Machine Learning Homework 3.1&3.2

专业：软件工程　　　　　姓名：沈金龙　　　　　学号：18214806

1. 作业题目

Hw 3.1

设 x 为一个 d 维的二值向量（即其分量取值为 0 或 1），服从多维伯努利分布

$$p(x_i \mid \Theta) = \prod_{i=1}^{d} \Theta_i^{x_i} (1-\Theta_i)^{1-x_i}$$

其中 $\Theta = (\Theta_1, \cdots, \Theta_d)^T$ 是未知的参数向量，而 Θ_i 为 x_i=1 的概率。证明，对于 Θ 的最大似然估计为

$$\hat{\Theta} = \frac{1}{n}\sum_{k=1}^{n} X_k$$

Hw 3.2

令 $D = \{x_1, \cdots, x_n\}$ 为 n 个独立的已标记的集合。令 $D_k(x) = \{x_1', \cdots, x_k'\}$ 为样本 x 的 k 个最邻近。根据 k-近邻规则，x 将归入 $D_k(x)$ 中出现次数最多的那个类别。考虑一个 2 类别问题，先验概率为 $P(\omega_1) = P(\omega_2) = 1/2$ 。进一步假设类条件概率密度 $P(X \mid \omega_i)$ 在 10 单位超球体内为均匀分布。

（a）证明如果 k 为奇数，那么平均误差率为

$$P_n(e) = \frac{1}{2^n}\sum_{j=0}^{(k-1)/2} \binom{n}{j}$$

（b）证明在这种情况下，如果 k>1，那么最近邻规则比 k-近邻规则有更低的误差率。

（c）如果 k 随着 n 的增加而增加，同时又受 $k < a\sqrt{n}$ 的限制，那么证明：

当 $n \to \infty$ 时 $P_n(e) \to 0$ 。

2. 答案

1） HW3.1

我们有 n 个样本 $\{x_1, \ldots, x_n\}$ 服从离散分布：

$$P(x \mid \theta) = \prod_{i=1}^{d} \theta_i^{x_i} (1-\theta_i)^{1-x_i}$$

n 个样本的特定序列的似然为：

$$P(x_1,...,x_n\,|\,\theta)=\prod_{k=1}^{n}\prod_{i=1}^{d}\theta_i^{x_{ki}}(1-\theta_i)^{1-x_{ki}}$$

对数似然函数为：

$$l(\theta)=\sum_{k=1}^{n}\sum_{i=1}^{d}x_{ki}\ln\theta_i+(1-x_{ki})\ln(1-\theta_i)$$

为了找到$l(\theta)$的最大值，我们假设$\nabla_\theta l(\theta)=0$再逐项求值$(i=1,...,d)$得到：

$$\begin{aligned}&[\nabla_\theta l(\theta)]_i=\nabla_{\theta_i}l(\theta)\\&=\frac{1}{\theta_i}\sum_{k=1}^{n}-\frac{1}{1-\theta_i}\sum_{k=1}^{n}(1-x_{ki})\\&=0.\end{aligned}$$

这意味着对于任意 i 有：

$$\frac{1}{\hat{\theta}_i}\sum_{k=1}^{n}x_{ki}=\frac{1}{1-\hat{\theta}_i}\sum_{k=1}^{n}(1-x_{ki})$$

上式可以被重写为：

$$(1-\hat{\theta}_i)\sum_{k=1}^{n}x_{ki}=\hat{\theta}_i(n-\sum_{k=1}^{n}x_{ki}).$$

即最终的结果为：

$$\hat{\theta}_i=\frac{1}{n}\sum_{k=1}^{n}x_{ki}.$$

由于该结果适用于所有的 i=1,...,d，我们可以将最后的这等式写成向量的形式：

$$\hat{\theta}_i=\frac{1}{n}\sum_{k=1}^{n}x_k.$$

因此θ的最大似然值仅仅是样本均值，正如我们所期望的。

2） HW3.2

我们已知$P(\omega_1)=P(\omega_2)=1/2$ ，且：

$$P(\omega_1 | x) = \begin{cases} 1, if \ \|x\| \leq 1 \\ 0, otherwise \end{cases}$$

$$P(\omega_2 | x) = \begin{cases} 1, if \ \|x-a\| \leq 1 \\ 0, otherwise \end{cases}$$

我们认为，不失一般性，这一类ω1 以原点 0 为中心，而类别ω2 以 a 为中心，是一个原点以外的点。

（a） 我们采用$P_n(e)$表示 n 个点的平均误差率，则：

$$\begin{aligned} P_n(e) &= \Pr[\text{true category is } \omega_1 \text{ while } \omega_2 \text{ is most frequently labeled}] \\ &\quad + \Pr[\text{true category is } \omega_2 \text{ while } \omega_1 \text{ is most frequently labeled}] \\ &= 2\Pr[\text{true category is } \omega_1 \text{ while } \omega_2 \text{ is most frequently labeled}] \\ &= 2P(\omega_1)\Pr[\text{label of } \omega_1 \text{ for fewer than } (k-1)/2 \text{points, and the rest labeled } \omega_2] \\ &= 2\frac{1}{2}\sum_{j=0}^{(k-1)/2} \Pr[j \text{ of } n \text{ chosen points are labeled } \omega_1, \text{the rest } \omega_2] \\ &= \sum_{j=0}^{(k-1)/2} \binom{n}{j} \frac{1}{2^j}\frac{1}{2^{(n-j)}} \\ &= \frac{1}{2^n}\sum_{j=0}^{(k-1)/2} \binom{n}{j}. \end{aligned}$$

（b） 我们明确 k 依赖于概率，$P_n(e) = P_n(e;k)$，然后有：

$$P_n(e;1) = \frac{1}{2^n} < P_n(e;k) = \frac{1}{2^n} \underbrace{\sum_{j=0}^{(k-1)/2} \binom{n}{j}}_{>0 \ for \ k>1}$$

（c） 在这种情况下：

$$P_n(e) = \frac{1}{2^n}\sum_{j=0}^{(k-1)/2}\binom{n}{j} = \Pr[B(n,1/2) \leq (k-1)/2]$$

$$= \Pr[\mathrm{Y}_1 + ... + \mathrm{Y}_n \leq \frac{k-1}{2}]$$

其中$Y_1,...Y_n$是独立的，$B(.,.)$为非正态分布，$\Pr[Y_i = 1] = \Pr[Y_i = 0] = 1/2$，如果 k 增加到 n，但由$k < a/\sqrt{n}$限制，则：

$$\begin{aligned} P_n(e) &\leq \Pr\left(Y_1 + \cdots + Y_n \leq \frac{a/\sqrt{n} - 1}{2}\right) \\ &= Pr(Y_1 + \cdots + Y_n \leq 0) \text{ for } n \text{ sufficiently large} \\ &= 0, \end{aligned}$$

这也说明当$n \to \infty$时，$P_n(e) \to 0$。