机器学习
Machine learning

第八章 信息论模型
Information-Theoretic Models

授课人：周晓飞
zhouxiaofei@iie.ac.cn
2021-11-26

第八章 信息论模型

8.1 概述

8.2 熵、最大熵

8.3 互信息

8.4 信息论优化模型

第八章 信息论模型

8.1 概述

8.2 熵、最大熵

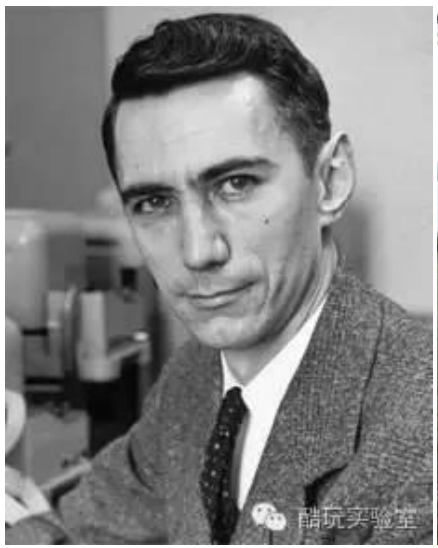
8.3 互信息

8.4 信息论优化模型

概述

香农（1916-2001）

美国数学家、信息论的创始人。1948 年发表的经典论文《通讯的数学原理》， 为信息论奠定了基础。信息论正是关于通信过程本质的深刻数学理论。这个理论提供一个对根本问题研究的总体框架，例如信息表示的效率以及通信信道可靠信息传输的极限问题。而且该理论包括很多有力的定理用以计算最佳表示和信号所携带信息的传输的理想界限。



[1] 杨义先博客--香农外传： http://blog.sciencenet.cn/blog-453322-978153.html

概述

信息论对机器学习的启发

- **熵：**

  不确定性的度量，类别不均匀程度的度量

- **最大熵：**

  一种状态的平衡分布，可看作一种自然法则

- **互信息：**

  随机变量相关性的度量

概述

本章主要内容

- **信息熵的相关定义：**熵、条件熵、联合熵、相对熵、互信息
- **最大熵模型：**求取类别后验概率分布 $P(y|x)$
- **最小互信息模型：**例如，独立成分分析

第八章 信息论模型

8.1 概述

8.2 熵、最大熵

8.3 互信息

8.4 信息论优化模型

熵、最大熵

信息熵

信息量（信息增益量）定义

$$I(x_k)=\log\left(\frac{1}{p_k}\right)=-\log p_k$$

$$X=\{x_k\,|k=0,\pm1,...,\pm N\},\quad p_k=P(X=x_k),\quad 0\le p_k\le 1,\quad \sum_{k=-N}^{N}p_k=1$$

信息量性质：概率越小的状态，信息量越大

$p_k=1,\quad I(x_k)=0;$

$0\le p_k\le 1,\quad I(x_k)\ge 0$

$p_k<p_i,\quad I(x_k)>I(x_i)$

熵、最大熵

信息熵

信息熵定义

信息量在全部数值域上的概率平均值。

- 离散熵：

$$H(X)=E[I(x)]=\sum_{k=-N}^{k=N} p_k I(x_k)=-\sum_{k=-N}^{k=N} p_k \log p_k$$

- 微分熵：

$$h(X)=-\int_{-\infty}^{\infty} p_x(x)\log p_x(x)dx=-E[\log p_x(x)]$$

微分熵不是严格意义的信息熵。

熵、最大熵

信息熵

微分熵推导

$$H(X) = -\lim_{\delta x \to 0} \sum_{k=-\infty}^{\infty} p_X(x_k)\delta x \log(p_X(x_k)\delta x)$$

$$\log(p_x(x_k)\delta x) = \log p_x(x_k) + \log \delta x$$

$$= -\lim_{\delta x \to 0} \left[\sum_{k=-\infty}^{\infty} p_X(x_k)(\log p_X(x_k))\delta x + \log \delta x \sum_{k=-\infty}^{\infty} p_X(x_k)\delta x \right]$$

$$= -\int_{-\infty}^{\infty} p_X(x)(\log p_X(x))dx - \lim_{\delta x \to 0} \log \delta x \int_{-\infty}^{\infty} p_X(x)dx$$

$$= h(X) - \lim_{\delta x \to 0} \log \delta x$$

第二项是所有微分熵都存在的项，且为无穷大，将其省略。

熵、最大熵

信息熵

微分熵性质

- 平移不变

$$h(X+c)=h(X)$$

- 尺度变化

$$h(cX)=h(X)+\log|c|$$

$$h(\boldsymbol{A}X)=h(X)+\log|\det(\boldsymbol{A})|$$

其中，c 为常数，$\boldsymbol{A}$ 为矩阵，$\det(\boldsymbol{A})$是矩阵 $\boldsymbol{A}$ 的行列式。

熵、最大熵

熵界限

离散熵情况: $H(X)=E[I(x_k)]=\sum_{k=-N}^{k=N} p_k I(x_k)=-\sum_{k=-N}^{k=N} p_k \log p_k$

$$0 \le H(X) \le \log(2N+1)$$

其中（$2N+1$）是总的离散值的数目

- $H(X)=0$，**当且仅当对于某一 k 概率 $p_k=1$ 时，不确定性为 0。**
- $H(X)=\log(2N+1)$，**当且仅当所有 k 概率 $p_k= 1/(2N+1)$时，不确定性最大。**

$$H(X)=-\sum_{k=-N}^{k=N} \frac{1}{2N+1} \log \frac{1}{2N+1}=\log(2N+1)$$

熵、最大熵

熵界限

微分熵情况

微分熵的值可以是负值，例如：

考虑在 [0,a] 区间上均匀分布的随机变量 X，其概率密度函数为：

$$p_X(x)=\begin{cases}\frac{1}{a}, & 0\leqslant x\leqslant a\\ 0, & \text{否则}\end{cases}$$

X 的微分熵为：

$$h(X)=-\int_0^a \frac{1}{a}\log\left(\frac{1}{a}\right)\mathrm{d}x=\log a$$

当 $a<1$，$\log a$ 为负，这意味着熵 $h(X)$ 是负的。

对于微分熵，只能讨论何种概率分布使得熵最大？将在 8.2 节讨论。

熵、最大熵

最大熵

最大熵（Jaynes，1957）

当根据不完整的信息作为依据进行推断时，应该由满足分布限制条件的具有最大熵的概率分布推得。

熵、最大熵

最大熵

最大微分熵问题

$$h(X)=-\int_{-\infty}^{\infty} p_x(x)\log p_x(x)dx$$

约束条件：

1. $p_x(x)\geq 0$
2. $\int_{-\infty}^{\infty} p_x(x)dx=1$
3. $\int_{-\infty}^{\infty} p_x(x)g_i(x)dx=\alpha_i\ ,i=1,2,...,m$

熵、最大熵

最大熵

拉格朗日函数

$$L(p_x(x))=\int_{-\infty}^{\infty}\left[p_x(x)\log p_x(x)-\lambda_0 p_x(x)-\sum_{i=1}^{m}\lambda_i g_i(x)p_x(x)\right]dx$$

被积分项求导，并置零

$$\frac{\partial L(p_x(x))}{\partial p_x(x)}=1+\log p_x(x)-\lambda_0-\sum_{i=1}^{m}\lambda_i g_i(x)=0$$

$$p_x(x)=\exp\left(-1+\lambda_0+\sum_{i=1}^{m}\lambda_i g_i(x)\right)$$

熵、最大熵

最大熵

例子：已知均值和方差，高斯分布的微分熵最大

假设已知随机变量 X 的均值和方差(μ,σ^2)，转化为约束条件

$$\int_{-\infty}^{\infty}(x-\mu)^2 p_x(x)dx=\sigma^2$$

对应约束条件 $g_1(x)=(x-\mu)^2$，$a_1=\sigma^2$ ；

代入 $p_x(x)=\exp\left(-1+\lambda_0+\sum_{i=1}^{m}\lambda_i g_i(x)\right)$，得到

$$p_x(x)=\exp\left(-1+\lambda_0+\lambda_1(x-\mu)^2\right) \qquad (1)$$

将（1）代入约束条件 2，3:

$\int_{-\infty}^{\infty}p_x(x)dx=1$,

$\int_{-\infty}^{\infty}(x-\mu)^2 p_x(x)dx=\sigma^2$, 解得 $\lambda_0=1-\log(2\pi\sigma^2)$, $\lambda_1=(-1/2\sigma^2)$

熵、最大熵

最大熵

- **得到 $p_x(x)$ 的概率分布：**

$$p_x(x)=\exp\left(-1+\lambda_0+\lambda_1(x-\mu)^2\right)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

给定均值和方差情况下，高斯分布的微分熵最大

- **最大熵值：** $h(x)=(1/2)\left(1+\log(2\pi\sigma^2)\right)$

- **结论也适用多维情况**

$$h(\boldsymbol{X})=\frac{1}{2}[m+m\log(2\pi)+\log\left|\det(\boldsymbol{\Sigma})\right|]$$

第八章 信息论模型

8.1 概述

8.2 熵、最大熵

8.3 互信息

8.4 信息论优化模型

互信息

条件熵

条件信息量： $L(x|y)=\log\frac{1}{p(x|y)}$

条件熵：

给定 y，X 的条件熵

$$H(X|y)=\sum_x p(x|y)\log\frac{1}{p(x|y)}$$

给定 Y，X 的条件熵

$$H(X|Y)=\sum_y p(y)H(X|y)=\sum_y p(y)\sum_x p(x|y)\log\frac{1}{p(x|y)}=\sum_y\sum_x p(x,y)\log\frac{1}{p(x|y)}$$

互信息

联合熵

联合概率密度: $p(x,y)=p(x|y)p(y)==p(y|x)p(x)$

联合信息量: $L(x,y)=\log\frac{1}{p(x,y)}$

联合微分熵: $H(X,Y)=\sum_y\sum_x p(x,y)\log\frac{1}{p(x,y)}$

$$H(X,Y)=H(X|Y)+H(Y)=H(Y|X)+H(X)$$

$$H(X,Y)=\sum_y\sum_x p(x,y)\log\frac{1}{p(x|y)p(y)}$$
$$=\sum_y\sum_x p(x,y)\log\frac{1}{p(x|y)}+\sum_y\sum_x p(x,y)\log\frac{1}{p(y)}$$
$$=H(X|Y)+H(Y)$$

$$H(X,Y)=\sum_y\sum_x p(x,y)\log\frac{1}{p(y|x)p(x)}$$
$$=\sum_y\sum_x p(x,y)\log\frac{1}{p(y|x)}+\sum_y\sum_x p(x,y)\log\frac{1}{p(x)}$$
$$=H(Y|X)+H(X)$$

互信息

互信息

互信息：信息熵与条件熵的差

$$I(X;Y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}p_{X,Y}(x,y)\log\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}dxdy$$

$$=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}p_{X,Y}(x|y)p_Y(y)\log\frac{p_{X,Y}(x|y)}{p_X(x)}dxdy$$

$$I(X;Y)=h(X)-h(X\mid Y)$$

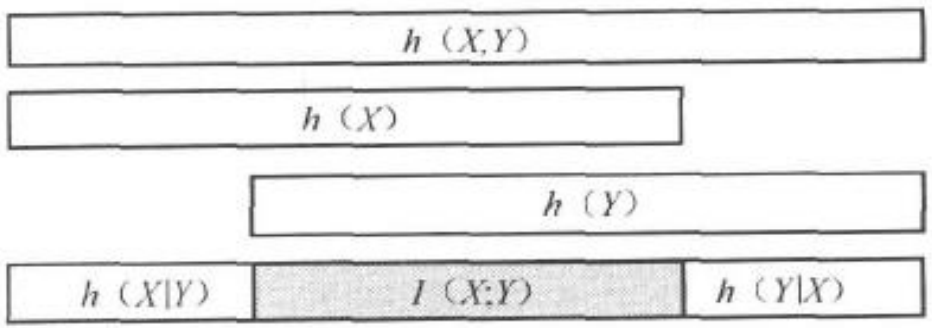
$$I(X;Y)=h(X)+h(Y)-h(X,Y)$$

互信息

互信息与熵之间的关系

$$I(X;Y) = h(X) - h(X \mid Y)$$
$$= h(Y) - h(Y \mid X)$$
$$= (h(X) + h(Y)) - h(X,Y)$$

$$h(X,Y) = (h(X) + h(Y)) - I(X;Y)$$



- H(X): 信源中每个符号的平均信息
- H(Y): 信宿中每个符号的平均信息
- H(X/Y): 输出端接受到 Y 的全部符号后，发送端 X 尚存的平均不确定性（信道疑义度，损失熵）
- H(Y/X): 在已知 X 的全部符号后，对于输出 Y 尚存的平均不确定性（噪声熵）
- H(X,Y): 表示整个信息传输系统的平均不确定性（联合熵）

互信息

互信息性质

性质 1　非负性

互信息$I(X;Y)$总是非负的，即$I(X;Y)\geq 0$

性质 2　对称性

这第二个性质说明　　$I(Y;X)=I(X;Y)$

性质 3　不变性

在随机变量的可逆变换下互信息是不变的。

考虑可逆变换

$$u = f(x)$$
$$v = g(y)$$
$$I(X;Y) = I(U;V)$$

互信息

相对熵

相对熵是衡量两个分布的平均信息差异

$$D_{p_x \| q_x} = \int_{-\infty}^{\infty} p_x(x) \log\left(\frac{p_x(x)}{q_x(x)}\right) dx = E\left[\log\left(\frac{p_x(x)}{q_x(x)}\right)\right]$$

性质 1 非负性:

$$D_{p_x \| q_x} \geq 0$$

性质 2 不变性:

考虑可逆变换 $y = f(x)$,

那么 $D_{p_x \| q_x} = D_{p_y \| q_y}$

互信息

相对熵

相对熵和互信息之间的关系

$$I(X;Y)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}p_{x,y}(x,y)\log\left(\frac{p_{x,y}(x,y)}{p_x(x)p_y(y)}\right)dxdy$$

互信息是一种相对熵

$$I(X;Y)=D_{p_{x,y}\|p_x p_y}$$

第八章 信息论模型

8.1 概述

8.2 熵、最大熵

8.3 互信息

8.4 信息论优化模型

信息论优化模型

优化原则

- **最大熵模型：**最大化 H(Y|X)，求取类别后验概率分布 p(y|x)，用于分类、预测等；
- **最大互信息模型：**最大化 $I(Y; X)$；最大化 $I(Y_a; Y_b)$；
- **最小互信息模型：**最小化 $I(Y_a; Y_b)$；最小化 $I(Y_1, Y_2, \ldots Y_m)$，独立分析（ICA）

信息论优化模型

优化原则

- **最大熵模型：**最大化 H(Y|X)，求取类别后验概率分布 p(y|x)，用于分类、预测等；
- **最大互信息模型：**最大化 $I(Y; X)$；最大化 $I(Y_a; Y_b)$；
- **最小互信息模型：**最小化 $I(Y_a; Y_b)$；最小化 $I(Y_1, Y_2, \ldots Y_m)$，独立分析（ICA）

信息论优化模型

最大熵模型

条件熵

$$H(Y|X)=-\sum_{x,y}P(x,y)\log P(y|x)$$

$$=-\sum_{x,y}P(x)P(y|x)\log P(y|x) \Rightarrow H(Y|X)=-\sum_{x,y}\tilde{P}(x)P(y|x)\log P(y|x)$$

1. **H(Y|X)应该是仅包含 *P(y|x)*的函数，** ⇨ 通过训练样本统计的经验概率：

$$\tilde{P}(x)=\frac{Number(X=x)}{N};\ \tilde{P}(x,y)=\frac{Number(X=x,Y=y)}{N};$$

2. **约束条件应当对 *P(y|x)*进行限定。**

⇩

$$E_p(f_i)=E_{\tilde{p}}(f_i),\quad i=1,2,...,n$$

$$\sum_y P(y|x)=1$$

信息论优化模型

最大熵模型

约束项

$$E_p(f_i)=E_{\tilde{p}}(f_i),\quad i=1,2,...,n$$

$$\sum_y P(y|x)=1$$

特征函数：$f(x,y)=\begin{cases}1, & x与y满足某一事实\\ 0, & 否则\end{cases}$

$$E_{\tilde{p}}(f_i)=\sum_{x,y}\tilde{p}(x)\tilde{p}(y|x)f(x,y)$$

$$E_p(f_i)=\sum_{x,y}p(x)p(y|x)f(x,y)$$

$$\Downarrow$$

$$=\sum_{x,y}\tilde{p}(x)p(y|x)f(x,y)$$

信息论优化模型

最大熵模型

最大熵模型

$$\max_{p(y|x)} H(p) = -\sum_{x,y} \tilde{P}(x) P(y|x) \log P(y|x)$$

$$s.t. \quad E_p(f_i) = E_{\tilde{p}}(f_i), \quad i = 1,2,...,n$$

$$\sum_y P(y|x) = 1$$

等价问题

$$\min_{p(y|x)} -H(p) = \sum_{x,y} \tilde{P}(x) P(y|x) \log P(y|x)$$

$$s.t. \quad E_p(f_i) = E_{\tilde{p}}(f_i), \quad i = 1,2,...,n$$

$$\sum_y P(y|x) = 1$$

拉格朗日函数

$$L(p,w) = -H(p) + w0\left(1 - \sum_y p(y|x)\right) + \sum_{i=1}^{n} w_i \left(E_{\tilde{p}}(f_i) - E_p(f_i)\right)$$

$$= \sum_{x,y} \tilde{P}(x) P(y|x) \log P(y|x) + w0\left(1 - \sum_y p(y|x)\right) + \sum_{i=1}^{n} w_i \left(\sum_{x,y} \tilde{p}(x) \tilde{p}(y|x) f_i(x,y) - \sum_{x,y} \tilde{p}(x) p(y|x) f_i(x,y)\right)$$

信息论优化模型

最大熵模型

原始的优化问题 ⇨ 对偶问题

$$\min_{p}\max_{w} L(p,w) \quad \Rightarrow \quad \max_{w}\min_{p} L(p,w)$$

$$\frac{\partial L(p,w)}{\partial P(y|x)} = \sum_{x,y}\tilde{p}(x)\left(\log P(y|x)+1\right) - \sum_{y} w_0 - \sum_{x,y}\left(\tilde{p}(x)\sum_{i=1}^{n} w_i f_i(x,y)\right)$$

$$= \sum_{x,y}\tilde{p}(x)\left(\log P(y|x)+1-w_0-\sum_{i=1}^{n} w_i f_i(x,y)\right)=0$$

解得：

$$P(y|x)=\exp\left(1-w_0-\sum_{i=1}^{n} w_i f_i(x,y)\right)=\frac{\exp\left(\sum_{i=1}^{n} w_i f_i(x,y)\right)}{\exp(1-w_0)}$$

由于 $\sum_{y} P(y|x)=1$，

归一化得：

$$P_w(y|x)=\frac{\exp\left(\sum_{i=1}^{n} w_i f_i(x,y)\right)}{\sum_{y}\exp\left(\sum_{i=1}^{n} w_i f_i(x,y)\right)},$$

归一化因子： $Z_w(x)=\sum_{y}\exp\left(\sum_{i=1}^{n} w_i f_i(x,y)\right)$

信息论优化模型

最大熵模型

原始的优化问题 ⇨ 对偶问题 （优化、对偶问题，参考《李航》附录 C）

$$\min_{p}\max_{w} L(p,w) \Rightarrow \max_{w}\min_{p} L(p,w)$$

求参数 w: $\max_{w} \Psi(w)$,

$$\Psi(w)=\min_{p} L(p,w)$$

$$=\sum_{x,y}\tilde{P}(x)P_w(y|x)\log P_w(y|x)+w0\left(1-\sum_{y}p_w(y|x)\right)+\sum_{i=1}^{n}w_i\left(\sum_{x,y}\tilde{p}(x)\tilde{p}(y|x)f_i(x,y)-\sum_{x,y}\tilde{p}(x)p_w(y|x)f_i(x,y)\right)$$

$$=\sum_{x,y}\tilde{P}(x)P_w(y|x)\log P_w(y|x)+\sum_{i=1}^{n}w_i\left(\sum_{x,y}\tilde{p}(x)\tilde{p}(y|x)f_i(x,y)-\sum_{x,y}\tilde{p}(x)p_w(y|x)f_i(x,y)\right)$$

$$=\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n}w_i f_i(x,y)+\sum_{x,y}\tilde{P}(x)P_w(y|x)\left(\log P_w(y|x)-\sum_{i=1}^{n}w_i f_i(x,y)\right)$$

$$=\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n}w_i f_i(x,y)+\sum_{x,y}\tilde{P}(x)P_w(y|x)\log Z_w(x)$$

$$=\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n}w_i f_i(x,y)+\sum_{x}\tilde{P}(x)\log Z_w(x)$$

优化方法：

梯度下降，迭代尺度，

牛顿法或拟牛顿法

参见《统计学习方法》p88.

$$P_w(y|x)=\frac{\exp\left(\sum_{i=1}^{n}w_i f_i(x,y)\right)}{\sum_{y}\exp\left(\sum_{i=1}^{n}w_i f_i(x,y)\right)}$$

等价于对数似然

$$L_{\tilde{p}}(p_w)=\log\prod_{x,y}P(x|y)^{P(x,y)}=\sum_{x,y}\tilde{P}(x,y)\log P(y|x)$$

$$=\sum_{x,y}\tilde{P}(x,y)\sum_{i=1}^{n}w_i f_i(x,y)+\sum_{x}\tilde{P}(x)\log Z_w(x)$$

信息论优化模型

最大熵模型

例子

假设随机变量 X 有 5 个取值 $\{A,B,C,D,E\}$，要估计取各个值的概率 $P(A),P(B),P(C),P(E),P(D)$.

基于最大熵原则，求解以下 2 个约束条件下的各率分布。

问题 1：约束条件为

$$P(A)+P(B)+P(C)+P(E)+P(D)=1$$

问题 2：约束条件为

$$P(A)+P(B)=\frac{3}{10}$$

$$P(A)+P(B)+P(C)+P(E)+P(D)=1$$

信息论优化模型

最大熵模型

- 问题 1：

$$\min -H(P)=\sum_{i=1}^{5}P(y_i)\log P(y_i)$$

$$s.t.\ \sum_{i=1}^{5}P(y_i)=1$$

基于前面讨论的熵界限，当各概率值相等时，H(P)最大.

$$P(A)=P(B)=P(C)=P(E)=P(D)=\frac{1}{5}$$

信息论优化模型

最大熵模型

- 问题 2:

$$\min -H(P)=\sum_{i=1}^{5}P(y_i)\log P(y_i)$$

$$s.t.\ P(y_1)+P(y_2)=\frac{3}{10}$$

$$\sum_{i=1}^{5}P(y_i)=1$$

引进拉格朗日 w_0, w_1 乘子，定义拉格拉朗日函数

$$L(P,w)=\sum_{i=1}^{5}P(y_i)\log P(y_i)+w_1\left(P(y_1)+P(y_2)-\frac{3}{10}\right)+w_0\left(\sum_{i=1}^{5}P(y_i)-1\right)$$

求解 $\max_{w}\min_{P}L(P,w)$

最大熵模型

首先求解 $L(P,w)$ 关于 P 的极小化问题。为此，固定 w_0, w_1，求偏导数：

$$\frac{\partial L(P,w)}{\partial P(y_1)} = 1 + \log P(y_1) + w_1 + w_0$$
$$\frac{\partial L(P,w)}{\partial P(y_2)} = 1 + \log P(y_2) + w_1 + w_0$$
$$\frac{\partial L(P,w)}{\partial P(y_3)} = 1 + \log P(y_3) + w_0$$
$$\frac{\partial L(P,w)}{\partial P(y_4)} = 1 + \log P(y_4) + w_0$$
$$\frac{\partial L(P,w)}{\partial P(y_5)} = 1 + \log P(y_5) + w_0$$

令各偏导数等于零解得：

$$P(y_1) = P(y_2) = e^{-w_1 - w_0 - 1}$$
$$P(y_3) = P(y_4) = P(y_5) = e^{-w_0 - 1}$$

$$\min_P L(P,w) = L(P_w, w) = -2e^{-w_1 - w_0 - 1} - 3e^{-w_0 - 1} - \frac{3}{10} w_1 - w_0$$

最大熵模型

再求解 $L(P,w)$ 关于 w 的极大化问题

分别求 $L(P,w)$ 对 w_0, w_1 的偏导数并令其为零，得到

$$e^{-w_1-w_0-1} = \frac{3}{20}$$

$$e^{-w_0-1} = \frac{7}{30}$$

于是得到所要求的概率分布为

$$P(y_1)=P(y_2)=\frac{3}{20}$$

$$P(y_3)=P(y_4)=P(y_5)=\frac{7}{30}$$

本讲参考文献

1. Simon Haykin，Neural Network and Learning Machine. 3rd
2. Simon Haykin，申富饶等译，神经网络与学习机器，第三版。
3. 统计学习方法，李航，清华大学，2012.