)=\dfrac{1}{2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)^T\sum\nolimits^{\mathbf{-1}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\ge 0$

由上式，正类 P 和负类 Q 以 w 为单位方向向量的直线上的相对位置固定

$\therefore$ 新样本点 x 在 w 上投影的位置有以下两种情况，会满足 $\left|\boldsymbol{w}^T\mathbf{x}-\boldsymbol{w}^T\boldsymbol{\mu}_2\right| \le \left|\boldsymbol{w}^T\mathbf{x}-\boldsymbol{w}^T\boldsymbol{\mu}_1\right|$

$Situation1$ x 在 w 上的投影在原点和负类 Q 在 w 投影的数据中心之间。

此时 x 离 Q 中心更近，所以满足 $\left|\boldsymbol{w}^T\mathbf{x}-\boldsymbol{w}^T\boldsymbol{\mu}_2\right| \le \left|\boldsymbol{w}^T\mathbf{x}-\boldsymbol{w}^T\boldsymbol{\mu}_1\right|$，以下证明右侧不等式成立

$2\mathbf{x}^T\sum\nolimits^{\mathbf{-1}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\le 2\boldsymbol{\mu}_1^T\sum\nolimits^{\mathbf{-1}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$

$\le(\boldsymbol{\mu}_1^T+\boldsymbol{\mu}_2^T)\sum\nolimits^{\mathbf{-1}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)=\boldsymbol{\mu}_1^T\sum\nolimits^{\mathbf{-1}}\boldsymbol{\mu}_1-\boldsymbol{\mu}_2^T\sum\nolimits^{\mathbf{-1}}\boldsymbol{\mu}_2$

$Situation2$ x 在 w 上的投影在正类 P 投影中心和负类 Q 投影中心之间。且离 Q 更近

此时 x 离 Q 中心更近，所以满足$\boldsymbol{w}^T\mathbf{x}-\boldsymbol{w}^T\boldsymbol{\mu}_2\le\boldsymbol{w}^T\mathbf{x}-\boldsymbol{w}^T\boldsymbol{\mu}_1$，以下证明右侧不等式成立

$\because \boldsymbol{w}^T\mathbf{x}-\boldsymbol{w}^T\boldsymbol{\mu}_2\le\boldsymbol{w}^T\mathbf{x}-\boldsymbol{w}^T\boldsymbol{\mu}_1$

$\therefore \mathbf{x}^T\boldsymbol{w}\le\boldsymbol{\mu}_1^T\boldsymbol{w}+\boldsymbol{\mu}_2^T\mathbf{2}$ 且 $\boldsymbol{w}=\dfrac{1}{2}\sum\nolimits^{\mathbf{-1}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$

$\therefore 2\mathbf{x}^T\boldsymbol{w}\sum\nolimits^{\mathbf{-1}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\le\boldsymbol{\mu}_1^T\sum\nolimits^{\mathbf{-1}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)+\boldsymbol{\mu}_2^T\sum\nolimits^{\mathbf{-1}}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)=\boldsymbol{\mu}_1^T\sum\nolimits^{\mathbf{-1}}\boldsymbol{\mu}_1-\boldsymbol{\mu}_2^T\sum\nolimits^{\mathbf{-1}}\boldsymbol{\mu}_2$

综上，Condition1 分为两种满足左侧不等式的两种情况 Situation1 和 Situation2，均满足等式右侧

同理，Condition2 也可以分为两种情况分开讨论得到本结论啦

(2) 解:

由书中推导，上述结论的最优解为 w,w 满足以下结论

w 是$S_w^{-1}S_b$的 d' 个最大非零特征值所对应的特征向量所组成的矩阵

$\because r(S_w^{-1}S_b) \leq min\{r(S_w^{-1}), r(S_b)\}$

$\therefore$ 以下证明$r(S_b) \leq N-1$

$$\because S_b = \sum_{i=1}^{N} m_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})(\boldsymbol{\mu}_i - \boldsymbol{\mu})^T$$

$\therefore$ 设$\boldsymbol{x_i} = m_i(\boldsymbol{\mu}_i - \boldsymbol{\mu})$且$\boldsymbol{y_i} = (\boldsymbol{\mu}_i - \boldsymbol{\mu})^T$

$$\therefore S_b = \sum_{i=1}^{N} \boldsymbol{x_i y_i}$$

因为均值向量的定义

$$\therefore \sum_{i=1}^{N} m_i(\boldsymbol{\mu}_i - \boldsymbol{\mu}) = 0$$

$$\begin{aligned}
\because \sum_{i=1}^{N} \boldsymbol{x_i y_i} &= \sum_{i=1}^{N-1} \boldsymbol{x_i y_i} + \boldsymbol{x_N y_N} = \sum_{i=1}^{N-1} \boldsymbol{x_i y_i} - \sum_{i=1}^{N-1} \mathrm{x}_i \boldsymbol{y_N} = \sum_{i=1}^{N-1} \boldsymbol{x_i}(\boldsymbol{y_i} - \boldsymbol{y_N}) \\
&= \sum_{i=1}^{N-1} \boldsymbol{x_i}\left((\boldsymbol{\mu}_i - \boldsymbol{\mu})^T - (\boldsymbol{\mu}_N - \boldsymbol{\mu})^T\right) = \sum_{i=1}^{N-1} \boldsymbol{x_i}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_N)^T \\
&= \sum_{i=1}^{N-1} m_i(\boldsymbol{\mu}_i - \boldsymbol{\mu}_N)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_N)^T = \sum_{i=1}^{N-1} m_i \alpha_i
\end{aligned}$$

其中上式中的$\alpha_i$可以看成是矩阵的一个基，一共至多有 N-1 个，即秩比 N-1 小

## 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) 解:

1. $OvO$

一共$C_k^2 = \dfrac{k(k-1)}{2}$个二分类任务，每个任务数据大小$\dfrac{2n}{k}$

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

2. $OvR$

一共$k$个二分类任务，每个任务数据大小$n$

$$T(n) = k \cdot \mathcal{O}(n^\alpha)$$

(2) 解:

1. $OvO$的编码矩阵如下所示

其中每行为不同的类别$C_i(i=1\dots K)$ 每列为不同的分类器$f_i(i=1\dots\frac{K(K-1)}{2})$

$$\begin{pmatrix} 1 & 1 & \cdots & 0 & 0 & \cdots & 0 \\ -1 & 0 & \cdots & 1 & 1 & \cdots & 0 \\ 0 & -1 & \cdots & -1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 0 & \cdots & -1 \end{pmatrix}$$

2. $OvR$的编码矩阵如下所示

其中每行为不同的类别$C_i(i=1\dots K)$ 每列为不同的分类器$f_i(i=1\dots K)$

$$\begin{pmatrix} 1 & -1 & -1 & \cdots & -1 \\ -1 & 1 & -1 & \cdots & -1 \\ -1 & -1 & 1 & \cdots & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -1 & -1 & -1 & \cdots & 1 \end{pmatrix}$$

(3) 解:

$\because$ 由题目的提示，使用 k-1 组对数几率回归模型预测

$$\therefore \begin{cases} \ln\frac{p(y=1|\mathbf{x})}{p(y=k|\mathbf{x})} = \boldsymbol{w_1^T x} + b_1 \\ \quad\vdots \\ \ln\frac{p(y=k-1|\mathbf{x})}{p(y=k|\mathbf{x})} = \boldsymbol{w_{k-1}^T x} + b_{k-1} \end{cases}$$

$$\text{设}p(y=i|\mathbf{x}) = x_i \quad \text{且}\sum_{i=1}^{k} x_i = 1 \qquad (*)$$

$$\therefore \begin{cases} \frac{x_1}{x_k} = e^{\boldsymbol{w_1^T x}+b_1} \\ \quad\vdots \\ \frac{x_{k-1}}{x_k} = e^{\boldsymbol{w_{k-1}^T x}+b_{k-1}} \end{cases} \qquad (**)$$

$$\text{综上 * 和 **，解得} \begin{cases} x_i = \frac{e^{\boldsymbol{w_i^T x}+b_i}}{e^{\boldsymbol{w_1^T x}+b_1}+\dots+e^{\boldsymbol{w_{k-1}^T x}+b_{k-1}}+1} \\ \quad\vdots \\ x_k = \frac{1}{e^{\boldsymbol{w_1^T x}+b_1}+\dots+e^{\boldsymbol{w_{k-1}^T x}+b_{k-1}}+1} \end{cases}$$

$$\because \text{由似然函数} \quad l(\boldsymbol{w}, b) = \sum_{i=1}^{m} \ln p(y_i|\mathbf{x}_i; \boldsymbol{w}, b)$$

$\because$ 类比二分类中的化简 $\quad p(y_i|\mathbf{x}_i;\boldsymbol{w},b) = (p_1(\widehat{x};\beta))_1^{y} \cdot (p_0(\widehat{x};\beta))^{1-y_1}$

在 K 分类的时候，同样引入$\boldsymbol{\beta_i} = \begin{pmatrix} w_i \\ b_i \end{pmatrix} \quad \widehat{\boldsymbol{x}} = \begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix}$ 就有$\boldsymbol{w_i^T}\mathbf{x} + b_i = \boldsymbol{\beta_i^T}\widehat{x}$

$$\therefore l(\boldsymbol{w},b) = l(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\ldots,\boldsymbol{\beta}_{k-1};\mathbf{x}) = \sum_{i=1}^{m}\ln\left(\prod_{j=1}^{j} p(y_i = j|\widehat{x}_i,\boldsymbol{\beta}_i)^{\mathbf{I}(y_i=j)}\right)$$

$$= \sum_{i=1}^{m}\sum_{j=1}^{k}\left(I(y_i=j)\cdot\,\ln p(y_i=j|\widehat{x}_i,\boldsymbol{\beta}_i)\right)$$

$$= \sum_{i=1}^{m}\left(\sum_{j=1}^{k-1}\left(\mathbf{I}(y_i=j)\cdot\,\ln p(y_i=j|\widehat{x}_i,\boldsymbol{\beta}_i)\right) + \mathbf{I}(y_i=k)\cdot\,\ln p(y_i=k|\widehat{x}_i,\boldsymbol{\beta}_i)\right)$$

$$\text{因为}\mathbf{I}(y_i=k) = 1 - \sum_{p=1}^{k-1}\mathbf{I}(y_i=p)$$

$$= \sum_{i=1}^{k}\left(\sum_{j=1}^{k-1}(\mathbf{I}(y_i=j)\cdot\,\ln\frac{p(y_i=j|\widehat{x}_i,\boldsymbol{\beta}_i)}{y_i=k|\widehat{x}_i,\boldsymbol{\beta}_i}) + \ln p(y_i=k|\widehat{x}_i,\boldsymbol{\beta})\right)$$

$$= \sum_{i=1}^{k}\left(\sum_{j=1}^{k-1}(\mathbf{I}(y_i=j)\cdot\,\ln(e^{\boldsymbol{\beta}_j^T\cdot\widehat{x}_j})) - \ln(1+\sum_{p=1}^{k-1}e^{\boldsymbol{\beta}_p^T\widehat{x}_p})\right)$$

$$= \sum_{i=1}^{k}\left(\sum_{j=1}^{k-1}(\mathbf{I}(y_i=j)\cdot\boldsymbol{\beta}_j^T\cdot\widehat{x}_j) - \ln(1+\sum_{p=1}^{k-1}e^{\boldsymbol{\beta}_p^T\widehat{x}_p})\right)$$

$$\therefore \frac{\partial l}{\partial\boldsymbol{\beta}_j} = \sum_{i=1}^{m}\left(\mathbf{I}(y_i=j)\widehat{x}_j - \frac{e^{\boldsymbol{\beta}_j^T\widehat{x}_j\widehat{x}_j}\widehat{x}_j}{1+\sum_{p=1}^{k-1}e^{\boldsymbol{\beta}_p^T\widehat{x}_p}}\right)$$

$$= \sum_{i=1}^{m}\left(\mathbf{I}(y_i=j)\widehat{x}_j - p(y=j|\widehat{x}_j)\widehat{x}_j\right)$$

## 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 $Y$, 条件熵 (conditional entropy)$H(Y|X)$ 定义如下:

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

诠释为 $Y$ 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 $I(X;Y)=H(X)-H(X|Y)=H(Y)-H(Y|X)\geq 0$, 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) 解:

1.1 证明: $\mathbf{I}(x;y)=H(X)-H(X|Y)=H(Y)-H(Y|X)$

以下证明$\mathbf{I}(x;y)=H(X)-H(X|Y)$ $\quad$ $\mathbf{I}(x;y)=H(Y)-H(Y|X)$同理

$$
\begin{aligned}
\mathbf{I}(X;Y) = H(X) - H(X|Y) &= -\sum_x p(x)\log_2 p(x) + \sum_y p(y)\sum_x p(x|y)\log_2 p(y|x) \\
&= -\sum_x p(x)\log_2 p(x) + \sum_{x,y} p(x,y)\log_2 \frac{p(x,y)}{p(x)} \\
&= \sum_{x,y} p(x,y)\log_2 p(x,y) - \sum_{x,y} p(x,y)\log_2 p(x) - \sum_x p(x)\log_2 p(x) \\
&\because \sum_x p(x)\log_2 p(x) = \sum_{x,y} p(x,y)\log_2 p(y) \\
\therefore \mathbf{I}(X;Y) &= \sum_{x,y} p(x,y)\left(\log_2 p(x,y) - \log_2 p(x) - \log_2 p(y)\right) \\
&= \sum_{x,y} p(x,y)\log_2 \frac{p(x,y)}{p(x)p(y)}
\end{aligned}
$$

1.2 证明：$I(X;Y) \geq 0$

$$
\begin{aligned}
\because \mathbf{I}(x,y) &= \sum_{x,y} p(x,y)\log_2 \frac{p(x,y)}{p(x)p(y)} \geq \sum_{x,y} p(x,y)\log_2 e^{1-\frac{p(x,y)}{p(x)p(y)}} \\
&= \sum_{x,y} \left(p(x,y) - p(x)p(y)\right)\log_2 e \geq 0
\end{aligned}
$$

其中上面第一个不等号的放缩用到了以下不等式。在 x>1 的时候，$x > e^{1-x}$

1.3 含义

$\mathbf{I}(X,Y)$表示在一直一个随机变量的信息后，让另外一个随机变量的不确定性减小的程度

(2) 解：

反例，书中 P77

根节点，Gain(D, 纹理)=0.381 是当前样本集合再所有属性集合中信息增益最大

$D_1$结点，Gain($D_1$, 根蒂)=0.458 是当前样本集合再所有属性集合中信息增益最大

此时，有 Gain($D_1$, 根蒂)>Gain(D, 纹理)，和题干命题矛盾

(3) 解：

设 X：训练集 D 中样本示例　Y：属性 a 取值

$$
\because \begin{cases}
\mathbf{I}(X;Y) = Ent(X) - Ent(X|Y) \geq 0 \\
Ent(X) = Ent(D) \\
Ent(X|Y) = -\sum_y p(y)Ent(X|Y=y) \\
p(y) = \frac{|D^v|}{|D|} \\
Ent(X|Y=y) = Ent(D^v)
\end{cases}
\qquad \therefore I(X,Y) = Gain(D,a) \geq 0
$$

(4) 解:

以下设 $miserror\_index$ 为题干中的五分类错误率

4.1　从决策树生成过程给出合理性

$\because miserror = 1 - max_k P_k$

$\therefore miserror\_index(D,a) = \sum_{i=1}^{V} \frac{|D^i|}{|D|} \cdot miserror(D^i)$

$Reason1:$

决策树生成中不断选择当前能把样本分的最开的一种属性，如果 $miserror\_index(D,a)$ 越小也即类别数量最多的被分错的概率都非常小，则其他类别被分错的概率更小啦

$Reason2:$

在决策树决定不再扩展的时候，以某节点对应的训练集子集中最多的类别作为分类结果

而此数量最多的类在此节点上的错误率就是 $miserror\_index$

4.2二分类时三种纯度表达式，分析各自特点

$Ent(D) = -(p\log_2 p + (1-p)\log_2(1-p))$

$Gini(D) = 1 - (p^2 + (1-p)^2)$

$Miserror(D) = 1 - max\{p, 1-p\}$

$Angle1:$ 取值范围

$Ent(D)$在 [0,1] 之间取值　$Gini(D), Miserror(D)$在 $[0,\frac{1}{2}]$ 之间取值

$Angle2:$ 函数特点

$Ent(D), Gini(D)$是非线性的

$Miserror(D)$是线性的

## 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) 解:

1.  Node C (root)

$Ent(C) = 1$

$Attempt1$ : Divide by X

$X = 0$  3 个 1,1 个 0  $Ent(C^1) = -(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4})$

$X = 1$  1 个 1,3 个 0  $Ent(C^2) = -(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4})$

$Gain(C, X) = 1 - \frac{1}{2}Ent(C^1) - \frac{1}{2}Ent(C^2) = 0.185$

$Attempt2$ : Divide by Y

$Y = 0$  4 个 1,1 个 0  $Ent(C^1) = -(\frac{4}{5}\log_2\frac{4}{5} + \frac{1}{5}\log_2\frac{1}{5})$

$Y = 1$  0 个 1,3 个 0  $Ent(C^2) = -1\log_2 1$

$Gain(C, Y) = 1 - \frac{5}{8}Ent(C^1) - \frac{3}{8}Ent(C^2) = 0.554$

$Attempt3$ : Divide by Z

$Z = 0$  2 个 1,2 个 0  $Ent(C^1) = -(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2})$

$Z = 1$  2 个 1,2 个 0  $Ent(C^2) = -(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2})$

$Gain(C, X) = 1 - \frac{1}{2}Ent(C^1) - \frac{1}{2}Ent(C^2) = 0$

综上，选择最大最大的基尼指数的 Y 作为根节点 C 的划分属性, 得到节点 A 和 B

2.Node A, 此时 A 所有的样本为 {1,2,4,7,8}

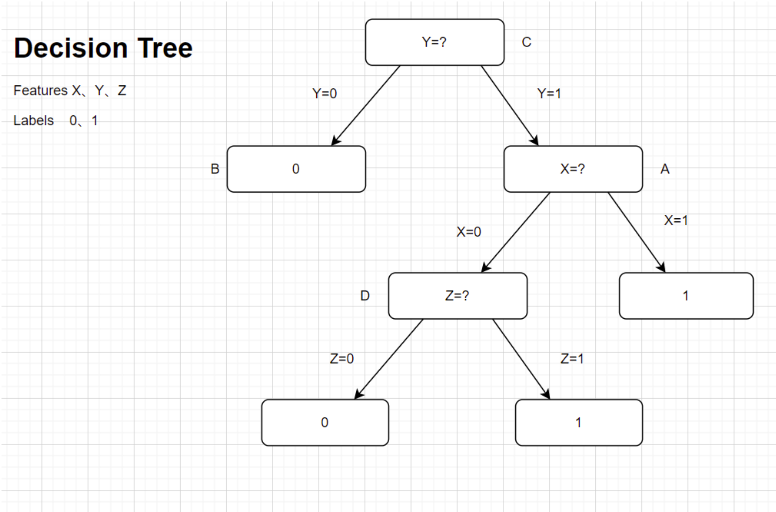
具体过程和根节点 C 的方式相同，此时只需要选择属性 X 和 Z 进行划分尝试

$Gain(A, X) = 0.32$

$Gain(A, Z) = 0.16$

综上，选择最大最大的基尼指数的 X 作为根节点 C 的划分属性, 得到节点 D 和 E

至此，所有结点所包含的样本均纯（全为 0 或者 1），决策树生成算法结束. 结果如下图所示




(2) 解：

1. 预剪枝

$situation1$ : Node C

不继续生成，C 就是叶节点—训练集中 4 个 1，4 个 0，好坏占比相同，取 1 作为本叶节点类别

验证集中 3 个 1，3 个 0。全被当成 1 了，正确率 50%

继续生成，依据第一问选择 Y 作为分类属性

验证集中 Y=1（2 个分类正确，1 个错误），Y=0（2 个分类正确，1 个错误），正确率 66.7%

综上，应该对节点 C 继续划分，决策树继续生成！

$situation2$ : Node A

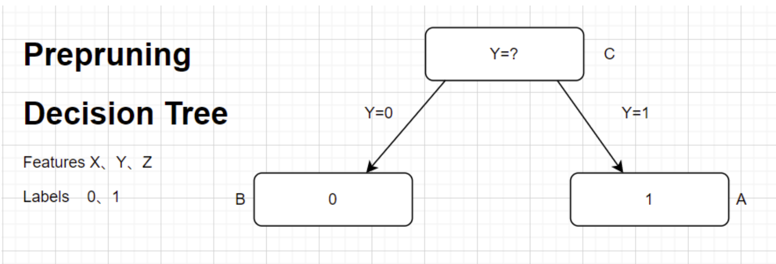
不继续生成，A 就是叶节点

正确率 66.7%

继续生成，依据第一问选择 X 作为分类属性

验证集中 X=1（1 个分类正确，0 个错误），X=0（1 个分类正确，1 个错误），正确率 66.7%

综上，不应该对节点 A 继续划分，决策树不继续生成啦！



2. 后剪枝

$situation1$ : Node D

不剪枝，验证集中 5 正确，1 错误 剪枝，验证集中 4 正确，2 错误

综上，不对节点 D 的生成进行剪枝

$situation2$ : Node A

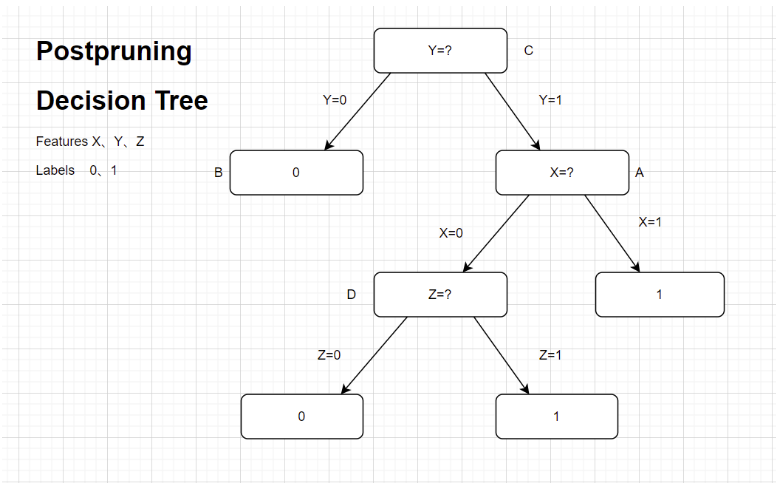
不剪枝，验证集中 5 正确，1 错误 剪枝，验证集中 4 正确，2 错误

综上，不对节点 A 的生成进行剪枝

$situation3$ : Node C

不剪枝，验证集中 5 正确，1 错误 剪枝，验证集中 3 正确，3 错误

综上，不对节点 C 的生成进行剪枝




(3) 解:

预剪枝，4 正确 2 错误，正确率 66.7%

后剪枝，5 正确 1 错误，正确率 83.3%

| | 欠拟合情况 | 泛化能力 | 训练时间 |
|---|---|---|---|
| 预剪枝 | 风险较大，可能当前分支没有对纯度提升<br>但被剪掉的后续分支对纯度有所提升 | 因欠拟合<br>泛化性能一般 | 许多分支未展开<br>训练、测试时间少 |
| 后剪枝 | 比预剪枝保留更多分支<br>欠拟合风险较小 | 展开较多<br>泛化性能好 | 自底向上对所有<br>非叶节点检查，时间长 |

# 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 $\kappa_1$ 和 $\kappa_2$ 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) 解:

反例如下

$$\mathbf{x} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad z = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$K = \begin{pmatrix} k(x,x) & k(x,z) \\ k(z,x) & k(z,z) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad \text{并非是半正定的}$$

(2) 解:

Review：高等代数定理

可逆矩阵 A 和 E 合同，并且和任一正对角矩阵合同

半正定矩阵和任一个非负对角矩阵合同

$$\because \text{由上面定理，存在可逆矩阵 C，有} C^T AC = \begin{pmatrix} a_1 & & & \\ & a_2 & & \\ & & \ddots & \\ & & & a_n \end{pmatrix} \quad (a_i \geq 0)$$

$$\begin{aligned}
\therefore A &= (C^T)^{-1} \begin{pmatrix} a_1 & & & \\ & a_2 & & \\ & & \ddots & \\ & & & a_n \end{pmatrix} C^{-1} \\
&= (C^T)^{-1} \begin{pmatrix} \sqrt{a_1} & & & \\ & \sqrt{a_2} & & \\ & & \ddots & \\ & & & \sqrt{a_n} \end{pmatrix} \begin{pmatrix} \sqrt{a_1} & & & \\ & \sqrt{a_2} & & \\ & & \ddots & \\ & & & \sqrt{a_n} \end{pmatrix} C^{-1} \\
&= \left( \begin{pmatrix} \sqrt{a_1} & & & \\ & \sqrt{a_2} & & \\ & & \ddots & \\ & & & \sqrt{a_n} \end{pmatrix} C^{-1} \right)^T \left( \begin{pmatrix} \sqrt{a_1} & & & \\ & \sqrt{a_2} & & \\ & & \ddots & \\ & & & \sqrt{a_n} \end{pmatrix} C^{-1} \right) \\
&= G^T G
\end{aligned}$$

其中 G 是半正定的

(3) 解:

$\because$ 由（2）可知，$K_1$ 和 $K_2$ 均可表示成半正定的转置乘以这个半正定矩阵

$\therefore$ 不妨设$K_1 = G_1^T G_1 \quad K_2 = G_2^T G_2$

其中$G_1 = (g_{ij})_{m*m} \quad G_2 = (g^{'}_{ij})_{m*m}$

$$\therefore K_1 = (x_{ij})_{m*m} = \sum_{k=1}^{m} g_{ki} g_{kj} \quad K_2 = (y_{ij})_{m*m} = \sum_{k=1}^{m} g^{'}_{ki} g^{'}_{kj}$$

$\therefore \forall Z \in \mathcal{R}^n$

$$\begin{aligned}
Z^T K Z &= \sum_{i=1}^{m}\sum_{j=1}^{m} z_i x_{ij} y_{ij} z_j \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m}\left( (\sum_{k=1}^{m} g_{ki} g_{kj})(\sum_{p=1}^{m} g^{'}_{pi} g^{'}_{pj}) z_i z_j \right) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m}\left( g_{ki} g_{kj} (\sum_{p=1}^{m} g^{'}_{pi} g^{'}_{pj}) z_i z_j \right) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{p=1}^{m}\sum_{k=1}^{m}\left( g_{ki} g_{kj} g^{'}_{pi} g^{'}_{pj} z_i z_j \right) \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{p=1}^{m}\sum_{k=1}^{m}\left( (g_{ki} g^{'}_{pi} z_i)(g_{kj} g^{'}_{pj} z_j) \right) \\
&= \sum_{i=1}^{m}\left( \sum_{k=1}^{m}\sum_{p=1}^{m} (g_{ki} g^{'}_{pi} z_i)^2 \right) \geq 0
\end{aligned}$$

(4) 解:

$\because$ 由（3）中的结论，以下只需要证明$k_1(x,z) =< x, z >$ 是核函数即可啦

$\because K = (k_{ij})_{m*m}$

$\therefore k_{ij} = x_i^T z_j$

$\therefore \forall Z \in \mathcal{R}$

$$\begin{aligned}
Z^T K Z &= \sum_{k=1}^{m}\sum_{j=1}^{m} z_i k_{ij} z_j \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m} z_i x_i^T x_j z_j \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m} z_i \left( \sum_{k=1}^{m} (x_i)_k (x_j)_k \right) z_j \\
&= \sum_{k=1}^{m}\left( \sum_{i=1}^{m} z_i (x_i)_k \right)^2 \geq 0
\end{aligned}$$

$\because$ 综上，由（3）中的结论

$\therefore k(x,z) =< x, z >^p = k_1 \otimes k_1 \otimes k_1 \cdots \otimes k_1$直积 p 次，仍然是核函数!