HW3 & HW4 Reference

注意：方法不唯一，言之成理即可！

0.1 [课本习题 3.2] 试证明，对于参数 w，对率回归的目标函数 (3.18) 是非凸的，但其对数似然函数 (3.27) 是凸的。

$$y = \frac{1}{1 + \mathrm{e}^{-\boldsymbol{w}^\top \boldsymbol{x} + b}} \tag{3.18}$$

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{m} (-y_i \boldsymbol{\beta}^\top \hat{\boldsymbol{x}}_i + \ln(1 + e^{\boldsymbol{\beta}^\top \hat{\boldsymbol{x}}_i})) \tag{3.27}$$

证明.
- 需要注意，标量函数先对向量变量的转置求导，再对向量变量求导，得到的是矩阵；标量函数先对向量变量求导，再对向量变量的转置求导，得到的是标量。很多同学混淆使用。《矩阵分析与应用（第 2 版)》张贤达
- 方法不唯一，但是需要注意符号书写清晰

$$\frac{\partial y}{\partial \boldsymbol{w}} = \frac{\boldsymbol{x} e^{-\left(\boldsymbol{w}^\top \boldsymbol{x} + b\right)}}{\left(1 + e^{-\left(\boldsymbol{w}^\top \boldsymbol{x} + b\right)}\right)^2} = \boldsymbol{x} y(1-y)$$

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

$\boldsymbol{x}^\top \boldsymbol{x} \geq 0$ 恒成立，当 $0.5 < y < 1$ 时，$y(1-2y)(1-y) < 0$，此时 $\frac{\partial^2 y}{\partial \boldsymbol{w}^\top \partial \boldsymbol{w}} < 0$，因此函数 (3.18) 非凸。

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{m} \left(-y_i \widehat{\boldsymbol{x}}_i + \frac{1}{1 + e^{\boldsymbol{\beta}^\top \widehat{\boldsymbol{x}}_i}} \widehat{\boldsymbol{x}}_i e^{\boldsymbol{\beta}^\top \widehat{\boldsymbol{x}}_i} \right)$$

$$\frac{\partial^2 \ell}{\partial \boldsymbol{\beta}^\top \partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}^\top} \frac{\partial \ell}{\partial \boldsymbol{\beta}} = \frac{\partial}{\partial \boldsymbol{\beta}^\top} \sum_{i=1}^{m} \left(-y_i \widehat{\boldsymbol{x}}_i + \frac{1}{1 + e^{\boldsymbol{\beta}^\top \widehat{\boldsymbol{x}}_i}} \widehat{\boldsymbol{x}}_i e^{\boldsymbol{\beta}^\top \widehat{\boldsymbol{x}}_i} \right) = \sum_{i=1}^{m} \frac{e^{\boldsymbol{\beta}^\top \widehat{\boldsymbol{x}}_i}}{\left(1 + e^{\boldsymbol{\beta}^\top \widehat{\boldsymbol{x}}_i}\right)^2} \widehat{\boldsymbol{x}}_i{}^\top \widehat{\boldsymbol{x}}_i$$

由于 $\widehat{\boldsymbol{x}}_i{}^\top \widehat{\boldsymbol{x}}_i \geq 0$ 且 $\frac{e^{\boldsymbol{\beta}^\top \widehat{\boldsymbol{x}}_i}}{\left(1+e^{\boldsymbol{\beta}^\top \widehat{\boldsymbol{x}}_i}\right)^2} \geq 0$，因此函数 (3.27) 为凸函数。

□

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9
C_1	1	1	1	1	1	1	1	x	x
C_2	0	0	0	0	1	1	1	x	x
C_3	0	0	1	1	0	0	1	x	x
C_4	0	1	0	1	0	1	0	x	x

0.2 [课本习题 3.7] 令码长为 9，类别数为 4，试给出海明距离意义下理论最优的 ECOC 二元码并证明之。

解.
- 任意两个类别间的海明距离足够大
- 任意两个分类器的输出应该尽量独立
- 最优定义 1: 任意两个编码之间的最小距离最大
- 最优定义 2: 任意两个类别之间的海明距离要大，并且任意两个类别间反码的距离最大
- 最优定义 3: 编码间距离和最大

(构造方法以及相关证明可参考文章“Solving Multiclass Learning Problems via Error-Correcting Output Codes”)

采用“Exhaustive Codes”方法构造。当类别为 4，可行编码有 7 种（即 $f_1 - f_7$），f_7 后的任意编码都是之前编码的反码, 因此 f_8、f_9 可以为任意编码。此时，类别间最小海明距离为 4。

□

0.3 在 LDA 多分类情形下，试计算类间散度矩阵 S_b 的秩，并证明

解. 令

$$\boldsymbol{S}_b = \sum_c m_c (\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^\top = \boldsymbol{A}\boldsymbol{M}\boldsymbol{A}^\top$$

其中

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{\mu}_1 - \boldsymbol{\mu} & \boldsymbol{\mu}_2 - \boldsymbol{\mu} & \cdots & \boldsymbol{\mu}_N - \boldsymbol{\mu} \end{pmatrix}, \quad \boldsymbol{M} = \operatorname{diag}(m_1, m_2, \cdots, m_N).$$

接着，可以得到

$$\operatorname{rank} \boldsymbol{S}_b = \operatorname{rank}\left(\boldsymbol{A}\boldsymbol{M}\boldsymbol{A}^\top\right) = \operatorname{rank}\left(\left(\boldsymbol{A}\boldsymbol{M}^{\frac{1}{2}}\right)\left(\boldsymbol{A}\boldsymbol{M}^{\frac{1}{2}}\right)^\top\right) = \operatorname{rank}\left(\boldsymbol{A}\boldsymbol{M}^{\frac{1}{2}}\right) = \operatorname{rank} \boldsymbol{A}$$

因为 $\sum_c m_i(\boldsymbol{\mu}_i - \boldsymbol{\mu}) = \mathbf{0}$，所以 $\operatorname{rank} \boldsymbol{S}_b = \operatorname{rank} \boldsymbol{A} \le N - 1$.

□

0.4 给出公式 (3.45) 的推导公式。

$$\mathbf{S}_b \mathbf{W} = \lambda \mathbf{S}_w \mathbf{W} \tag{3.45}$$

解. 问题:

$$\max_{\mathbf{W}} \frac{\operatorname{tr}\left(\mathbf{W}^\top \mathbf{S}_b \mathbf{W}\right)}{\operatorname{tr}\left(\mathbf{W}^\top \mathbf{S}_w \mathbf{W}\right)} \tag{3.44}$$

令上式分母为 1，则可上述问题转化为

$$\begin{aligned} \min_{\mathbf{W}} \quad & -\operatorname{tr}\left(\mathbf{W}^\top \mathbf{S}_b \mathbf{W}\right) \\ \text{s.t.} \quad & \operatorname{tr}\left(\mathbf{W}^\top \mathbf{S}_w \mathbf{W}\right) = 1 \end{aligned} \tag{1}$$

引入拉格朗日乘子法：

$$L(\mathbf{W}, \lambda) = -\operatorname{tr}\left(\mathbf{W}^\top \mathbf{S}_b \mathbf{W}\right) + \lambda\left(\operatorname{tr}\left(\mathbf{W}^\top \mathbf{S}_w \mathbf{W}\right) - 1\right) \tag{2}$$

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{W}} &= -\frac{\partial \operatorname{tr}\left(\mathbf{W}^\top \mathbf{S}_b \mathbf{W}\right)}{\partial \mathbf{W}} + \lambda \frac{\partial \operatorname{tr}\left(\mathbf{W}^\top \mathbf{S}_w \mathbf{W}\right)}{\partial \mathbf{W}} \\ &= -\left(\mathbf{S}_b + \mathbf{S}_b^\top\right)\mathbf{W} + \lambda\left(\mathbf{S}_w + \mathbf{S}_w^\top\right)\mathbf{W} \\ &= -2\mathbf{S}_b\mathbf{W} + 2\lambda\mathbf{S}_w\mathbf{W} \end{aligned} \tag{3}$$

令 $\frac{\partial L}{\partial \mathbf{W}} = 0$，即可得到公式 (3.45)。 □

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

证明. 令 $P = X(X^\top X)^{-1}X^\top$，那么

$$P^\top = \left(X(X^\top X)^{-1}X^\top\right)^\top = X\left(X(X^\top X)^{-1}\right)^\top = X(X^\top X)^{-1}X^\top = P$$

因此 P 是一个对称矩阵，又因为

$$P^2 = X(X^\top X)^{-1}X^\top X(X^\top X)^{-1}X^\top = X(X^\top X)^{-1}(X^\top X)(X^\top X)^{-1}X^\top = X(X^\top X)^{-1}X^\top = P$$

因此 P 是一个幂等矩阵，所以 P 是一个投影矩阵。 □

解释. 线性回归模型：$\hat{\boldsymbol{y}} = X^\top(X^\top X)^{-1}X^\top\boldsymbol{y}$。可以发现，$\hat{\boldsymbol{y}}$ 其实是 $\boldsymbol{y}$ 在线性空间的投影。 □

1 HW4

1.1 [课本习题 4.1] 试证明对于不含冲突数据（即特征向量完全相同但标记不同）的训练集，必存在与训练集一致（即训练误差为 0）的决策树。

证明. (反证法) 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必然含有冲突数据，这与假设矛盾，因此必然存在与训练集一致决策树。 □

1.2 [课本习题 4.9] 试将 4.4.2 节对缺失值的处理机制推广到基尼指数的计算中去。

解.

$$\begin{aligned} \operatorname{Gini}(D, a) &= \rho \times \operatorname{Gini\_index}(\tilde{D}, a) \\ &= \rho \times \sum_{v=1}^{V} \tilde{r}_v \operatorname{Gini}(\tilde{D}^v) \\ &= \rho \times \sum_{v=1}^{V} \tilde{r}_v \left(1 - \sum_{i=1}^{k} \tilde{p}_k^2\right) \end{aligned}$$

□

1.3 **假设离散随机变量 $X \in \{1, \ldots, K\}$，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分布为均匀分布。**

证明.

$$L(p, \lambda) = -\sum_{i=1}^{k} p_i \log_2 p_i + \lambda \left(\sum_{i=1}^{k} p_i - 1 \right)$$

$$\frac{\partial L}{\partial p_i} = -\log_2 p_i - \frac{1}{\ln 2} + \lambda = 0 \Longrightarrow p_1 = p_2 = \cdots = p_k = 2^{\lambda - \frac{1}{\ln 2}}$$

$$\frac{\partial L}{\partial \lambda} = \sum_{i=1}^{k} p_i - 1 = 0 \Longrightarrow p_1 = p_2 = \cdots = p_k = \frac{1}{k}$$

□

1.4 **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	+

1.4.1 整个训练样本关于类属性的熵是多少?

解. 类别 $+$ 的概率为 $p^+ = \frac{5}{10}$，类别 $-$ 的概率为 $p^- = \frac{5}{10}$，因此熵为

$$\mathrm{Ent}(D) = -p^+ \log_2 p^+ - p^- \log_2 p^- = 1.$$

□

1.4.2 数据集中 A、B 两个属性的信息增益各是多少?

解.

$$\begin{aligned}
\mathrm{Gain}(D, A) &= \mathrm{Ent}(D) - \sum_v \frac{|D^v|}{|D|} \mathrm{Ent}\left(D^v\right) \\
&= 1 - \left(\frac{4}{10} \left(-\frac{3}{4} \log_2 \frac{3}{4} - \frac{1}{4} \log_2 \frac{1}{4} \right) + \frac{6}{10} \left(-\frac{4}{6} \log_2 \frac{4}{6} - \frac{2}{6} \log_2 \frac{2}{6} \right) \right) \\
&= 0.125
\end{aligned}$$

$$
\begin{aligned}
\operatorname{Gain}(D, B) &= \operatorname{Ent}(D)-\sum_{v} \frac{\left|D^{v}\right|}{|D|} \operatorname{Ent}\left(D^{v}\right) \\
&= 1-\left(\frac{5}{10}\left(-\frac{2}{5} \log _{2} \frac{2}{5}-\frac{3}{5} \log _{2} \frac{3}{5}\right)+\frac{5}{10}\left(-\frac{2}{5} \log _{2} \frac{2}{5}-\frac{3}{5} \log _{2} \frac{3}{5}\right)\right) \\
&= 0.029
\end{aligned}
$$

□

1.4.3 对于属性 C，计算所有可能划分的信息增益?

解. 可取划分点为 $\{1.5,2.5,3.5,4.5,5.5,6.5,7.5\}$，然后对应划分信息增益为:

$$
\begin{aligned}
&\operatorname{Gain}(D, C, 1.5)=1-\tfrac{9}{10}\left(-\tfrac{4}{9} \log _{2} \tfrac{4}{9}-\tfrac{5}{9} \log _{2} \tfrac{5}{9}\right) \approx 0.108 \\
&\operatorname{Gain}(D, C, 2.5)=1-\tfrac{8}{10}\left(-\tfrac{3}{8} \log _{2} \tfrac{3}{8}-\tfrac{5}{8} \log _{2} \tfrac{5}{8}\right) \approx 0.236 \\
&\operatorname{Gain}(D, C, 3.5)=1-\left[\tfrac{3}{10}\left(-\tfrac{2}{3} \log _{2} \tfrac{2}{3}-\tfrac{1}{3} \log _{2} \tfrac{1}{3}\right)+\tfrac{7}{10}\left(-\tfrac{3}{7} \log _{2} \tfrac{3}{7}-\tfrac{4}{7} \log _{2} \tfrac{4}{7}\right)\right] \approx 0.035 \\
&\operatorname{Gain}(D, C, 4.5)=1-\left[\tfrac{4}{10}\left(-\tfrac{3}{4} \log _{2} \tfrac{3}{4}-\tfrac{1}{4} \log _{2} \tfrac{1}{4}\right)+\tfrac{6}{10}\left(-\tfrac{2}{6} \log _{2} \tfrac{2}{6}-\tfrac{4}{6} \log _{2} \tfrac{4}{6}\right)\right] \approx 0.125 \\
&\operatorname{Gain}(D, C, 5.5)=1-\left[\tfrac{6}{10}\left(-\tfrac{3}{6} \log _{2} \tfrac{3}{6}-\tfrac{3}{6} \log _{2} \tfrac{3}{6}\right)+\tfrac{4}{10}\left(-\tfrac{2}{4} \log _{2} \tfrac{2}{4}-\tfrac{2}{4} \log _{2} \tfrac{2}{4}\right)\right]=0 \\
&\operatorname{Gain}(D, C, 6.5)=1-\left[\tfrac{7}{10}\left(-\tfrac{4}{7} \log _{2} \tfrac{4}{7}-\tfrac{3}{7} \log _{2} \tfrac{3}{7}\right)+\tfrac{3}{10}\left(-\tfrac{1}{3} \log _{2} \tfrac{1}{3}-\tfrac{2}{3} \log _{2} \tfrac{2}{3}\right)\right] \approx 0.035 \\
&\operatorname{Gain}(D, C, 7.5)=1-\tfrac{9}{10}\left(-\tfrac{5}{9} \log _{2} \tfrac{5}{9}-\tfrac{4}{9} \log _{2} \tfrac{4}{9}\right) \approx 0.108
\end{aligned}
$$

□

1.4.4 根据 Gini 指数，A 和 B 两个属性哪个是最优划分?

解.

$$
\begin{aligned}
\text{Gini\_index}(D, A) &= \frac{4}{10}\left(1-\left(\frac{3}{4}\right)^{2}-\left(\frac{1}{4}\right)^{2}\right)+\frac{6}{10}\left(1-\left(\frac{2}{6}\right)^{2}-\left(\frac{4}{6}\right)^{2}\right) \\
&= 0.417 \\
\text{Gini\_index}(D, B) &= \frac{5}{10}\left(1-\left(\frac{2}{5}\right)^{2}-\left(\frac{3}{5}\right)^{2}\right)+\frac{5}{10}\left(1-\left(\frac{3}{5}\right)^{2}-\left(\frac{2}{5}\right)^{2}\right) \\
&= 0.48
\end{aligned}
\tag{4}
$$

因此，A 是最优划分。

□

1.4.5 采用算法 C4.5，构造决策树。

解.
- 指标：信息增益率，不是信息增益
- 构造方法和构造结果不唯一，建议大家构造前简述自己的构造方法，可以拿一半的过程分。

初始：$D=\{1,2,3,4,5,6,7,8,9,10\}$

第一层划分:

$$
\text{Gain\_ratio}(D, A)=\frac{\operatorname{Gain}(D, A)}{\operatorname{IV}(D, A)}=\frac{0.125}{0.971}=0.129
$$

$$\text{Gain\_ratio}(D,B)=\frac{\text{Gain}(D,B)}{\text{IV}(D,B)}=\frac{0.029}{1}=0.029$$

$$\text{Gain\_ratio}(D,C,2.5)=\frac{\text{Gain}(D,C,2.5)}{\text{IV}(D,C,2.5)}=\frac{0.236}{0.722}=0.326$$

选择 $(C,2.5)$ 作为 D 的划分，得到 $D^1_{c\leq 2.5}=\{1,10\}$，$D^1_{c>2.5}=\{2,3,4,5,7,8,9\}$。

第二层划分: 因为 $D^1_{c\leq 2.5}$ 元素类别一致，不再划分，主要针对 $D^1_{c>2.5}$ 继续划分。

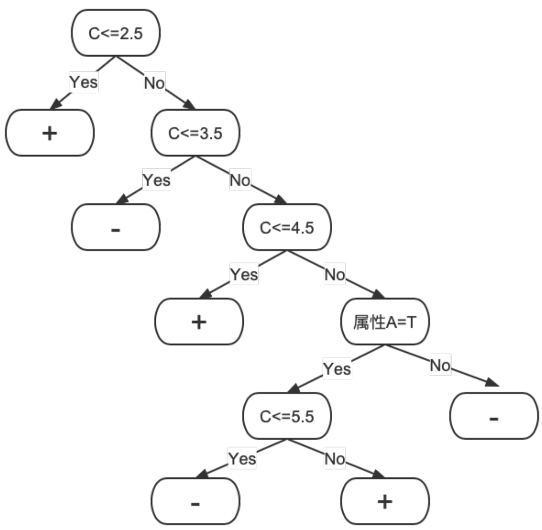
$$\text{Gain\_ratio}(D^1_{c>2.5},A)=\frac{\text{Gain}(D^1_{c>2.5},A)}{\text{IV}(D^1_{c>2.5},A)}=\frac{0.159}{0.954}=0.167$$

$$\text{Gain\_ratio}(D^1_{c>2.5},B)=\frac{\text{Gain}(D^1_{c>2.5},B)}{\text{IV}(D^1_{c>2.5},B)}=\frac{0.049}{1}=0.049$$

$$\text{Gain\_ratio}(D^1_{c>2.5},C,3.5)=\frac{\text{Gain}(D^1_{c>2.5},C,3.5)}{\text{IV}(D^1_{c>2.5},C,3.5)}=\frac{0.092}{0.544}=0.169$$

选择 $(C,3.5)$ 作为 $D^1_{c>2.5}$ 的划分，得到 $D^2_{2.5<c\leq 3.5}=\{6\}$，$D^2_{c\geq 3.5}=\{2,3,4,5,7,8,9\}$。

第三层划分: 以此类推...



$\square$