这个不等式是 Azuma 鞅不等式的一个推论，下面的证明不用复杂的理论。以后再补上随机过程中的证明。

从 wikipedia 摘抄的。注意，markov 不等式中的 y 是 x，不等式右边的 E(X)，换成 E(|X|)。证明过程假设 X 是非负随机变量

Hoeffding 不等式如下：

对于任意 t > 0，都有

$$P(S_n - E[S_n] \geq t) \leq exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

$$P(E[S_n] - S_n \geq t) \leq exp\left(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

先介绍下Markov不等式:

$$p(|X| \geq \epsilon) \leq \frac{E(X)}{\epsilon}$$

证明如下:

$$E(X) = \int_0^\infty yp(x)\,dx \geq \int_\epsilon^\infty yp(x)\,dx$$

因为对于$x \in [\epsilon, \infty]$都有$x \geq \epsilon$,所以:

$$E(X) \geq \epsilon \int_\epsilon^\infty p(x)dx = \epsilon P(x \geq \epsilon)$$

即

$$p(|X| \geq \epsilon) \leq \frac{E(X)}{\epsilon}$$

对于离散情况，把积分变成求和即可。

在证明Hoeffding不等式之前，要用到一个Hoeffding的一个引理:

对于一个随机变量X，$P(X \in [a,b]) = a, E(X) = 0$,有

$$E(e^{sX}) \leq e^{\frac{1}{8}s^2(b-a)^2}$$

注意到e^{sX}是关于X的一个凸函数和条件$E(X) = 0$有:

$$e^{sX} \leq \frac{b-X}{b-a}e^{sb} + \frac{X-a}{b-a}e^{sa}$$

两边对X取期望:

$$\begin{aligned} E(e^{sX}) \leq \frac{b-E(X)}{b-a}e^{sa} + \frac{E(X)-a}{b-a}e^{sb} &= \frac{b}{b-a}e^{sa} - \frac{a}{b-a}e^{sb} \\ &= (-\frac{a}{b-a})e^{sa}(-\frac{b}{a} + e^{sb-sa}) \end{aligned}$$

令$\theta = -\frac{a}{b-a} > 0$,上式变成了：

$$\theta e^{-s\theta(b-a)}(\frac{1}{\theta} - 1 + e^{s(b-a)}) = (1 - \theta + \theta e^{s(b-a)})e^{-s\theta(b-a)}$$

令 $u = s(b-a)$定义:

$$\begin{cases} \varphi : R \to R \\ \varphi(u) = -\theta u + log(1 - \theta + \theta e^u) \end{cases}$$

它是良定义的，因为由$e^u \geq 0, a < 0, b > 0$和 $\theta > 0$有：

$$1 - \theta + \theta e^u = \theta(\frac{1}{\theta} - 1 + e^u) = \theta(-\frac{a}{b} + e^u) > 0$$

由定义知：

$$E(e^{sX}) \leq e^{\varphi(u)}$$

根据泰勒展开,和泰勒中值定理,存在一个$v \in [0, u]$使得：

$$\varphi(u) = \varphi(0) + u\varphi'(0) + \frac{1}{2}u^2\varphi''(v)$$

意识到：

$$\varphi(0) = 0$$

$$\varphi'(0) = -\theta + \frac{\theta e^u}{1 - \theta + \theta e^u}|_{u=0} = 0$$

$$\varphi''(v) = \frac{\theta e^v(1 - \theta + \theta e^v) - \theta^2 e^{2v}}{(1 - \theta + \theta e^v)^2} = \frac{\theta e^v}{1 - \theta + \theta e^v}(1 - \frac{\theta e^v}{1 - \theta + \theta e^v}) = t(1 - t) \leq 1/4$$

其中$t = \frac{\theta e^v}{1 - \theta + \theta e^v} > 0.$

因此$\varphi(u) \leq 0+0+\frac{1}{2}u^2 * \frac{1}{4} = \frac{1}{8}u^2 = \frac{1}{8}s^2(b-a)^2$ 这便证明了引理

$$E(e^{sX}) \leq e^{\frac{1}{8}s^2(b-a)^2}$$

接下来便容易得多了。

对于$X_1, X_2 ... X_n, n$个随机变量，其中 $P(X_i \in [a_i, b_i]) = 1, 1 \leq i \leq n$

令

$$S_n = \sum_{i=1}^{n} X_i$$

根据Markov不等式，有：

$$\begin{aligned}
P(S_n - E[S_n] \geq t) &= P(e^{s(S_n - E[S_n])} \geq e^s t) \\
&\leq e^{-st} E[e^{s(S_n - E[S_n])}] \\
&= e^{-st} \prod_{i=1}^{n} P(e^{s(X_i - E[X_i])}) \\
&\leq e^{-st} \prod_{i=1}^{n} e^{\frac{s^2 (b_i - a_i)^2}{8}} \\
&= exp(-st + \frac{1}{8} s^2 \sum_{i=1}^{n} (b_i - a_i)^2)
\end{aligned}$$

上面的推导都只假定$s > 0$,定义：

$$\begin{cases}
g : R_+ \to R \\
g(s) = -st + \frac{1}{8} s^2 \sum_{i=1}^{n} (b_i - a_i)^2
\end{cases}$$

求$g'(s) = 0$得到$s = \frac{4t}{\sum_{i=1}^{n} (b_i - a_i)^2}$,代入不等式，即可得到：

$$P(S_n - E[S_n] \geq t) \leq exp\left(-\frac{2t^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right)$$

取$S = -S_n$即可得到

$$P(E[S_n] - S_n \geq t) \leq exp\left(-\frac{2t^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right)$$

1.2 原题：通过经验最小化推导极大似然估计，证明模型是条件概率分布，当损失函数是对数损失函数时，经验最小化等价于极大似然估计。

以下答案写起来比较简洁清晰。但存在一个问题，因为两个目标函数实际上不是相等的，不能直接划等号，但他们在找到最优参数的效果上是相等的，所以称为等价。而其它方式说明等价正确性没问题，但写起来会比较麻烦。

要说明需要用到样本独立同分布的假设，而题目中没有提到这一点（我认为这是题目不严谨的地方）。需要自己假定样本独立同分布。

解答 根据定义，条件概率分布模型下，当损失函数为对数损失函数时，经验风险最小化目标函数为

$$\min \frac{1}{N}\sum_{i=1}^{N} -\log P(y_i|x_i,\boldsymbol{\theta}) \tag{1}$$

假设$(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)$独立同分布，有极大似然估计的目标函数为：

$$\mathrm{p}(y_1,y_2,\ldots,y_N|x_1,x_2,\ldots,x_N,\theta)=\frac{\mathrm{p}(y_1,y_2,\ldots,y_N,x_1,x_2,\ldots,x_N,\theta)}{\mathrm{p}(x_1,x_2,\ldots,x_N,\theta)}=\frac{\prod_{i=1}^{N}\mathrm{p}(y_i,x_i,\theta)}{\prod_{i=1}^{N}\mathrm{p}(x_i,\theta)}=\prod_{i=1}^{N}p(y_i|x_i,\theta)$$

其中 $\boldsymbol{\theta}\in\mathbb{R}^n$ 表示模型参数.

那么由模型求最优解的过程可以得到

$$\begin{aligned}\hat{\boldsymbol{\theta}}&=\arg\min_{\boldsymbol{\theta}\in\mathbb{R}^n}\frac{1}{N}\sum_{i=1}^{N}-\log P(y_i|x_i,\boldsymbol{\theta})\\&=\arg\min_{\boldsymbol{\theta}\in\mathbb{R}^n}\sum_{i=1}^{N}-\log P(y_i|x_i,\boldsymbol{\theta})\\&=\arg\max_{\boldsymbol{\theta}\in\mathbb{R}^n}\sum_{i=1}^{N}\log P(y_i|x_i,\boldsymbol{\theta})\\&=\arg\max_{\boldsymbol{\theta}\in\mathbb{R}^n}\prod_{i=1}^{N}P(y_i|x_i,\boldsymbol{\theta})\end{aligned}$$

由此可见在题设条件下，经验风险最小化等价于极大似然估计.

2.3 题目：证明一下定理：样本集线性可分的充分必要条件是正实例点集和负实例点集所构成的凸壳互不相交。

这里给出比较精确的数学证明，主要参考凸优化相关理论

充分性证明：

若样本线性可分，令正实例点为X_i、凸壳为C_x，负实例点为Y_i、凸壳为C_y。由题意得：

$$\forall i,\ \omega \cdot x_i + b > 0$$

∴ 对于凸包C_x来说有：

$$\begin{aligned}\omega \cdot x + b &= \omega \cdot (\Sigma_{i=1}^{k} \lambda_i x_i) + (\Sigma_{i=1}^{k} \lambda_i) \cdot b \\ &= \Sigma_{i=1}^{k} \lambda_i \omega_i x_i + (\Sigma_{i=1}^{k} \lambda_i) \cdot b \\ &= \Sigma_{i=1}^{k} \lambda_i (\omega_i x_i + b) > 0\end{aligned}$$

∴ 凸包C_x位于超平面$\omega \cdot x + b > 0$的一侧，同理可证凸包C_y位于超平面$\omega \cdot x + b < 0$的另一侧。

必要性证明：

* 先证凸壳具有凸性：对于凸壳内任意两点x_1，x_2组成的线段$\lambda_1 x_1 + \lambda_2 x_2 (\lambda_1 + \lambda_2 = 1)$ 也在凸壳内，即凸壳具有凸性。

∵ 两个凸壳互不相交,可令：

$$d = \|x_0 - y_0\|_2 = inf_{x \in C_X,\ y \in C_y} \|x - y\|_2 \quad p = (x_0 + y_0)/2$$

以$\overrightarrow{(x_0 - y_0)}$为法向量，包含点$p$做一超平面

下用反证法证明凸壳C_x在超平面的一侧，并壳内任意点到超平面的距离均大于d/2：

若点集分布在超平面两侧，分别取两侧一点x_1，x_2并连接两点，上证明凸壳内两点连线线段也必在凸壳内

∴ 两点连线与超平面的交点也属于凸壳，但该交点到超平面的距离为零，矛盾。

以点p为圆心半径$r = d/2$做一超球, 如图1.1：

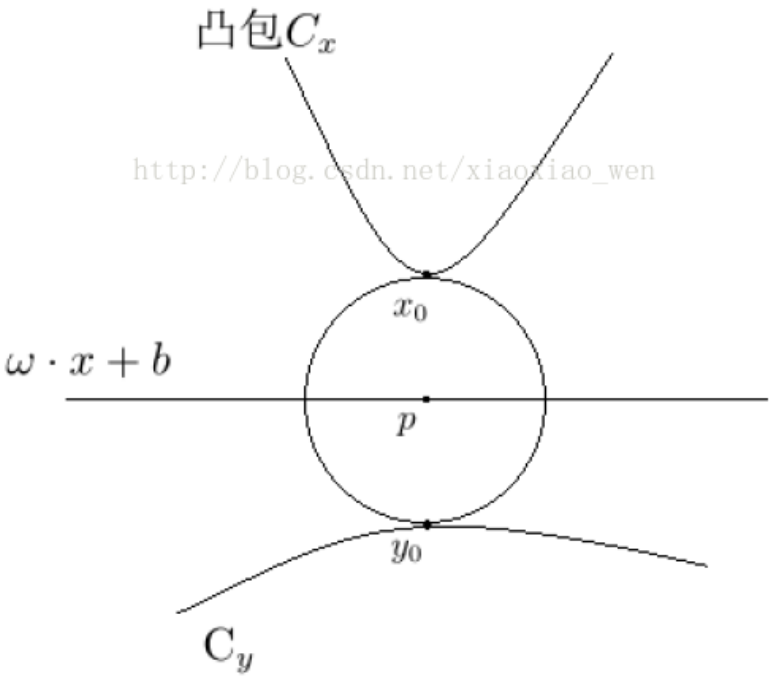

图1.1

(i) 若点x_p使得$\|x_p - p\|_2 > d/2$，且在球内或球上，连接点$x_p y_0$，$y_0 p$，$x_p p$。(图1.2)

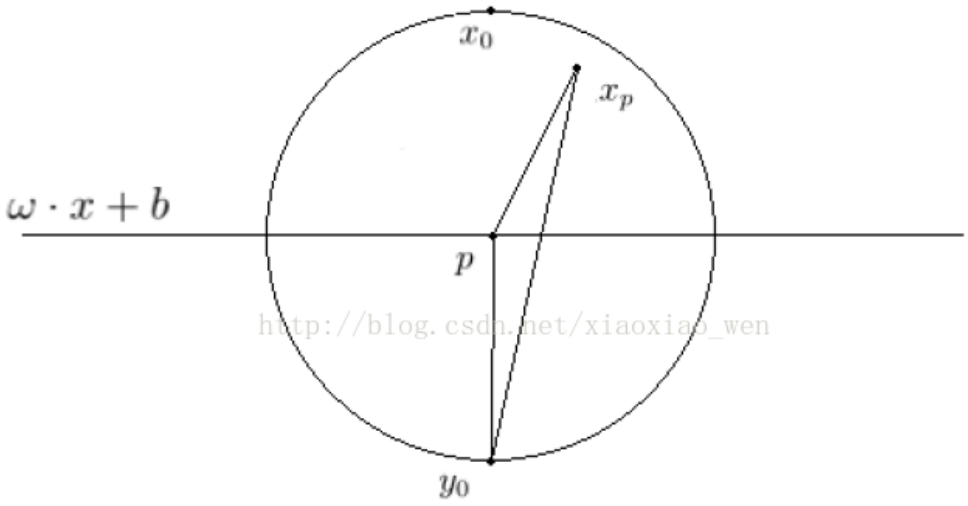

图1.2

显然$\|x_p - y_0\|_2 < \|y_0 - p\|_2 + \|x_p - p\|_2 < d/2 + d/2 = d$，矛盾

(ii) 若点x_p在球外，连接$x_p x_0$交球与点x_m再连接$x_m y_0$，$y_0 p$，$x_m p$。如图(1.3)

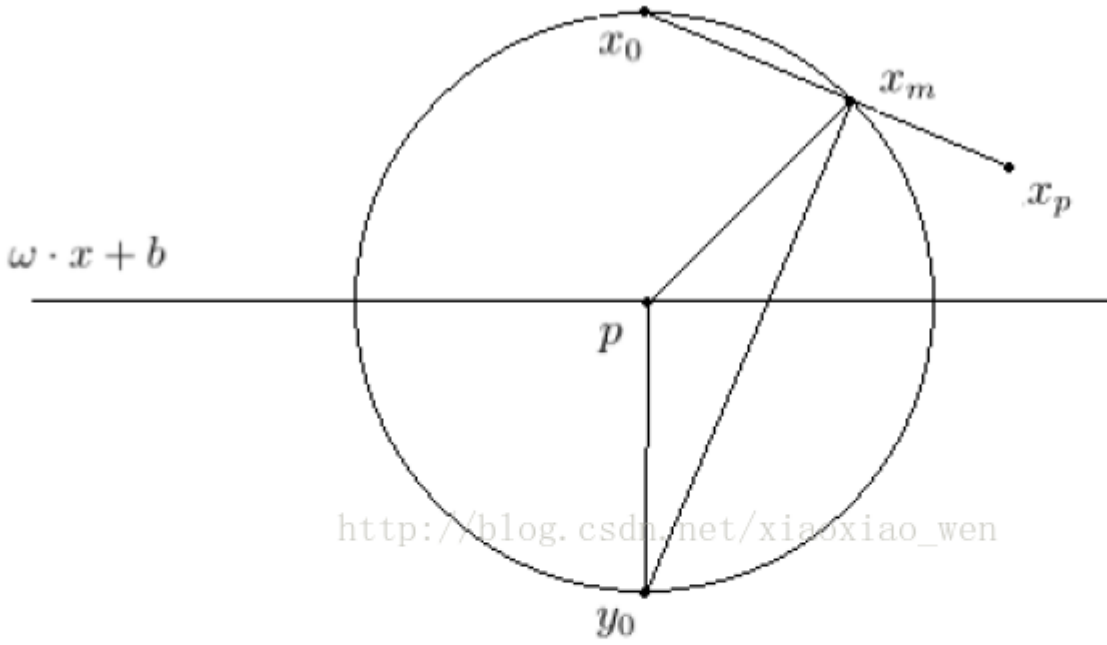

图1.3

∵ 凸壳是凸性的，∴ $x_m \in C_x$，(i)证球上的点到y_0的距离小于
盾。

即证其必要性。