3.2

$$y=\frac{1}{1+e^{-(w^Tx+b)}}$$

求导

$$\frac{\partial y}{\partial w^T}=-\frac{e^{(-w^Tx+b)}(-x)}{(1+e^{-(w^Tx+b)})^{-2}}=x(y-y^2)$$

二阶导

$$\frac{\partial}{\partial w^T}(\frac{\partial y}{\partial w^T})=x(1-2y)(\frac{\partial y}{\partial w})^T=x(1-2y)x^T(y-y^2)$$

对于 xPx^T，对任意向量Z有$Z^TxPx^TZ=\sum_i P_{ii}v_i^2\geq 0$，因此其海森矩阵半正定

在二阶导中，xx^T的秩为1，非零特征向量为1，y的值域为$(0,1)$，当$y\in(0.5,1)$时

$$y(y-1)(1-2y)<0$$

二阶导半负定，则公式(3.18)非凸

$$l(\beta)=\sum_{i=1}^{m}(-y_i\beta^T\hat{x}_i+\ln(1+e^{\beta^T\hat{x}_i}))$$

$$\frac{\partial}{\partial\beta^T}(\frac{\partial l}{\partial\beta})=\sum_{i=1}^{m}\hat{x}_i\hat{x}_i^Tp_1(\hat{x}_i;\beta)(1-p_1(\hat{x}_i;\beta))\geq 0$$

所以公式(3.27)为凸函数

3.7

在类别为4时，其可行的编码有7种

	f0	f1	f2	f3	f4	f5	f6
c1	1	1	1	1	1	1	1
c2	0	0	0	0	1	1	1
c3	0	0	1	1	0	0	1
c4	0	1	0	1	0	1	0

当码长为9时，那么 f_6之后加任意两个编码，即为最优编码，因为此时再加任意的编码都是先有编码的反码，此时，类别之间最小的海明距离都为4，不会再增加。

▲在LDA多分类情形下，试计算类间散度矩阵S_b的秩并证明

$$rank(S_b)\leq k-1$$

其中k为属性的数量

因为.

$$S_b=\sum_{j=1}^{k}N_j(\mu_j-\mu)(\mu_j-\mu)^T$$

对于单独的$\mu_j-\mu$，它的秩为1，因此协方差矩阵相加后的秩$\leq k$，但是由于均值的性质，最后一个μ_k能够用前$k-1$个μ_j线性表示，因此

$$rank(S_b) \leq k-1$$

▲给出公式3.45的推导过程

$$\max_W \frac{tr(W^t S_b W)}{tr(W^T S_w W)} = \max_W \frac{\sum w_i^T S_b w_i}{\sum w_i^T S_w w_i}$$

上式为广义瑞利熵，其中w_i表示W只有第i行不为0的矩阵，$W = \sum w_i$

$$w_i^T = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ & & & \cdots & \\ w_{i1} & w_{i2} & w_{i3} & \cdots & w_{in} \\ & & & \cdots & \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}$$

此时最优化问题同式(3.35)，利用拉格朗日乘子法，即等价于求

$$S_b W = \lambda S_w W$$

▲证明$X(X^TX)^{-1}X^T$是投影矩阵，并对线性回归模型从投影角度解释

假设向量b在空间X上的投影为p，则

$$p = \begin{bmatrix} X_1 & X_2 & \dots & X_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} = Xx$$

根据投影的性质，$b-p$应该和空间X垂直

$$X^T(b-p) = 0$$
$$X^T(b-Xx) = 0$$
$$X^Tb = X^TXx$$
$$x = (X^TX)^{-1}X^Tb$$

代回

$$p = X(X^TX)^{-1}X^Tb$$

即

$$X(X^TX)^{-1}X^T$$

为投影矩阵

我们可以将特征矩阵X看作是一个向量组，每一列（特征）都是一个 n 维向量，我们有 d 个这样的向量。我们假设 $d < n$且所有特征都线性无关，那X 张成的空间是个d 维度空间。真实值 y 是一个 $n \times 1$ 的向量，处于 n维空间中。多元线性回归就是在 X 张成的 d 维空间中，寻找 n 维空间中 y 的投影。

4.1

假设不存在与训练集一致的决策树，那么训练集训练得到的决策树至少有一个节点上存在无法划分的多个数据（若节点上没有冲突数据，那么总是能够将数据分开的） 这与前提-不含冲突数据 矛盾，因此必存在与训练集一致的决策树

4.9

给定训练集 D 和属性 a, 令 $\tilde{D}$ 表示 D 中在属性 a 上没有缺失值的样本子集. 对问题(1), 显然我们仅可根据 $\tilde{D}$ 来判断属性 a 的优劣. 假定属性 a 有 V 个可取值 $\{a^1, a^2, \ldots, a^V\}$, 令 $\tilde{D}^v$ 表示 $\tilde{D}$ 中在属性 a 上取值为 a^v 的样本子集, $\tilde{D}_k$ 表示 $\tilde{D}$ 中属于第 k 类 $(k = 1, 2, \ldots, |\mathcal{Y}|)$ 的样本子集, 则显然有 $\tilde{D} = \bigcup_{k=1}^{|\mathcal{Y}|} \tilde{D}_k$, $\tilde{D} = \bigcup_{v=1}^{V} \tilde{D}^v$. 假定我们为每个样本 $\boldsymbol{x}$ 赋予一个权重 $w_{\boldsymbol{x}}$, 并定义

$$\rho = \frac{\sum_{\boldsymbol{x} \in \tilde{D}} w_{\boldsymbol{x}}}{\sum_{\boldsymbol{x} \in D} w_{\boldsymbol{x}}}, \tag{4.9}$$

$$\tilde{p}_k = \frac{\sum_{\boldsymbol{x} \in \tilde{D}_k} w_{\boldsymbol{x}}}{\sum_{\boldsymbol{x} \in \tilde{D}} w_{\boldsymbol{x}}} \quad (1 \leqslant k \leqslant |\mathcal{Y}|), \tag{4.10}$$

$$\tilde{r}_v = \frac{\sum_{\boldsymbol{x} \in \tilde{D}^v} w_{\boldsymbol{x}}}{\sum_{\boldsymbol{x} \in \tilde{D}} w_{\boldsymbol{x}}} \quad (1 \leqslant v \leqslant V). \tag{4.11}$$

将基尼指数推广为:

$$Gini_index(D, a) = \rho \times (\sum_{v=1}^{V} \tilde{r}_v Gini(\tilde{D}^v))$$

$$Gini(\tilde{D}) = 1 - \sum_{k=1}^{|Y|} \tilde{p}_k^2$$

▲ 假设离散随机变量$X \in \{1, \cdots, K\}$，其取值为k的概率$P(X = k) = p_k$，其熵为$H(p) = -\sum_k p_k log_2 p_k$，试用拉格朗日乘子法证明熵最大的分布为均匀分布

写出拉格朗日函数:

$$L(p_1, \ldots, p_K) = -\sum_i p_i log_2 p_i - \lambda(\sum_i p_i - 1)$$

对p_1到p_K求偏导可得方程

$$\begin{cases} -\sum_i p_i log_2 p_i - \lambda(\sum_i p_i - 1) = 0 \\ -log_2 p_i - ln2 - \lambda = 0 \\ \sum_i p_i - 1 = 0 \end{cases}$$

可得到p_i平均分布时，可以得到熵的最大值

▲习题

- 下表表示的二分类数据集，具有三个属性A,B,C，样本标记为两类“+”，“-”。请运用你学过的知识完成如下问题：

实例	A	B	C	类别
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	8.0	-
8	T	F	7.0	+
9	F	T	5.0	-
10	F	F	2.0	+

(a) 整个训练样本关于类属性的熵是多少
(b) 数据集中A，B两个属性的信息增益各是多少
(c) 对于属性C，计算所有可能划分的信息增益
(d) 根据Gini指数，A和B两个属性哪个是最优划分
(e) 采用算法C4.5，构造决策树

(a)

$$Ent(D) = -\frac{5}{10}log\frac{5}{10} - \frac{5}{10}log\frac{5}{10} = 1$$

(b)

$$\begin{aligned} Gain(D,A) &= Ent(D) - [\frac{4}{10}Ent(D_1) + \frac{6}{10}Ent(D_2)] \\ &= 1 - [\frac{4}{10}(-\frac{3}{4}log\frac{3}{4} - \frac{1}{4}log\frac{1}{4}) + \frac{6}{10}(-\frac{2}{6}log\frac{2}{6} - \frac{4}{6}log\frac{4}{6})] \\ &= 0.1245 \end{aligned}$$

$$\begin{aligned} Gain(D,B) &= Ent(D) - [\frac{5}{10}Ent(D_1) + \frac{5}{10}Ent(D_2)] \\ &= 1 - [\frac{5}{10}(-\frac{2}{5}log\frac{2}{5} - \frac{3}{5}log\frac{3}{5}) + \frac{5}{10}(-\frac{3}{5}log\frac{3}{5} - \frac{2}{5}log\frac{2}{5})] \\ &= 0.0290 \end{aligned}$$

(c)

$$t = 1.5 \quad Gain(D,C,t) = 1 + \frac{9}{10}\frac{4}{9}log\frac{4}{9} + \frac{5}{10}\frac{5}{9}log\frac{5}{9} = 0.1080$$

$$t = 2.5 \quad Gain(D,C,t) = 1 + \frac{8}{10}\frac{3}{8}log\frac{3}{8} + \frac{8}{10}\frac{5}{8}log\frac{5}{8} = 0.2365$$

$$t = 3.5 \quad Gain(D,C,t) = 1 + \frac{3}{10}\frac{2}{3}log\frac{2}{3} + \frac{3}{10}\frac{1}{3}log\frac{1}{3} + \frac{7}{10}\frac{3}{7}log\frac{3}{7} + \frac{7}{10}\frac{4}{7}log\frac{4}{7} = 0.0349$$

$$t = 4.5 \quad Gain(D,C,t) = 1 + \frac{4}{10}\frac{3}{4}log\frac{3}{4} + \frac{4}{10}\frac{1}{4}log\frac{1}{4} + \frac{6}{10}\frac{2}{6}log\frac{2}{6} + \frac{6}{10}\frac{4}{6}log\frac{4}{6} = 0.1245$$

$$t = 5.5 \quad Gain(D,C,t) = 0$$

$$t = 7.5 \quad Gain(D,C,t) = 1 + \frac{9}{10}\frac{5}{9}log\frac{5}{9} + \frac{9}{10}\frac{4}{9}log\frac{4}{9} = 0.1080$$

(d)

$$Gini\_index(D,A) = \frac{4}{10}(1 - \frac{3}{4}^2 - \frac{1}{4}^2) + \frac{6}{10}(1 - \frac{2}{6}^2 - \frac{4}{6}^2) = 0.4167$$

$$Gini\_index(D,B)=\frac{5}{10}(1-\frac{2}{5}^2-\frac{3}{5}^2)+\frac{5}{10}(1-\frac{2}{5}^2-\frac{3}{5}^2)=0.48$$

A是最优化分

(e)