机器学习导论 习题二

211300063, 张运吉, 211300063@smail.nju.edu.cn

2023 年 4 月 17 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、编程题 .ipynb 文件; **请将二者打包为 .zip 文件上传**. 注意命名规则, 三个文件均命名为"学号_姓名" + ". 后缀" (例如 211300001_张三" + ".pdf"、".ipynb"、".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:

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

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 $d' \le N - 1$. (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) 当 $P(x) = Q(x)$ 时，

$$\frac{1}{(2\pi)^{\frac{n}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}}} exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)\right) = \frac{1}{(2\pi)^{\frac{n}{2}}\boldsymbol{\Sigma}^{\frac{1}{2}}} exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)\right)$$

化简得：

$$\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1) - \frac{1}{2}(\boldsymbol{\mu}_2+\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1) = 0$$

上式即为满足最优分类性质的分类平面。

通过 LDA 得到的分类器:

$$\begin{aligned} \boldsymbol{w} &= \mathbf{S}_w^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) \\ &= (\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) \\ &= \frac{1}{2}\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) \end{aligned}$$

两类样本投影中心

$$\boldsymbol{\mu}_1' = \frac{1}{2}\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$$

$$\boldsymbol{\mu}_2' = \frac{1}{2}\boldsymbol{\mu}_2^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)$$

分类点:

$$\begin{aligned} \boldsymbol{\mu}' &= \frac{1}{2}(\boldsymbol{\mu}_1'+\boldsymbol{\mu}_2') \\ &= \frac{1}{4}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2) \end{aligned}$$

所以由 LDA 得到的分类平面:

$$\boldsymbol{x}^T\boldsymbol{w} - \boldsymbol{\mu}' = 0$$

即：

$$\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)-\frac{1}{2}(\boldsymbol{\mu}_2+\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_2-\boldsymbol{\mu}_1)=0$$

所以在 $\boldsymbol{\Sigma_1}=\boldsymbol{\Sigma_2}=\boldsymbol{\Sigma}$ 时，通过 LDA 得到的分类器可实现最优分类.

(2) 由课本上定义：$\mathbf{S}_b=\sum_{i=1}^{N}m_i(\boldsymbol{\mu}_i-\boldsymbol{\mu})(\boldsymbol{\mu}_i-\boldsymbol{\mu})^T$ 以及 rank $(m_i(\boldsymbol{\mu}_i-\boldsymbol{\mu})(\boldsymbol{\mu}_i-\boldsymbol{\mu})^T)\leq \min\{\boldsymbol{\mu}_i-\boldsymbol{\mu},(\boldsymbol{\mu}_i-\boldsymbol{\mu})^T\}=1$
所以: rank $(\mathbf{S}_b)\leq N$
因为一共有 N 个类别，所以第 N 个类别的均值 μ_N 可以由前 $N-1$ 个类别的均值线性表示，因此 $A_N=m_N(\boldsymbol{\mu}_N-\boldsymbol{\mu})(\boldsymbol{\mu}_N-\boldsymbol{\mu})^T$ 可以由 $A_i=m_i(\boldsymbol{\mu}_i-\boldsymbol{\mu})(\boldsymbol{\mu}_i-\boldsymbol{\mu})^T,i=0,1,...,N-1$ 线性表示。
所以: rank $(\mathbf{S}_b)\leq N-1$
rank $(\mathbf{S}_w^{-1}\mathbf{S}_b)\leq$ rank $(\mathbf{S}_b)\leq N-1$
因为：对于任意 n 阶方阵, 其非零特征值个数小于等于其秩大小
所以：矩阵 $\mathbf{S}_w^{-1}\mathbf{S}_b$ 的非零特征值个数 $\leq N-1$
所以：$d^{'}\leq N-1$

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) OvO:

$$\begin{aligned} T(n) &= \frac{K(K-1)}{2} \times \mathcal{O}((\frac{2n}{K})^\alpha) \\ &= \mathcal{O}(K^{2-\alpha} n^\alpha) \end{aligned}$$

OvR:

$$\begin{aligned} T(n) &= K \times \mathcal{O}(n^\alpha) \\ &= \mathcal{O}(K n^\alpha) \end{aligned}$$

(2) OvO:

	f_1	f_2	f_3	f_4	f_5	f_6
C_1	1	1	1	0	0	0
C_2	-1	0	0	1	1	0
C_3	0	-1	0	-1	0	1
C_4	0	0	-1	0	-1	-1

OvR:

	f_1	f_2	f_3	f_4
C_1	1	-1	-1	-1
C_2	-1	1	-1	-1
C_3	-1	-1	1	-1
C_4	-1	-1	-1	1

(3) 设:

$$\ln \frac{p(y=k|\mathbf{x})}{p(y=K|\mathbf{x})} = \beta_k^\top \mathbf{x}, \quad k=1,...,K-1, \beta_k = (\omega_k, b_k)$$

显然有:

$$p(y=k|\mathbf{x}) = \frac{e^{\beta_k^\top \mathbf{x}}}{1+\sum_{i=1} e^{\beta_i^\top \mathbf{x}}}, \quad k=1,...,K-1$$
$$p(y=K|\mathbf{x}) = \frac{1}{1+\sum_{i=1} e^{\beta_i^\top \mathbf{x}}}$$

对数似然函数:

$$l(\beta_k) = \sum_{i=1}^{m} \ln p(y_i|\mathbf{x}_i; \beta_\mathbf{k})$$

定义指示函数 $\mathbb{I}(\cdot)$ 使得答案简洁,

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

似然项可重写为:

$$p(y_i|\mathbf{x}_i; \beta) = \sum_{i=1}^{K} \mathbb{I}(y=i) p_i(\mathbf{x}_i; \beta_i)$$

对数似然函数可重写为:

$$
\begin{aligned}
l(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \ln p\left(y_i|\mathbf{x_i};\boldsymbol{\beta}\right) \\
&= \sum_{i=1}^{n} \ln \sum_{j=1}^{K} \mathbb{I}(y_i=j) p_j\left(\mathbf{x}_i;\boldsymbol{\beta}_j\right) \\
&= \sum_{i=1}^{n} \ln \left(\sum_{j=1}^{K-1} \mathbb{I}(y_i=j) \frac{\exp\left(\boldsymbol{\beta}_j^{\top}\mathbf{x}_i\right)}{1+\sum_{j=1}^{K-1}\exp\left(\boldsymbol{\beta}_j^{\top}\mathbf{x}_i\right)} + \mathbb{I}(y_i=K) \frac{1}{1+\sum_{j=1}^{K-1}\exp\left(\boldsymbol{\beta}_j^{\top}\mathbf{x}_i\right)} \right) \\
&= \sum_{i=1}^{n} \left(\sum_{j=1}^{K-1} \mathbb{I}(y_i=j)\boldsymbol{\beta}_j^{\top}\mathbf{x}_i - \ln\left(1+\sum_{j=1}^{K-1}\exp(\boldsymbol{\beta}_j^{\top}\mathbf{x}_i)\right) \right).
\end{aligned}
$$

其梯度:

$$
\nabla l(\beta) = \sum_{i=1}^{m} \mathbf{x}_i \left(\frac{\sum_{j=1}^{K-1}\mathbb{I}(y_i=j)e^{\beta_j^{\top}\mathbf{x}_i}}{\sum_{j=1}^{K-1}\mathbb{I}(y_i=j)e^{\beta_j^{\top}\mathbf{x}_i}+\mathbb{I}(y_i=K)} - \sum_{j=1}^{K-1} p(y=j|\mathbf{x}_i) \right)
$$

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) 先证 $I(X;Y)=H(X)-H(X|Y)=H(Y)-H(Y|X)$

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

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

下证 $I(X;Y)\geqslant 0$

先证明 KL 散度非负, 即

$$KL(p\|q) = -\int p(x)\ln\left(\frac{q(x)}{p(x)}\right) \geq 0$$

对凸函数 f(x) , 由 Jensen 不等式:

$$f(E(x)) \leq E(f(x))$$

对于连续变量

$$f\left(\int xp(x)dx\right) \leq \int f(x)p(x)dx$$

因为 $-\ln y$ 为凸函数，令 $y=\frac{q(x)}{p(x)}$, 得 $f(x)=-\ln y=-\ln\frac{q(x)}{p(x)}$

对 f 使用 jensen 不等式：

$$KL=\int p(x)\left[-\ln\left(\frac{q(x)}{p(x)}\right)\right] \geq -\ln\left(\int \frac{q(x)}{p(x)}p(x)dx\right) = -\ln\left(\int q(x)dx\right) = 0$$

得 KL 散度非负.

利用以上结论

$$\begin{aligned}
I(X;Y) &= \sum_{x,y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)\\
&= KL[p(x,y)\|p(x)p(y)]\\
&\geqslant 0
\end{aligned}$$

等号成立的条件: 存在常数 a, b, 使得 $-\ln \frac{p(x)p(y)}{p(x,y)} = ax + b$
互信息含义：用来评价两个随机变量之间的依赖程度的一个度量.

(2) 错误的, 考虑如下反例:
假设我们有一个简单的数据集, 其中包含两个属性 1 和 2, 以及二元分类标签 (0 或 1)
数据集如下：
$D = \{(0,0,1), (0,1,0), (1,0,0), (1,1,1)\}$
其中 (a, b, c) 表示样本在属性 1 和属性 2 分别取值 a, b，标签为 c
经过计算, 属性 1 和属性 2 的信息增益相同, 不妨选择特征 a 作为划分属性, 根节点的信息增益为:

$$Gain(D, 1) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v) = 1 - (\frac{1}{2} + \frac{1}{2}) = 0$$

使用 1 属性对 D 集合进行划分后, 可得两个子集 $D_1 = \{(0,0,1), (0,1,0)\}$ 和 $D_2 = \{(1,0,0), (1,1,1)\}$, 分别记为左节点和右节点. 对于左节点, 划分属性只剩下 2, 故选择 2 作为划分属性, 左节点的信息增益为:

$$Gain(D_1, 2) = Ent(D_1) - \sum_{v=1}^{V} \frac{|D_1^v|}{|D_1|} Ent(D_1^v) = 1 - 0 = 1$$

左节点的信息增益大于根节点的信息增益, 故题干的结论是错误的.

(3) X_C 表示分类标记的随机向量,X_{a_i} 表示分类属性 a_i 取值的随机向量

$$\begin{aligned}
I(X_C; X_{a_i}) &= H(X_C) - H(X_C|X_{a_i}) \\
&= -\sum_{x_c} p(x_c) \log(p(x_c)) - \sum_{x_{a_i}} p(x_{a_i}) H(X_C \mid X_{a_i} = x_{a_i}) \\
&= Gain(C, a_i)
\end{aligned}$$

其中，$p(x_c)$ 是分类标记 C 中出现 x_c 的概率,$p(x_{a_i})$ 是属性 a_i 取值为 x_{a_i} 的概率,$p(x_c|x_{a_i})$ 是在已知属性 a_i 取值为 x_{a_i} 的条件下分类标记 C 中出现 x_c 的概率.
由 (1) 可知互信息 $I(X_C; X_{a_i})$ 是非负的, 因此信息增益 $Gain(C, a_i)$ 也是非负的.

(4) 误分类错误率表示子集中被错误分类的样本所占的比例，在决策树生成的过程中，每次进行划分的时候，需要尽可能找到最优的属性，如果一个属性能够很好地分开不同类别的样本，则分裂后的子节点中同类别样本的比例会更高，误分类错误率会更低. 因此，可以误分类错误率作为纯度指标。

(a) 信息熵：$Ent(D) = -p_1 \log_2 p_1 - p_2 \log_2 p_2$
信息熵是度量样本集合纯度最常用的一种指标。信息熵越小，表示子集的纯度越高。它的优点是能够充分考虑每个类别的权重，有效处理样本不平衡问题, 并且在决策树构建的过程中能够产生较为平衡的树结构。但是，计算信息熵的代价较高, 对于连续变量需要进行离散化。

(b) 基尼指数：$Gini(D) = 1 - p_1^2 - p_2^2$

基尼指数 $Gini(D)$ 反映了从数据集 D 中随机抽取两个样本，其分类标记不一致的概率，因此基尼指数越小表示子集的纯度越高。与信息熵类似，考虑了各个类别的分布情况，但相对于信息熵计算更简单，对于连续变量也更加方便处理。但是它对于类别权重的处理不如信息熵。

(c) 误分类错误率：$E(D) = 1 - \max(p_1, p_2)$

误分类错误率是一种较为简单的指标，如果一个属性能够很好地分开不同类别的样本，则分裂后的子节点中同类别样本的比例会更高，误分类错误率会更低. 误分类错误率计算简单，结果直观，但是其忽略了样本中不同类别的比例，若数据集较为不均衡，容易出现过拟合或者欠拟合的现象.

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) 原始数据集信息熵为

$$\mathrm{Ent}(D) = -(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}) = 1$$

若用属性 X 划分:

$$\mathrm{Ent}(D^1) = -\left(\frac{3}{4}\log_2\frac{3}{4} + \frac{1}{4}\log_2\frac{1}{4}\right) \approx 0.811$$
$$\mathrm{Ent}(D^2) = -\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{3}{4}\log_2\frac{3}{4}\right) \approx 0.811$$
$$\mathrm{Gain}(D, X) = 1 - \left(\frac{1}{2}\times 0.811 + \frac{1}{2}\times 0.811\right) = 0.189$$

若用属性 Y 划分:

$$\mathrm{Ent}(D^1) = -\left(\frac{4}{5}\log_2\frac{4}{5} + \frac{1}{5}\log_2\frac{1}{5}\right) \approx 0.722$$
$$\mathrm{Ent}(D^2) = -(0 + 1\cdot\log_2 1) = 0$$
$$\mathrm{Gain}(D, Y) = 1 - \left(\frac{5}{8}\times 0.722\right) = 0.549$$

若用属性 Z 划分:

$$\mathrm{Ent}(D^1) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1$$
$$\mathrm{Ent}(D^2) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1$$
$$\mathrm{Gain}(D, Z) = 1 - (\frac{1}{2}\times 1 + \frac{1}{2}\times 1) = 0$$

使用属性 Y 划分得到的信息增益最大,使用属性 Y 划分得到:$D_1 = \{1, 2, 4, 7, 8\}, D_2 = \{3, 5, 6\}$

$$\mathrm{Ent}(D_1) = 0.722$$

Y=0 时，所有节点的标签都是 0，标记为叶子节点
对 Y=1 节点进行划分.
若用属性 X 划分:

$$\mathrm{Ent}(D^1) = 0$$
$$\mathrm{Ent}(D^2) = 1$$
$$\mathrm{Gain}(D_1, X) = 0.322$$

若用属性 Z 划分:

$$\mathrm{Ent}(D^1) = 0$$
$$\mathrm{Ent}(D^2) = -(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}) = 0.918$$
$$\mathrm{Gain}(D_1, Z) = 0.722 - 0.918\times 0.6 = 0.1712$$

使用属性 X 划分得到的信息增益最大，根据属性 X 进行划分
X=1 时，标记为叶子节点。
X=0 由于只剩下属性 Z, 所以使用 Z 进行划分，Z=1，Z=0 各为一个叶子节点.

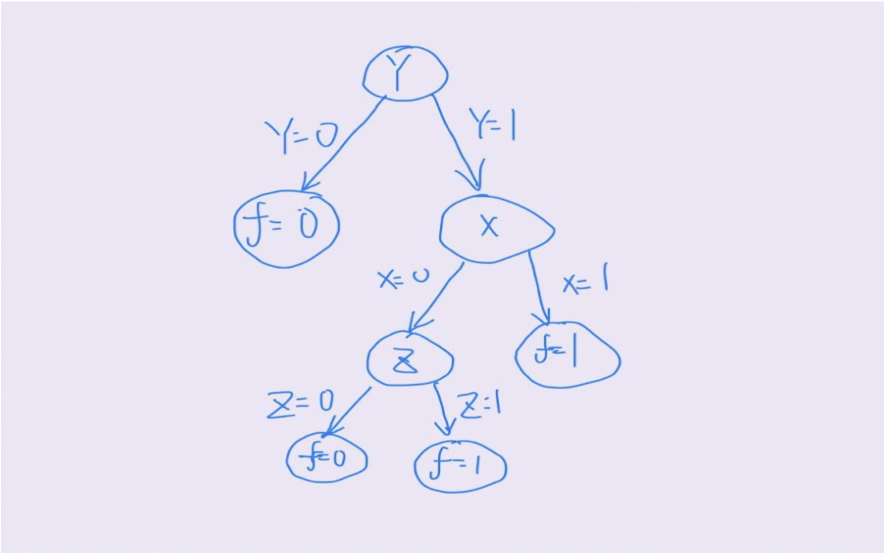

图 1: ID3 决策树

(2) 预剪枝:

属性 Y 划分前精准度为 $\frac{1}{2}$，划分后精准度为 $\frac{2}{3}$，所以应该划分.

属性 X 划分前精准度为 $\frac{2}{3}$，划分后精准度为 $\frac{2}{3}$，所以不应划分.

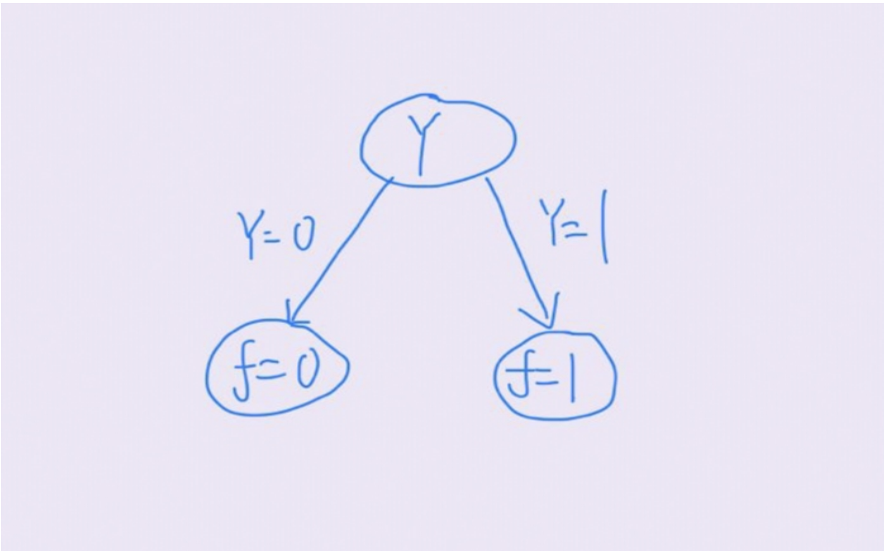

图 2: 预剪枝决策树

后剪枝：

该决策树在验证集上的精度为 $\frac{5}{6}$.

如果不划分 Z，验证集精度为 $\frac{2}{3}$，所以不进行剪枝.

如果不划分 X，验证集精度为 $\frac{5}{6}$，所以可以不进行剪枝.

如果不划分 Y，验证集精度为 $\frac{1}{2}$，所以不进行剪枝.

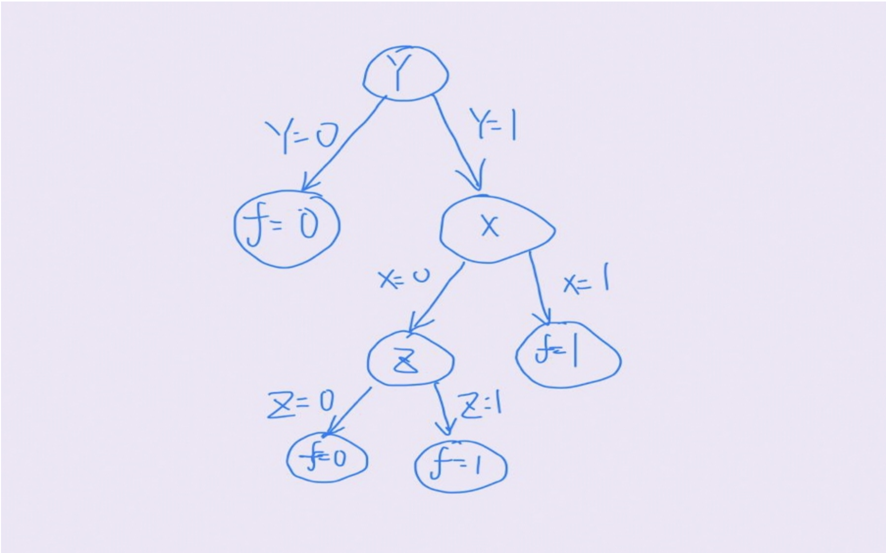

图 3: 后剪枝决策树

(3) 预剪枝决策树训练集精度为 87.5%，验证集精度为 66%.
后剪枝决策树训练集精度为 100%，验证集精度为 83%
预剪枝使得决策树的很多分支都没有“展开”，这不仅降低了过拟合的风险，而且减少了决策树的训练时间开销和预测时间开销，但另一方面，有些分支的当前划分虽然不能提升泛化性能甚至会降低泛化性能，但是在其基础上进行的后续划分却有可能导致性能显著提高，所以预剪枝有欠拟合的风险。
后剪枝通常比预剪枝保留更多的分支，一般情况下，后剪枝的欠拟合风险很小，泛化性能往往优于预剪枝决策树，但是后剪枝决策树的训练时间开销比预剪枝大得多。

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) 不是核函数.
反例：令数据集 $D=\{(1,0),(0,1)\}$
则 $k(x_1,x_1)=0,k(x_1,x_2)=1,k(x_2,x_1)=1,k(x_2,x_2)=0$
核矩阵:

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \tag{5.1}$$

不是半正定矩阵.

(2) 因为 $\mathbf{A}$ 为半正定矩阵，则其特征值 $\lambda_1,\lambda_2,...,\lambda_n$ 均大于等于 0，则存在正交矩阵 $\mathbf{P}$ 使得

$$\begin{aligned}\mathbf{A}&=\mathbf{P}\operatorname{diag}\left\{\lambda_1,\lambda_2,\ldots,\lambda_n\right\}\mathbf{P}^{-1}\\&=\mathbf{P}\operatorname{diag}\left\{\sqrt{\lambda_1},\sqrt{\lambda_2},\ldots,\sqrt{\lambda_n}\right\}\mathbf{P}^{-1}\mathbf{P}\operatorname{diag}\left\{\sqrt{\lambda_1},\sqrt{\lambda_2},\ldots,\sqrt{\lambda_n}\right\}\mathbf{P}^{-1}\\&=\mathbf{P}\operatorname{diag}\left\{\sqrt{\lambda_1},\sqrt{\lambda_2},\ldots,\sqrt{\lambda_n}\right\}\mathbf{P}^{-1}\left(\mathbf{P}^{-1}\right)\top\operatorname{diag}\left\{\sqrt{\lambda_1},\sqrt{\lambda_2},\ldots,\sqrt{\lambda_n}\right\}\mathbf{P}\\&=\mathbf{P}\operatorname{diag}\left\{\sqrt{\lambda_1},\sqrt{\lambda_2},\ldots,\sqrt{\lambda_n}\right\}\mathbf{P}^{-1}\left(\mathbf{P}\operatorname{diag}\left\{\sqrt{\lambda_1},\sqrt{\lambda_2},\ldots,\sqrt{\lambda_n}\right\}\mathbf{P}^{-1}\right)\top\end{aligned} \tag{5.2}$$

令 $\mathbf{C}=\mathbf{P}\left\{\sqrt{\lambda_1},\sqrt{\lambda_2},\ldots,\sqrt{\lambda_n}\right\}\mathbf{P}^{-1}$, 则 $\mathbf{A}=\mathbf{C}^\top\mathbf{C}$, $\mathbf{C}$ 是半正定矩阵。

(2) $\because A$ 为半正定矩阵
$\therefore$ 存在可逆矩阵 D, 使得

$$A=D^T\begin{pmatrix} E_r & 0 \\ 0 & 0 \end{pmatrix}D \tag{5.3}$$

其中 $r = rank(A)$

$\therefore$ 下式成立：

$$A = D^T \begin{pmatrix} E_r & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} E_r & 0 \\ 0 & 0 \end{pmatrix} D = \left(\begin{pmatrix} E_r & 0 \\ 0 & 0 \end{pmatrix} D \right)^T \left(\begin{pmatrix} E_r & 0 \\ 0 & 0 \end{pmatrix} D \right) \quad (5.4)$$

令

$$C = \begin{pmatrix} E_r & 0 \\ 0 & 0 \end{pmatrix} D \quad (5.5)$$

则 $A = C^T C$

(3) 设 K_1, K_2 分别表示 k_1, k_2 的核矩阵

$K = K_1 \otimes K_2$ 表示 $\kappa_1 \otimes \kappa_2(\mathbf{x}, \mathbf{z})$ 的核矩阵

$$\begin{aligned}
yK^Ty &= y \begin{pmatrix} k_1(x_1,z_1)k_2(x_1,z_1) & k_1(x_1,z_2)k_2(x_1,z_2) & ... & k_1(x_1,z_m)k_2(x_1,z_m) \\ k_1(x_2,z_1)k_2(x_2,z_1) & k_1(x_2,z_2)k_2(x_2,z_2) & ... & k_1(x_1,z_m)k_2(x_1,z_m) \\ ... & ... & ... & ... \\ k_1(x_m,z_1)k_2(x_m,z_1) & k_1(x_m,z_2)k_2(x_m,z_2) & ... & k_1(x_m,z_m)k_2(x_m,z_m) \end{pmatrix} y^T \\
&= \sum_{i=1}^{m}\sum_{j=1}^{m} k_1(x_i,z_j)k_2(x_i,z_j)y_iy_j \\
&= tr\left(\begin{pmatrix} k_1(x_1,z_1)y_1 & k_1(x_1,z_2)y_1 & ... & k_1(x_1,z_m)y_1 \\ k_1(x_2,z_1)y_2 & k_1(x_2,z_2)y_2 & ... & k_1(x_1,z_m)y_2 \\ ... & ... & ... & ... \\ k_1(x_m,z_1)y_m & k_1(x_m,z_2)y_m & ... & k_1(x_m,z_m)y_m \end{pmatrix} \begin{pmatrix} k_2(x_1,z_1)y_1 & k_2(x_2,z_1)y_1 & ... \\ k_2(x_1,z_2)y_2 & k_2(x_2,z_2)y_2 & ... \\ ... & ... & ... \\ k_2(x_1,z_m)y_m & k_2(x_2,z_m)y_m & ... \end{pmatrix} \right. \\
&= tr\left(\begin{pmatrix} y_1 & & & \\ & y_2 & & \\ & & ... & \\ & & & y_m \end{pmatrix} K_1 \begin{pmatrix} y_1 & & & \\ & y_2 & & \\ & & ... & \\ & & & y_m \end{pmatrix} K_2^T \right) \\
&= tr\left(\begin{pmatrix} y_1 & & & \\ & y_2 & & \\ & & ... & \\ & & & y_m \end{pmatrix} C^TC \begin{pmatrix} y_1 & & & \\ & y_2 & & \\ & & ... & \\ & & & y_m \end{pmatrix} D^TD \right)
\end{aligned} \quad (5.6)$$

上式最后一个等式是由于 K_1, K_2 是半正定矩阵，由 (2) 中结论：

$K_1 = C^TC, K_2 = D^TD$

因为交换矩阵顺序不改变矩阵的迹，所以:

$$
\begin{aligned}
yK^Ty &= tr\left(\left(\begin{matrix} y_1 & & & \\ & y_2 & & \\ & & ... & \\ & & & y_m \end{matrix}\right)C^TC\left(\begin{matrix} y_1 & & & \\ & y_2 & & \\ & & ... & \\ & & & y_m \end{matrix}\right)D^TD\right) \\
&= tr\left(\left(C\left(\begin{matrix} y_1 & & & \\ & y_2 & & \\ & & ... & \\ & & & y_m \end{matrix}\right)D^T\right)^T\left(C\left(\begin{matrix} y_1 & & & \\ & y_2 & & \\ & & ... & \\ & & & y_m \end{matrix}\right)D^T\right)\right) \\
&= tr(Q^TQ) \geq 0
\end{aligned}
\tag{5.7}
$$

所以 K 半正定，$\kappa_1 \otimes \kappa_2(\mathbf{x}, \mathbf{z})$ 是核函数。

(4) 先证明 $p = 1$ 时，$k(x, z) = x^Tz$ 是核函数
由核函数的定义, 假设输入空间为 $\mathcal{X}$
定义映射 $\phi : \mathcal{X} \to \mathcal{H}$, $\phi(x) = x$
那么对所有的 $x, z \in \mathcal{X}$
$k(x, z) = x^Tz = \phi(x) \cdot \phi(z)$
所以 k 是核函数
假设当 $p = n - 1$ 的时候，$\kappa(\mathbf{x}, \mathbf{z}) = \langle \mathbf{x}, \mathbf{z} \rangle^p$ 也是核函数
由问 (3) 得到的结论当 $p = n$ 的时候，$\kappa(\mathbf{x}, \mathbf{z}) = \kappa_{n-1} \bigotimes \kappa_1(\mathbf{x}, \mathbf{z}) = \langle \mathbf{x}, \mathbf{z} \rangle^{n-1} \cdot \langle \mathbf{x}, \mathbf{z} \rangle$，所以 $p = n$ 时也是核函数.
数学归纳法成立，$\kappa(\mathbf{x}, \mathbf{z}) = \langle \mathbf{x}, \mathbf{z} \rangle^p$ 对 $\forall p \in \mathbb{Z}_+ (p < \infty)$ 均为核函数.