机器学习的数学笔记

X

目录

Notation

This section provides a concise reference describing notation used throughout this document. If you are unfamiliar with any of the corresponding mathematical concepts, Goodfellow *et al.* (2016) describe most of these ideas in chapters 2–4.

Numbers and Arrays

a	A scalar (integer or real)
$\boldsymbol{a}$	A vector
$\boldsymbol{A}$	A matrix
$\mathsf{\mathbf{A}}$	A tensor
$\boldsymbol{I}_n$	Identity matrix with n rows and n columns
$\boldsymbol{I}$	Identity matrix with dimensionality implied by context
$\boldsymbol{e}^{(i)}$	Standard basis vector $[0, \dots, 0, 1, 0, \dots, 0]$ with a 1 at position i
$\text{diag}(\boldsymbol{a})$	A square, diagonal matrix with diagonal entries given by $\boldsymbol{a}$
a	A scalar random variable
$\mathbf{a}$	A vector-valued random variable
$\mathbf{A}$	A matrix-valued random variable

Sets and Graphs

$\mathbb{A}$	A set
$\mathbb{R}$	The set of real numbers
$\{0, 1\}$	The set containing 0 and 1
$\{0, 1, \dots, n\}$	The set of all integers between 0 and n
$[a, b]$	The real interval including a and b
$(a, b]$	The real interval excluding a but including b
$\mathbb{A} \backslash \mathbb{B}$	Set subtraction, i.e., the set containing the elements of $\mathbb{A}$ that are not in $\mathbb{B}$
$\mathcal{G}$	A graph
$Pa_{\mathcal{G}}(\mathrm{x}_i)$	The parents of x_i in $\mathcal{G}$

Indexing

a_i	Element i of vector $\boldsymbol{a}$, with indexing starting at 1
a_{-i}	All elements of vector $\boldsymbol{a}$ except for element i
$A_{i,j}$	Element i, j of matrix $\boldsymbol{A}$
$\boldsymbol{A}_{i,:}$	Row i of matrix $\boldsymbol{A}$
$\boldsymbol{A}_{:,i}$	Column i of matrix $\boldsymbol{A}$
$\mathsf{A}_{i,j,k}$	Element (i, j, k) of a 3-D tensor $\mathsf{\mathbf{A}}$
$\mathsf{\mathbf{A}}_{:,:,i}$	2-D slice of a 3-D tensor
a_i	Element i of the random vector $\mathbf{a}$

Linear Algebra Operations

$\boldsymbol{A}^\top$	Transpose of matrix $\boldsymbol{A}$
$\boldsymbol{A}^+$	Moore-Penrose pseudoinverse of $\boldsymbol{A}$
$\boldsymbol{A} \odot \boldsymbol{B}$	Element-wise (Hadamard) product of $\boldsymbol{A}$ and $\boldsymbol{B}$
$\det(\boldsymbol{A})$	Determinant of $\boldsymbol{A}$

Calculus

$\dfrac{dy}{dx}$	Derivative of y with respect to x
$\dfrac{\partial y}{\partial x}$	Partial derivative of y with respect to x
$\nabla_{\boldsymbol{x}} y$	Gradient of y with respect to $\boldsymbol{x}$
$\nabla_{\boldsymbol{X}} y$	Matrix derivatives of y with respect to $\boldsymbol{X}$
$\nabla_{\mathsf{\mathbf{X}}} y$	Tensor containing derivatives of y with respect to $\mathsf{\mathbf{X}}$
$\dfrac{\partial f}{\partial \boldsymbol{x}}$	Jacobian matrix $\boldsymbol{J} \in \mathbb{R}^{m \times n}$ of $f : \mathbb{R}^n \rightarrow \mathbb{R}^m$
$\nabla^2_{\boldsymbol{x}} f(\boldsymbol{x})$ or $\boldsymbol{H}(f)(\boldsymbol{x})$	The Hessian matrix of f at input point $\boldsymbol{x}$
$\int f(\boldsymbol{x}) d\boldsymbol{x}$	Definite integral over the entire domain of $\boldsymbol{x}$
$\int_{\mathbb{S}} f(\boldsymbol{x}) d\boldsymbol{x}$	Definite integral with respect to $\boldsymbol{x}$ over the set $\mathbb{S}$

Probability and Information Theory

$\mathrm{a} \bot \mathrm{b}$	The random variables a and b are independent
$\mathrm{a} \bot \mathrm{b} \mid \mathrm{c}$	They are conditionally independent given c
$P(\mathrm{a})$	A probability distribution over a discrete variable
$p(\mathrm{a})$	A probability distribution over a continuous variable, or over a variable whose type has not been specified
$\mathrm{a} \sim P$	Random variable a has distribution P
$\mathbb{E}_{\mathrm{x} \sim P}[f(x)]$ or $\mathbb{E} f(x)$	Expectation of $f(x)$ with respect to $P(\mathrm{x})$
$\mathrm{Var}(f(x))$	Variance of $f(x)$ under $P(\mathrm{x})$
$\mathrm{Cov}(f(x), g(x))$	Covariance of $f(x)$ and $g(x)$ under $P(\mathrm{x})$
$H(\mathrm{x})$	Shannon entropy of the random variable x
$D_{\mathrm{KL}}(P \Vert Q)$	Kullback-Leibler divergence of P and Q
$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	Gaussian distribution over $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$

Functions

$f : \mathbb{A} \rightarrow \mathbb{B}$	The function f with domain $\mathbb{A}$ and range $\mathbb{B}$
$f \circ g$	Composition of the functions f and g
$f(\boldsymbol{x}; \boldsymbol{\theta})$	A function of $\boldsymbol{x}$ parametrized by $\boldsymbol{\theta}$. (Sometimes we write $f(\boldsymbol{x})$ and omit the argument $\boldsymbol{\theta}$ to lighten notation)
$\log x$	Natural logarithm of x
$\sigma(x)$	Logistic sigmoid, $\dfrac{1}{1+\exp(-x)}$
$\zeta(x)$	Softplus, $\log(1+\exp(x))$
$\|\|\boldsymbol{x}\|\|_p$	L^p norm of $\boldsymbol{x}$
$\|\|\boldsymbol{x}\|\|$	L^2 norm of $\boldsymbol{x}$
x^+	Positive part of x, i.e., $\max(0, x)$
$\mathbf{1}_{\text{condition}}$	is 1 if the condition is true, 0 otherwise

Sometimes we use a function f whose argument is a scalar but apply it to a vector, matrix, or tensor: $f(\boldsymbol{x})$, $f(\boldsymbol{X})$, or $f(\mathsf{X})$. This denotes the application of f to the array element-wise. For example, if $\mathsf{C} = \sigma(\mathsf{X})$, then $C_{i,j,k} = \sigma(X_{i,j,k})$ for all valid values of i, j and k.

Datasets and Distributions

p_{data}	The data generating distribution
$\hat{p}_{\text{data}}$	The empirical distribution defined by the training set
$\mathbb{X}$	A set of training examples
$\boldsymbol{x}^{(i)}$	The i-th example (input) from a dataset
$y^{(i)}$ or $\boldsymbol{y}^{(i)}$	The target associated with $\boldsymbol{x}^{(i)}$ for supervised learning
$\boldsymbol{X}$	The $m \times n$ matrix with input example $\boldsymbol{x}^{(i)}$ in row $\boldsymbol{X}_{i,:}$

Chapter 1

逻辑回归

1.1 二项逻辑回归模型

二项逻辑回归模型是如下的条件概率分布

$$
\begin{aligned}
P(Y=1|\boldsymbol{x}) =& \frac{\exp(\boldsymbol{\theta}^T\boldsymbol{x}+b)}{1+\exp(\boldsymbol{\theta}^T\boldsymbol{x}+b)} \\
P(Y=0|\boldsymbol{x}) =& \frac{1}{1+\exp(\boldsymbol{\theta}^T\boldsymbol{x}+b)}
\end{aligned}
$$

其中$\boldsymbol{x}\in\mathbb{R}^n$是输入变量， $Y\in\{0,1\}$是输出变量，$\boldsymbol{\theta}\in\mathbb{R}^n$和$b\in\mathbb{R}$是参数。 $\boldsymbol{x}$和$\boldsymbol{\theta}$为n维列向量。

若令$\boldsymbol{\theta}=(\theta^{(1)},...,\theta^{(n)},b)^T$， $\boldsymbol{x}=(x^{(1)},...,x^{(n)},1)^T$，那么条件概率可以表示为

$$
\begin{aligned}
P(Y=1|\boldsymbol{x}) =& \frac{\exp(\boldsymbol{\theta}^T\boldsymbol{x})}{1+\exp(\boldsymbol{\theta}^T\boldsymbol{x})} \\
P(Y=0|\boldsymbol{x}) =& \frac{1}{1+\exp(\boldsymbol{\theta}^T\boldsymbol{x})}
\end{aligned}
\tag{1.1}
$$

1.1.1 模型的参数估计

对于给定的训练集$\mathbb{X}=\{(\boldsymbol{x}_1,y_1),...,(\boldsymbol{x}_N,y_N)\}$，可应用极大似然估计法估计模型参数。

为表示方便，令$P(Y=1|\boldsymbol{x})=\pi(\boldsymbol{x}),P(Y=0|\boldsymbol{x})=1-\pi(\boldsymbol{x})$，似然函数为

$$
L(\boldsymbol{\theta})=\prod_{i=1}^{N}\left(\pi(\boldsymbol{x}_i)\right)^{y_i}\left(1-\pi(\boldsymbol{x}_i)\right)^{1-y_i}
$$

那么对数似然函数为

$$
\begin{aligned}
\log L(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \left(y_i \log \pi(\boldsymbol{x}_i) + (1-y_i)\log(1-\pi(\boldsymbol{x}_i))\right) \\
&= \sum_{i=1}^{N} \left(y_i \log \frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)} + \log(1-\pi(\boldsymbol{x}_i))\right) \\
&= \sum_{i=1}^{N} \left(y_i(\boldsymbol{\theta}^T \boldsymbol{x}_i) - \log(1+\exp(\boldsymbol{\theta}^T \boldsymbol{x}_i))\right)
\end{aligned}
\tag{1.2}
$$

1.1.1.1 参数估计：梯度下降法

根据公式 (1.2)，对数似然函数对$\boldsymbol{\theta}$的偏导为

$$
\begin{aligned}
\nabla_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta}) &= \sum_{i=1}^{N} \left(y_i \boldsymbol{x}_i - \frac{\exp(\boldsymbol{\theta}^T x_i)\boldsymbol{x}_i}{1+\exp(\boldsymbol{\theta}^T \boldsymbol{x}_i)}\right) \\
&= \sum_{i=1}^{N} \left(y_i - \pi(x_i)\right) x_i
\end{aligned}
$$

由此此处求对数似然函数的最大值，故需要沿着梯度上升的方向进行迭代，迭代公式为

$$
\begin{aligned}
\boldsymbol{\theta} :=& \boldsymbol{\theta} + \alpha \frac{\partial}{\partial \boldsymbol{\theta}} \log L(\boldsymbol{\theta}) \\
=& \boldsymbol{\theta} + \alpha \sum_{i=1}^{N} \left(y_i - \pi(\boldsymbol{x}_i)\right) \boldsymbol{x}_i
\end{aligned}
\tag{1.3}
$$

其中α称为学习率，是一个正常数。

公式 (1.3)可以用矩阵表示

$$
\boldsymbol{\theta} := \boldsymbol{\theta} + \alpha X^T \boldsymbol{\Lambda} \tag{1.4}
$$

其中$\boldsymbol{\Lambda} = \begin{pmatrix} y_1 - \pi(\boldsymbol{x}_1) \\ y_2 - \pi(\boldsymbol{x}_2) \\ \dots \\ y_N - \pi(\boldsymbol{x}_N) \end{pmatrix}_{N\times 1}$，$X$是由训练数据构成的$N \times (n+1)$矩阵(每一行对应一个样本，每一列对应样本的一个维度，其中还包括一维常数项)。

1.1.1.2 参数估计：随机梯度下降法

梯度下降算法在每次更新回归系数时需要遍历整个数据集，当数据集数量庞大或者

特征过多时，该方法的计算复杂度太高。改进方法是每次迭代仅用一个样本来更新回归系数，称为随机梯度下降法。

具体而言，对于训练集中的每一个样本(x_i, y_i)，计算该样本梯度，并依据迭代公式：

$$\boldsymbol{\theta} := \boldsymbol{\theta} + \alpha \left(y_i - \pi(\boldsymbol{x}_i) \right) \boldsymbol{x}_i \tag{1.5}$$

与公式 (1.3)相比，随机梯度下降的迭代公式 (1.5)中

- 误差变量是数值，而不是向量
- 不再有矩阵变换的过程

所以随机梯度下降算法的计算效率较高，缺点是存在解的不稳定性(如解存在周期性波动)的问题。为了解决这一问题，并进一步加快收敛速度，可以通过随机选取样本来更新回归系数。

1.2 Softmax回归模型

Softmax模型是二项回归模型在多分类问题上的推广，在多分类问题中，类标签Y可以取两个以上的值。

假设Y的取值集合是$\{1, 2, ..., K\}$，Softmax模型是如下的条件概率分布

$$P(Y = k|\boldsymbol{x}) = \frac{\exp(\boldsymbol{\theta}_k^T \boldsymbol{x})}{\sum\limits_{j=1}^{K} \exp(\boldsymbol{\theta}_j^T \boldsymbol{x})} \tag{1.6}$$

其中$\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K \in \mathbb{R}^{n+1}$ 是模型的参数。

为方便起见，下文用矩阵$\boldsymbol{\Theta}_{K\times(n+1)}$表示全部的模型参数

$$\boldsymbol{\Theta} = \begin{bmatrix} \boldsymbol{\theta}_1^T \\ . \\ . \\ \boldsymbol{\theta}_K^T \end{bmatrix}$$

1.2.1 模型的参数估计

令$P(Y = k|\boldsymbol{x}) = \pi_k(\boldsymbol{x})$，与二项逻辑回归类似，Softmax的似然函数可以表示为

$$L(\boldsymbol{\Theta}) = \prod_{i=1}^{N} \prod_{k=1}^{K} \left(\pi_k(\boldsymbol{x}_i) \right)^{\mathbf{1}_{\mathrm{y_i=k}}}$$

对数似然函数为

$$\log L(\boldsymbol{\Theta}) = \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbf{1}_{\mathrm{y_i=k}} \log \pi_k(\boldsymbol{x}_i) \tag{1.7}$$

1.2.1.1 参数估计：梯度下降法

首先求

$$\frac{\partial \pi_k(\boldsymbol{x}_i)}{\partial \boldsymbol{\theta}_k} = \frac{\boldsymbol{x}_i \exp(\boldsymbol{\theta}_k^T \boldsymbol{x}_i) \left(\sum\limits_{j=1}^{K} \exp(\boldsymbol{\theta}_j^T \boldsymbol{x}) - \exp(\boldsymbol{\theta}_k^T \boldsymbol{x}_i) \right)}{\left(\sum\limits_{j=1}^{K} \exp(\boldsymbol{\theta}_j^T \boldsymbol{x}) \right)^2} \tag{1.8}$$

故根据公式 (1.7)，得到Softmax模型的对数似然函数的梯度

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}_k} \log L(\boldsymbol{\Theta}) &= \sum_{i=1}^{N} \mathbf{1}_{\mathrm{y_i=k}} \frac{1}{\pi_k(\boldsymbol{x}_i)} \frac{\partial \pi_k(\boldsymbol{x}_i)}{\partial \boldsymbol{\theta}_k} \\
&= \sum_{i=1}^{N} \mathbf{1}_{\mathrm{y_i=k}} \frac{1}{\pi_k(\boldsymbol{x}_i)} \frac{\boldsymbol{x}_i \exp(\boldsymbol{\theta}_k^T \boldsymbol{x}_i) \left(\sum\limits_{j=1}^{K} \exp(\boldsymbol{\theta}_j^T \boldsymbol{x}_i) - \exp(\boldsymbol{\theta}_k^T \boldsymbol{x}_i) \right)}{\left(\sum\limits_{j=1}^{K} \exp(\boldsymbol{\theta}_j^T \boldsymbol{x}_i) \right)^2} \\
&= \sum_{i=1}^{N} \mathbf{1}_{\mathrm{y_i=k}} \frac{\boldsymbol{x}_i \left(\sum\limits_{j=1}^{K} \exp(\boldsymbol{\theta}_j^T \boldsymbol{x}) - \exp(\boldsymbol{\theta}_k^T \boldsymbol{x}_i) \right)}{\sum\limits_{j=1}^{K} \exp(\boldsymbol{\theta}_j^T \boldsymbol{x})} \\
&= \sum_{i=1}^{N} \mathbf{1}_{\mathrm{y_i=k}} \boldsymbol{x}_i \left(1 - \pi_k(\boldsymbol{x}_i)\right)
\end{aligned} \tag{1.9}$$

对于任意第k个分类的参数$\boldsymbol{\theta}_k$，可沿着梯度上升的方向进行迭代

$$\boldsymbol{\theta}_k := \boldsymbol{\theta}_k + \alpha \sum_{i=1}^{N} \mathbf{1}_{\mathrm{y_i=k}} \boldsymbol{x}_i \left(1 - \pi_k(\boldsymbol{x}_i)\right) \tag{1.10}$$

公式 (1.10)的迭代关系用矩阵可以表示为

$$\boldsymbol{\theta}_k := \boldsymbol{\theta}_k + \alpha X^T \boldsymbol{\Lambda} \tag{1.11}$$

其中$\mathbf{\Lambda} = \begin{pmatrix} \mathbf{1}_{\mathrm{y_1=k}}\left(1-\pi_k(\boldsymbol{x}_1)\right) \\ \mathbf{1}_{\mathrm{y_2=k}}\left(1-\pi_k(\boldsymbol{x}_2)\right) \\ \dots \\ \mathbf{1}_{\mathrm{y_N=k}}\left(1-\pi_k(\boldsymbol{x}_N)\right) \end{pmatrix}_{N\times 1}$，$X$是由训练数据构成的$N\times(n+1)$矩阵(每一行对应一个样本，每一列对应样本的一个维度，其中还包括一维常数项)。

Chapter 2

附录：信息熵

假设 [1] X是一个取有限值的离散随机变量(本文只考虑离散情况)，概率分布为P。

那么$I(X = x_i) = -\log P(X = x_i)$称为事件$x_i$的**自信息量**，随机变量$X$的**熵**定义为$X$的自信息量的数学期望，即

$$H(X) = \mathbb{E}(I(X)) = -\sum_x P(x)\log P(x)$$

熵反映的是随机变量不确定程度的大小：熵的值越大，不确定程度越高。

2.1 相关概念

2.1.1 条件熵

条件熵是指在联合概率空间上熵的条件自信息的数学期望。在已知X时，Y的条件熵为

$$H(Y|X) = \mathbb{E}_{\mathrm{x,y}} I(y_j|x_i) = -\sum_x \sum_y P(x,y)\log P(y|x) \tag{2.1}$$

Lemma 2.1.1. 与公式 (2.1)等价的定义为给定X条件下Y的条件分布概率的熵的数学期望

$$H(Y|X) = \mathbb{E}_{\mathrm{x}} H(Y|X = x) = \sum_x P(x)H(Y|X = x)$$

[1] 本章参考了信息论与编码(http://www.docin.com/p-957983839-f6.html)和信息论基础(https://wenku.baidu.com/view/5319fed3b9f3f90f76c61b1a.html)

证明.

$$
\begin{aligned}
H(Y|X) &= -\sum_x \sum_y P(x,y)\log P(y|x) \\
&= -\sum_x \sum_y P(x)P(y|x)\log P(y|x) \\
&= -\sum_x P(x) \sum_y P(y|x)\log P(y|x) \quad (P(x)\text{与}y\text{无关}) \\
&= \sum_x P(x)[-\sum_y P(y|x)\log P(y|x)] \\
&= \sum_x P(x)H(Y|X=x)
\end{aligned}
$$

□

$H(Y|X)$的含义是已知在X发生的前提下，Y发生**新带来的熵**。

2.1.2 相对熵

相对熵，也称**KL散度**，交叉熵等，定义为两个概率分布之比的数学期望。

设$Q(x), P(x)$是随机变量X中取值的两个概率分布，则P对Q的相对熵是

$$
D_{\mathrm{KL}}(P\|Q) = \sum_x P(x)\log\frac{P(x)}{Q(x)} = \mathbb{E}_{\mathrm{x}}\log\frac{P(x)}{Q(x)} \tag{2.2}
$$

相对熵可以用来度量两个随机变量的“距离”。

Lemma 2.1.2. 相对熵恒大于等于零。

证明. 对于任意分布P, Q，根据公式 (2.2)，可知

$$
\begin{aligned}
D_{\mathrm{KL}}(P\|Q) &= \sum_x P(x)\log\frac{P(x)}{Q(x)} \\
&= -\sum_x P(x)\log\frac{Q(x)}{P(x)} \\
&\geq -\log(\sum_x P(x)\frac{Q(x)}{P(x)}) \quad (\text{对}-logx\text{应用Jensen不等式}) \\
&= -\log\sum_x Q(x) \\
&= -\log 1 \\
&=0
\end{aligned}
$$

□

2.1.3 互信息

两个随机变量X, Y的**互信息**，定义为X, Y的联合分布和独立分布乘积的相对熵

$$I(X, Y) = D_{\mathrm{KL}}(P(X, Y) \| P(X)P(Y)) \tag{2.3}$$

Lemma 2.1.3. 互信息与条件熵满足如下关系

$$H(X|Y) = H(X) - I(X, Y) \tag{2.4}$$

证明. 根据公式 (2.2)以及互信息的定义可知

$$I(X, Y) = \sum_{x,y} P(x, y) \log \frac{P(x, y)}{P(x)P(y)}$$

那么

$$\begin{aligned}
H(X) - I(X, Y) &= -\sum_{x} P(x) \log P(x) - \sum_{x,y} P(x, y) \log \frac{P(x, y)}{P(x)P(y)} \\
&= -\sum_{x} \left(\sum_{y} P(x, y) \right) \log P(x) - \sum_{x,y} P(x, y) \log \frac{P(x, y)}{P(x)P(y)} \\
&= -\sum_{x,y} P(x, y) \log P(x) - \sum_{x,y} P(x, y) \log \frac{P(x, y)}{P(x)P(y)} \\
&= -\sum_{x,y} P(x, y) \left(\log P(x) + \log \frac{P(x, y)}{P(x)P(y)} \right) \\
&= -\sum_{x,y} P(x, y) \log \frac{P(x, y)}{P(y)} \\
&= -\sum_{x,y} P(x, y) \log P(x \mid y) \\
&= H(X|Y) \quad (\text{根据公式 } (2.1))
\end{aligned}$$

□

2.2 熵的性质

X的熵具有如下几个性质

- 非负性：$H(X) \geq 0$。[2]
- 对称性: 当随机变量的概率取值任意互换时，熵不变。

$$H(p_1, p_2 ... p_n) = H(p_2, p_1 ... p_n) = H(p_3, p_1 ... p_n) = ..$$

- 可加性: 如果随机变量X, Y相互独立，则$H(X, Y) = H(X) + H(Y)$。
- 极值性: 对于任意概率分布$P(X = x_i) = p_i$和$P(Y = y_i) = q_i$，$i = 1...n$，都有

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i \leq -\sum_{i=1}^{n} p_i \log q_i \tag{2.5}$$

  当X和Y的概率分布相同时，公式 (2.5)取等号。

  该性质表明，任意概率分布，它对其他概率分布的自信息取数学期望时，必大于它本身的熵。

- 凸性：对于任意概率分布$P(X = x_i) = p_i$和$P(Y = y_i) = q_i$，$i = 1...n$，假设随机变量Z的分布为$P(Z = z_i) = \gamma_i = \alpha p_i + (1 - \alpha) q_i, \alpha \in [0, 1]$，那么$Z$的熵满足

$$H(Z) \geq \alpha H(X) + (1 - \alpha) H(Y) \tag{2.6}$$

Theorem 2.2.1 (最大熵定理). 离散随机变量X的概率分布为$P(X = x_i) = p_i, i = 1...n$，那么

$$H(X) \leq \log n \tag{2.7}$$

当$p_1 = p_2 = ... = \frac{1}{n}$ 时，等号成立。

证明. 求熵的最大值等价于以下优化问题

$$\begin{aligned} \max \quad & H(x) = -\sum_{i=1}^{n} p_i \log p_i \\ s.t. \quad & \sum_{i=1}^{n} p_i = 1 \end{aligned}$$

利用拉格朗日乘子法构造函数

$$G(p, \lambda) = -\sum_{i=1}^{n} p_i \log p_i + \lambda \left(\sum_{i=1}^{n} p_i - 1 \right) \tag{2.8}$$

[2]实际上这种非负性对于离散随机变量X成立，对连续随机变量X不一定成立。这是本文只考虑离散情况的原因。

公式 (2.8)中分别对p_i和λ求导，令其为零，得到

$$
\begin{aligned}
\frac{\partial G(p,\lambda)}{\partial p_i} = -\log p_i - 1 + \lambda = 0 \\
\sum_{i=1}^{n} p_i - 1 = 0
\end{aligned}
\tag{2.9}
$$

由$-\log p_i - 1 + \lambda = 0$可得到 $p_i = e^{\lambda-1}, i = 1, 2..n$, 由此可知$p_1 = p_2 = ... = \frac{1}{n}$

□

Lemma 2.2.1 (熵的强可加性)**.** 当随机变量X, Y相关的情况下，联合熵满足强可加性，即

$$
\begin{aligned}
H(X,Y) =& H(Y) + H(X|Y) \\
H(X,Y) =& H(X) + H(Y|X)
\end{aligned}
\tag{2.10}
$$

证明.

$$
\begin{aligned}
H(Y) + H(X|Y) =& -\sum_y P(y)\log P(y) - \sum_x \sum_y P(x,y)\log P(x|y) \\
=& -\sum_x \sum_y P(x,y)\log P(y) - \sum_x \sum_y P(x,y)\log P(x|y) \\
=& -\sum_x \sum_y P(x,y)\log P(x,y) \\
=& H(X,Y)
\end{aligned}
$$

同理可证

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

□

Lemma 2.2.2 (熵的凸性)**.** 证明公式 (2.6)

证明.

$$
\begin{aligned}
H(Z) =& -\sum_{i=1}^{n} \gamma_i \log \gamma_i \\
=& -\sum_{i=1}^{n} \alpha p_i \log \gamma_i - \sum_{i=1}^{n} (1-\alpha) q_i \log \gamma_i \\
=& -\sum_{i=1}^{n} \alpha p_i \log \left(\gamma_i \frac{p_i}{p_i} \right) - \sum_{i=1}^{n} (1-\alpha) q_i \log \left(\gamma_i \frac{q_i}{q_i} \right) \\
=& -\alpha \sum_{i=1}^{n} p_i \log p_i - (1-\alpha) \sum_{i=1}^{n} q_i \log q_i - \alpha \sum_{i=1}^{n} p_i \log \frac{\gamma_i}{p_i} - (1-\alpha) \sum_{i=1}^{n} q_i \log \frac{\gamma_i}{q_i} \\
=& \alpha H(X) + (1-\alpha) H(Y) - \alpha \sum_{i=1}^{n} p_i \log \frac{\gamma_i}{p_i} - (1-\alpha) \sum_{i=1}^{n} q_i \log \frac{\gamma_i}{q_i}
\end{aligned}
\tag{2.11}
$$

其中公式 (2.11)的倒数第二项

$$
\begin{aligned}
-\alpha \sum_{i=1}^{n} p_i \log \frac{\gamma_i}{p_i} =& \alpha \left(-\sum_{i=1}^{n} p_i \log \gamma_i + \sum_{i=1}^{n} p_i \log p_i \right) \\
\geq& 0 \quad (\text{根据公式 } (2.5)))
\end{aligned}
$$

同理可知公式 (2.11)的倒数第一项

$$
-(1-\alpha) \sum_{i=1}^{n} q_i \log \frac{\gamma_i}{q_i} \geq 0
$$

所以得到

$$
H(Z) \geq \alpha H(X) + (1-\alpha) H(Y)
$$

□

Theorem 2.2.2. 条件熵小于无条件熵，即$H(X|Y) \leq H(X)$

证明.

$$
\begin{aligned}
H(X|Y) - H(X) &= -\sum_{x,y} P(x,y)\log P(x|y) + \sum_{x} P(x)\log P(x) \\
&= -\sum_{x,y} P(x,y)\log P(x|y) + \sum_{x}\left(\sum_{y} P(x,y)\right)\log P(x) \\
&= -\sum_{x,y} P(x,y)\log P(x|y) + \sum_{x,y} P(x,y)\log P(x) \\
&= -\sum_{x,y} P(x,y)\left(\log P(x|y) - \log P(x)\right) \qquad (2.12) \\
&= -\sum_{x,y} P(x,y)\log\frac{P(x,y)}{P(x)P(y)} \\
&= -\sum_{x,y} P(x,y)\log P(x,y) - \left(-\sum_{x,y} P(x,y)\log P(x)P(y)\right) \\
&\leq 0 \quad \text{(根据熵的极值性)}
\end{aligned}
$$

□

2.2.1 整理得到的公式

根据本节内容整理得到的重要公式

- 根据条件熵定义可得

$$H(X|Y) = H(X,Y) - H(Y) \tag{2.13}$$

- 根据互信息定义展开可得

$$H(X|Y) = H(X) - I(X,Y) \tag{2.14}$$

- 根据公式 (2.13)和公式 (2.14)得到的对偶形式

$$
\begin{aligned}
H(Y|X) =& H(X,Y) - H(X) \\
H(Y|X) =& H(Y) - I(X,Y)
\end{aligned}
$$

- 多数文献将下式作为互信息的定义公式

$$I(X,Y) = H(X) + H(Y) - H(X,Y)$$

- $H(X|Y) \leq H(X)$

参考文献

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. ii

索引