机器学习的数学笔记

X

目录

符号表

下面列出了本文所用的数学符号对照表，具体可参照Goodfellow *et al.* (2016)的2-4章节。

数字和数组

a	标量(整数或实数)
$\boldsymbol{a}$	向量(矢量)
$\boldsymbol{A}$	矩阵
$\mathsf{\mathbf{A}}$	张量
$\boldsymbol{I}_n$	n 行 n 列的单位矩阵
$\boldsymbol{I}$	单位矩阵，维度参见上下文
$\boldsymbol{e}^{(i)}$	标准基向量 $[0,\dots,0,1,0,\dots,0]$ ，其中第 i 位为 1
$\mathrm{diag}(\boldsymbol{a})$	正方形对角矩阵，其中对角线元素为 $\boldsymbol{a}$
a	随机变量，标量
$\mathbf{a}$	随机变量，向量
$\mathbf{A}$	随机变量，矩阵

集合和图

$\mathbb{A}$	集合
$\mathbb{R}$	实数集合
$\{0,1\}$	包含 0 和 1 的集合
$\{0,1,\ldots,n\}$	从 0 到 n 的所有整数的集合
$[a,b]$	a 到 b 的实数闭区间
$(a,b]$	a 到 b 的实数半开区间
$\mathbb{A}\backslash\mathbb{B}$	差集, 即集合包含了在 $\mathbb{A}$ 中但不在 $\mathbb{B}$ 中的元素
$\mathcal{G}$	图
$Pa_{\mathcal{G}}(\mathrm{x}_i)$	图$\mathcal{G}$ 中节点 x_i 的双亲

角标

a_i	向量 $\boldsymbol{a}$ 的第 i 个元素, 其中角标从 1 开始
a_{-i}	向量 $\boldsymbol{a}$ 除第i个元素以外的所有元素
$A_{i,j}$	矩阵 $\boldsymbol{A}$ 的第 i 行第 j 列的元素
$\boldsymbol{A}_{i,:}$	矩阵 $\boldsymbol{A}$ 的第 i 行
$\boldsymbol{A}_{:,i}$	矩阵 $\boldsymbol{A}$ 的第 i 列
$\mathsf{A}_{i,j,k}$	3-D 张量 $\mathbf{\mathsf{A}}$ 的元素 (i,j,k)
$\mathbf{\mathsf{A}}_{:,:,i}$	3-D 张量的 2-D 切片
a_i	随机向量 $\mathbf{a}$ 的第 i 个元素

线性代数

$\boldsymbol{A}^\top$	矩阵 $\boldsymbol{A}$ 的转置
$\boldsymbol{A}^+$	矩阵 $\boldsymbol{A}$ 的摩尔彭罗斯伪逆（广义逆）
$\boldsymbol{A}\odot\boldsymbol{B}$	矩阵 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 的元素积（Hadamard乘积）
$\det(\boldsymbol{A})$	矩阵 $\boldsymbol{A}$ 的行列式

微积分

$\dfrac{dy}{dx}$	y 关于 x 的导数
$\dfrac{\partial y}{\partial x}$	y 关于 x 的偏导数
$\nabla_{\boldsymbol{x}} y$	y 关于向量$\boldsymbol{x}$ 的梯度
$\nabla_{\boldsymbol{X}} y$	y 关于矩阵$\boldsymbol{X}$ 的导数
$\nabla_{\mathsf{X}} y$	y 关于张量X 的导数
$\dfrac{\partial f}{\partial \boldsymbol{x}}$	$f : \mathbb{R}^n \to \mathbb{R}^m$ 的雅克比矩阵 $\boldsymbol{J} \in \mathbb{R}^{m \times n}$
$\nabla^2_{\boldsymbol{x}} f(\boldsymbol{x})$ or $\boldsymbol{H}(f)(\boldsymbol{x})$	f 在输入向量 $\boldsymbol{x}$ 的海森矩阵
$\displaystyle\int f(\boldsymbol{x}) d\boldsymbol{x}$	在整个定义域上f关于 $\boldsymbol{x}$ 的定积分
$\displaystyle\int_{\mathbb{S}} f(\boldsymbol{x}) d\boldsymbol{x}$	在集合 $\mathbb{S}$ 上f关于 $\boldsymbol{x}$ 的定积分

概率论与信息论

$\mathrm{a} \perp \mathrm{b}$	随机变量 a 与 b 相互独立
$\mathrm{a} \perp \mathrm{b} \mid \mathrm{c}$	随机变量 a 与 b对于给定的 c 条件独立
$P(\mathrm{a})$	离散变量的概率分布
$p(\mathrm{a})$	连续变量的概率分布或类型不确定的变量的概率分布
$\mathrm{a} \sim P$	随机变量 a 服从 P 分布
$\mathbb{E}_{\mathrm{x} \sim P}[f(x)]$ or $\mathbb{E} f(x)$	$f(x)$ 在概率分布 $P(\mathrm{x})$ 的期望
$\mathrm{Var}(f(x))$	$f(x)$ 在概率分布 $P(\mathrm{x})$ 下的方差
$\mathrm{Cov}(f(x), g(x))$	$f(x)$ 和 $g(x)$ 在概率分布 $P(\mathrm{x})$ 下的协方差
$H(\mathrm{x})$	随机变量 x 的信息熵
$D_{\mathrm{KL}}(P \Vert Q)$	随机变量P 与 Q 的相对熵(KL散度)
$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	均值为 $\boldsymbol{\mu}$ 方差为 $\boldsymbol{\Sigma}$ 的 $\boldsymbol{x}$ 的高斯分布

函数

$f: \mathbb{A} \rightarrow \mathbb{B}$	定义域为 $\mathbb{A}$ 值域为 $\mathbb{B}$ 的函数 f
$f \circ g$	函数 f 和 g 的复合函数
$f(\boldsymbol{x};\boldsymbol{\theta})$	参数为 $\boldsymbol{\theta}$ 的关于$\boldsymbol{x}$ 的函数. （有时我们写作 $f(\boldsymbol{x})$ 而忽略参数 $\boldsymbol{\theta}$ 来简化符号）
$\log x$	x 的自然对数
$\sigma(x)$	Logistic sigmoid 函数, $\dfrac{1}{1+\exp(-x)}$
$\zeta(x)$	Softplus 函数, $\log(1+\exp(x))$
$\|\|\boldsymbol{x}\|\|_p$	$\boldsymbol{x}$ 的 L^p 范数
$\|\|\boldsymbol{x}\|\|$	$\boldsymbol{x}$ 的 L^2 范数
x^+	x 的正值部分, 即 $\max(0,x)$
$\mathbf{1}_{\text{condition}}$	如果条件为真则值为1，条件为假则值为0

有时我们把参数为标量的函数 f 应用到矢量、矩阵或张量中：$f(\boldsymbol{x})$, $f(\boldsymbol{X})$, 或 $f(\mathsf{X})$。这代表着将 f 应用到数组元素层面，例如，如果 $\mathsf{C}=\sigma(\mathsf{X})$, 那么任意 i, j 和 k ，都有 $\mathsf{C}_{i,j,k}=\sigma(\mathsf{X}_{i,j,k})$

数据集和分布

p_{data}	数据生成的概率分布
$\hat{p}_{\text{data}}$	训练集生成(定义)的经验分布
$\mathbb{X}$	训练集
$\boldsymbol{x}^{(i)}$	数据集中的第 i 个（输入）实例
$y^{(i)}$ or $\boldsymbol{y}^{(i)}$	有监督学习下 $\boldsymbol{x}^{(i)}$ 对应的标记
$\boldsymbol{X}$	$m \times n$ 的矩阵，其中输入实例 $\boldsymbol{x}^{(i)}$ 在 $\boldsymbol{X}_{i,:}$ 行

Chapter 1

方法概论

本章讨论**监督学习**的基本方法与概念，内容参考了李航 (2012)和Rigollet (2015)。

1.1 基本概念

1.1.1 模型

用X和Y表示输入空间$\mathcal{X}$和输出空间$\mathcal{Y}$上的变量，用$\boldsymbol{\theta}$表示参数向量，模型的假设空间一般有两种情况

- 决策函数的集合$\mathcal{F} = \{f|Y = f_{\boldsymbol{\theta}}(X)\}$，此类模型称为**非概率模型**
- 条件概率的集合$\mathcal{F} = \{P|P_{\boldsymbol{\theta}}(Y|X)\}$，此类模型称为**概率模型**

1.1.2 策略

1.1.2.1 损失函数

度量输出的预测值$f(X)$与真实值Y差异程度的函数称为**损失函数**，记做$L(Y, f(X))$，通常的损失函数有0-1损失，平方损失等。

常用的损失函数有0-1损失，平方损失，对数损失等。

用损失函数的期望来度量模型f在联合分布$P(X, Y)$下的平均损失，也成为称为**风险**

函数或**期望风险**。

$$R_{\exp}(f) = \mathbb{E}[L(Y, f(X))] = \int_{\mathcal{X}\times\mathcal{Y}} L(Y, f(X))dP(X, Y)$$

学习的目标是选择期望最小的模型。然而实践中联合概率分布$P(X, Y)$是未知的，只能用**经验风险**来估计。

给定一个训练集$\mathbb{X} = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2).., (\boldsymbol{x}_N, y_N)\}$

$$R_{\mathrm{emp}}(f) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(\boldsymbol{x}_i))$$

根据监督学习的基本假设，训练数据和测试数据都是依$P(X, Y)$独立同分布产生的，所以$R_{\mathrm{emp}}(f)$是$R_{\exp}(f)$的无偏估计。根据大数定律，$R_{\mathrm{emp}}(f)$收敛于$R_{\exp}(f)$。

1.2 泛化误差

假设学习到的模型为$\hat{f}$，那么该模型的泛化误差定义为该模型的期望风险：

$$R_{\exp}(\hat{f}) = \mathbb{E}[L(Y, \hat{f}(X))] = \int_{\mathcal{X}\times\mathcal{Y}} L(Y, \hat{f}(X))dP(X, Y) \tag{1.1}$$

1.2.1 二类分类问题

下文以二类分类问题为例，讨论模型的泛化误差。

假设模型的输入空间$\mathcal{X} = \mathbb{R}^n$，输出空间$\mathcal{Y} = \{-1, 1\}$。模型的决策函数空间是有限集合$\mathcal{F} = \{f_1, f_2, ...f_d\}$，损失函数是0-1损失。

Theorem 1.2.1. 对于二类分类问题，当假设空间是有限个函数的集合$\mathcal{F} = \{f_1, f_2, ...f_d\}$时，对于任意$f \in \mathcal{F}$，至少以概率$1-\delta$，以下关于期望风险$R(f)$和经验风险$\hat{R}(f)$关系的不等式成立

$$R(f) \le \hat{R}(f) + \sqrt{\frac{\log(d/\delta)}{2N}} \tag{1.2}$$

Chapter 2

逻辑回归

2.1 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

$$\begin{aligned}P(Y=1|\boldsymbol{x})=&\frac{\exp(\boldsymbol{\theta}^T\boldsymbol{x})}{1+\exp(\boldsymbol{\theta}^T\boldsymbol{x})}\\P(Y=0|\boldsymbol{x})=&\frac{1}{1+\exp(\boldsymbol{\theta}^T\boldsymbol{x})}\end{aligned}\tag{2.1}$$

2.1.1 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

$$\begin{aligned}\log L(\boldsymbol{\theta}) &= \sum_{i=1}^{N}\left(y_i \log \pi(\boldsymbol{x}_i) + (1-y_i)\log(1-\pi(\boldsymbol{x}_i))\right)\\ &= \sum_{i=1}^{N}\left(y_i \log \frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)} + \log(1-\pi(\boldsymbol{x}_i))\right)\\ &= \sum_{i=1}^{N}\left(y_i(\boldsymbol{\theta}^T\boldsymbol{x}_i) - \log(1+\exp(\boldsymbol{\theta}^T\boldsymbol{x}_i))\right)\end{aligned} \tag{2.2}$$

2.1.1.1 参数估计：梯度下降法

根据公式 (2.2)，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

$$\begin{aligned}\boldsymbol{\theta} :=&\boldsymbol{\theta} + \alpha\frac{\partial}{\partial\boldsymbol{\theta}}\log L(\boldsymbol{\theta})\\ =&\boldsymbol{\theta} + \alpha\sum_{i=1}^{N}\left(y_i - \pi(\boldsymbol{x}_i)\right)\boldsymbol{x}_i\end{aligned} \tag{2.3}$$

其中α称为学习率，是一个正常数。

公式 (2.3)可以用矩阵表示

$$\boldsymbol{\theta} := \boldsymbol{\theta} + \alpha X^T\boldsymbol{\Lambda} \tag{2.4}$$

其中$\boldsymbol{\Lambda} = \begin{pmatrix} y_1 - \pi(\boldsymbol{x}_1)\\ y_2 - \pi(\boldsymbol{x}_2)\\ \cdots\\ y_N - \pi(\boldsymbol{x}_N)\end{pmatrix}_{N\times 1}$，$X$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)。

2.1.1.2 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：

$$\boldsymbol{\theta} := \boldsymbol{\theta} + \alpha\left(y_i - \pi(\boldsymbol{x}_i)\right)\boldsymbol{x}_i \tag{2.5}$$

与公式 (2.3)相比，随机梯度下降的迭代公式 (2.5)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。

2.2 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

$$P(Y = k|\boldsymbol{x}) = \frac{\exp(\boldsymbol{\theta}_k^T \boldsymbol{x})}{\sum\limits_{j=1}^{K} \exp(\boldsymbol{\theta}_j^T \boldsymbol{x})} \tag{2.6}$$

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

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

2.2.1 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

$$\log L(\boldsymbol{\Theta}) = \sum_{i=1}^{N}\sum_{k=1}^{K} \mathbf{1}_{\mathrm{y_i=k}} \log \pi_k(\boldsymbol{x}_i) \tag{2.7}$$

2.2.1.1 参数估计：梯度下降法

首先求

$$\frac{\partial \pi_k(\boldsymbol{x}_i)}{\partial \boldsymbol{\theta}_k} = \frac{\boldsymbol{x}_i \exp(\boldsymbol{\theta}_k^T \boldsymbol{x}_i)\left(\sum\limits_{j=1}^{K}\exp(\boldsymbol{\theta}_j^T \boldsymbol{x}) - \exp(\boldsymbol{\theta}_k^T \boldsymbol{x}_i)\right)}{\left(\sum\limits_{j=1}^{K}\exp(\boldsymbol{\theta}_j^T \boldsymbol{x})\right)^2} \tag{2.8}$$

故根据公式 (2.7)，得到Softmax模型的对数似然函数的梯度

$$\begin{aligned}
\nabla_{\boldsymbol{\theta}_k} \log L(\boldsymbol{\Theta}) &= \sum_{i=1}^{N} \mathbf{1}_{\mathrm{y_i=k}} \frac{1}{\pi_k(\boldsymbol{x}_i)} \frac{\partial \pi_k(\boldsymbol{x}_i)}{\partial \boldsymbol{\theta}_k} \\
&= \sum_{i=1}^{N} \mathbf{1}_{\mathrm{y_i=k}} \frac{1}{\pi_k(\boldsymbol{x}_i)} \frac{\boldsymbol{x}_i \exp(\boldsymbol{\theta}_k^T \boldsymbol{x}_i)\left(\sum\limits_{j=1}^{K}\exp(\boldsymbol{\theta}_j^T \boldsymbol{x}_i) - \exp(\boldsymbol{\theta}_k^T \boldsymbol{x}_i)\right)}{\left(\sum\limits_{j=1}^{K}\exp(\boldsymbol{\theta}_j^T \boldsymbol{x}_i)\right)^2} \\
&= \sum_{i=1}^{N} \mathbf{1}_{\mathrm{y_i=k}} \frac{\boldsymbol{x}_i \left(\sum\limits_{j=1}^{K}\exp(\boldsymbol{\theta}_j^T \boldsymbol{x}) - \exp(\boldsymbol{\theta}_k^T \boldsymbol{x}_i)\right)}{\sum\limits_{j=1}^{K}\exp(\boldsymbol{\theta}_j^T \boldsymbol{x})} \\
&= \sum_{i=1}^{N} \mathbf{1}_{\mathrm{y_i=k}} \boldsymbol{x}_i \left(1 - \pi_k(\boldsymbol{x}_i)\right)
\end{aligned} \tag{2.9}$$

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

$$\boldsymbol{\theta}_k := \boldsymbol{\theta}_k + \alpha \sum_{i=1}^{N} \mathbf{1}_{\mathrm{y_i=k}} \boldsymbol{x}_i \left(1 - \pi_k(\boldsymbol{x}_i)\right) \tag{2.10}$$

公式 (2.10)的迭代关系用矩阵可以表示为

$$\boldsymbol{\theta}_k := \boldsymbol{\theta}_k + \alpha X^T \boldsymbol{\Lambda} \tag{2.11}$$

其中$\mathbf{\Lambda} = \begin{pmatrix} \mathbf{1}_{\mathrm{y_1=k}}\left(1-\pi_k(\boldsymbol{x}_1)\right) \\ \mathbf{1}_{\mathrm{y_2=k}}\left(1-\pi_k(\boldsymbol{x}_2)\right) \\ ... \\ \mathbf{1}_{\mathrm{y_N=k}}\left(1-\pi_k(\boldsymbol{x}_N)\right) \end{pmatrix}_{N\times 1}$，$X$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 3

主成分分析

主成分分析（Principal Component Analysis, PCA）是一种常见的**数据降维**方法，其目的是在信息量损失较小的前提下，将高维的数据转换到低维，从而减小计算量。实质就是找到一些投影方向，使得数据在这些投影方向上包含的信息量最大，而且这些投影方向是相互正交的。选择其中一部分包含最多信息量的投影方向作为新的数据空间，同时忽略包含较小信息量的投影方向，从而达到降维的目的。

样本的**信息量**可以理解为是样本在特征方向上投影的方差。方差越大，则样本在该特征上的差异就越大，因此该特征就越重要。参见《机器学习实战》上的图，在分类问题里，样本的方差越大，越容易将不同类别的样本区分开。

PCA的数学原理，就是对原始的空间中顺序地找一组相互正交的坐标轴，第一个轴是使得方差最大的，第二个轴是在与第一个轴正交的平面中使得方差最大的，第三个轴是在与第1、2个轴正交的平面中方差最大的，这样假设在N维空间中，可以找到N个这样的坐标轴，取前r 个去近似这个空间，这样就从一个N 维的空间压缩到r 维的空间了，但是最终选择的r 个坐标轴能够使得数据的损失最小。

3.1 主成分分析的算法

假设

- 存在n个原始数据，每个数据有p个特征，用矩阵表示为$\boldsymbol{Z}_{n\times p} = (\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_n)^T$, 其中$\boldsymbol{z}_i$ 为p维列向量。

1. 去除平均值，即中心化，将数据**中心化**变换为$\boldsymbol{X}_{n\times p} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n)^T$, 其中$\boldsymbol{X} =$

$\boldsymbol{Z} - \mathbb{E}\boldsymbol{Z}$(具体而言$\boldsymbol{x}_i = \boldsymbol{z}_i - \boldsymbol{\mu}$, $\boldsymbol{\mu} = \frac{1}{n}\sum\limits_{i=1}^{n}\boldsymbol{z}_i$)。

2. 计算X的协方差矩阵，用$\boldsymbol{\Sigma}_{p\times p}$ 表示

$$\mathrm{Var}X = \mathrm{Var}(Z - \mathbb{E}Z) = \mathrm{Var}Z = \boldsymbol{\Sigma}$$

   实际上X的协方差矩阵就是原始数据Z的协方差矩阵。

3. 计算协方差矩阵$\boldsymbol{\Sigma}$的特征向量$\{\boldsymbol{\xi}_j\}$和特征值$\{\lambda_j\}$, $j = 1..p$。

4. 将特征值从小到大排序。

5. 保留前若干个特征值对应的特征向量, 假设保留的特征值为$\{\lambda_j^*\}$, $j = 1..q$, 对应的特征向量构成的矩阵为$\boldsymbol{\Xi}_{p\times q} = (\boldsymbol{\xi}_1^*, \boldsymbol{\xi}_2^*, ..., \boldsymbol{\xi}_q^*)$

6. 将数据集X转换到上述q个特征向量构建的新的空间中, 得到新的数据集$\boldsymbol{X}_{n\times q}^* = \boldsymbol{X}\boldsymbol{\Xi} = \begin{pmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ ... \\ \boldsymbol{x}_i \\ ... \\ \boldsymbol{x}_n \end{pmatrix} (\boldsymbol{\xi}_1^*, \boldsymbol{\xi}_2^*, ..., \boldsymbol{\xi}_q^*)$

3.2 主成分分析的数学原理

3.2.1 几个重要的定理

Theorem 3.2.1. $\boldsymbol{\Sigma}$为对称矩阵，如下优化问题的解$\boldsymbol{u}^*$是$\boldsymbol{\Sigma}$的最大特征值对应的特征向量。

$$\boldsymbol{u}^* = \mathop{\arg\max}_{\|u\|=1} \left(\boldsymbol{u}^T\boldsymbol{\Sigma}\boldsymbol{u}\right)$$

证明. 实际上约束条件$\|\boldsymbol{u}\| = 1$等价于$\boldsymbol{u}^T\boldsymbol{u} = 1$

利用拉格朗日乘子法，得到

$$G(\boldsymbol{u};\lambda) = \boldsymbol{u}^T\boldsymbol{\Sigma}\boldsymbol{u} + \lambda(\boldsymbol{u}^T\boldsymbol{u} - 1)$$

对G求u的偏导得到

$$\nabla_{\boldsymbol{u}} G(\boldsymbol{u};\lambda) = 2\boldsymbol{\Sigma}\boldsymbol{u} + 2\lambda\boldsymbol{u}$$

如果$\boldsymbol{u}^*$是优化问题的解，那么$\boldsymbol{u}^*$满足

$$
\begin{aligned}
&\nabla_{\boldsymbol{u}} G(\boldsymbol{u};\lambda)\mid_{\boldsymbol{u}=\boldsymbol{u}^*} = 0\\
\Rightarrow &\boldsymbol{\Sigma}\boldsymbol{u}^* = -\lambda\boldsymbol{u}^*\\
\Rightarrow &\boldsymbol{\Sigma}\boldsymbol{u}^* = \lambda^*\boldsymbol{u}^* \quad (\text{令}\lambda^* = -\lambda)
\end{aligned}
$$

所以$\boldsymbol{u}^*$是矩阵$\boldsymbol{\Sigma}$的特征向量，对应的特征值为λ^*。

当$\boldsymbol{u} = \boldsymbol{u}^*$时，目标函数

$$
\boldsymbol{u}^{*T}\boldsymbol{\Sigma}\boldsymbol{u}^* = \lambda^* \tag{3.1}
$$

为使得公式 (3.1)最大，必须使得λ^*最大，即λ^*等于矩阵$\boldsymbol{\Sigma}$最大的特征值，那么对应的特征向量便是真正的解$\boldsymbol{u}^*$。

□

3.2.2 最大方差投影

用$\boldsymbol{u}_{p\times 1}$表示某投影方向上的单位向量，那么$\boldsymbol{x}_i$在$\boldsymbol{u}$ 上的投影可以表示为

$$
<\boldsymbol{x}_i, \boldsymbol{u}> = \boldsymbol{x}_i^T\boldsymbol{u}
$$

那么数据集$\boldsymbol{X}$在$\boldsymbol{u}$ 上的投影向量为$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{u}$，可知$\boldsymbol{Y}$的均值和方差为

$$
\begin{aligned}
\mathbb{E}\boldsymbol{Y} &= \mathbb{E}\boldsymbol{X}\boldsymbol{u}\\
\mathrm{Var}\boldsymbol{Y} &= \mathrm{Var}\boldsymbol{X}\boldsymbol{u} = \boldsymbol{u}^T(\mathrm{Var}\boldsymbol{X})\boldsymbol{u} = \boldsymbol{u}^T\boldsymbol{\Sigma}\boldsymbol{u}
\end{aligned}
$$

主成分分析就是要到一个方向，使得数据集$\boldsymbol{X}$在该方向上投影方差最大。如果用单位向量$\boldsymbol{u}_1$来表示这个方向，那么$\boldsymbol{u}_1$是如下优化问题的解

$$
\underset{\|u\|=1}{\arg\max} \quad \boldsymbol{u}^T\boldsymbol{\Sigma}\boldsymbol{u}
$$

根据定理 3.2.1，u_1就是$\boldsymbol{\Sigma}$的最大特征值对应的特征向量，也是第一个**主成分**的单位向量。

随后要求第二个主成分，用单位向量$\boldsymbol{u}_2$表示这个方向。根据原理，第二个主成分依然是要最大化投影的方差，但是约束条件要多一个，即$\boldsymbol{u}_2$与$\boldsymbol{u}_1$正交。所以$\boldsymbol{u}_2$是如下优化问题的解

$$
\begin{aligned}
&\underset{\|u\|=1}{\arg\max} \quad \boldsymbol{u}^T\boldsymbol{\Sigma}\boldsymbol{u}\\
&s.t. \quad \boldsymbol{u}^T\boldsymbol{u}_1 = 0
\end{aligned}
$$

为求解上述优化问题，继续使用拉格朗日乘子法，得到

$$G(\boldsymbol{u};\boldsymbol{u}_1,\lambda,\gamma)=\boldsymbol{u}^T\boldsymbol{\Sigma}\boldsymbol{u}+(\boldsymbol{u}^T\boldsymbol{u}-1)+\gamma\boldsymbol{u}^T\boldsymbol{u}_1$$

对G求$\boldsymbol{u}$的偏导得到，

$$\nabla_{\boldsymbol{u}}G(\boldsymbol{u};\boldsymbol{u}_1,\lambda,\gamma)=2\boldsymbol{\Sigma}\boldsymbol{u}+2\lambda\boldsymbol{u}+\gamma\boldsymbol{u}_1$$

如果$\boldsymbol{u}^*$是优化问题的解，那么$\boldsymbol{u}^*$满足

$$2\boldsymbol{\Sigma}\boldsymbol{u}^*+2\lambda\boldsymbol{u}^*+\gamma\boldsymbol{u}_1=0 \tag{3.2}$$

对公式 (3.2)两边同乘以u_1^T，得到

$$\begin{aligned}
&2\boldsymbol{\Sigma}\boldsymbol{u}^*\boldsymbol{u}_1^T+2\lambda\boldsymbol{u}^*\boldsymbol{u}_1^T+\gamma\boldsymbol{u}_1\boldsymbol{u}_1^T=0\\
\Rightarrow\quad&\gamma\boldsymbol{u}_1\boldsymbol{u}_1^T=0\quad(\text{因为}u_1,u^*\text{正交})\\
\Rightarrow\quad&\gamma=0\quad\big(\text{因为}u_1u_1^T=1\big)
\end{aligned}$$

于是公式 (3.2)等于

$$\begin{aligned}
&\boldsymbol{\Sigma}\boldsymbol{u}^*+\lambda\boldsymbol{u}^*=0\\
\Rightarrow&\boldsymbol{\Sigma}\boldsymbol{u}^*=-\lambda\boldsymbol{u}^*\\
\Rightarrow&\boldsymbol{\Sigma}\boldsymbol{u}^*=\lambda^*\boldsymbol{u}^*\quad(\text{令}\lambda^*=-\lambda)
\end{aligned} \tag{3.3}$$

$\boldsymbol{u}^*$也是矩阵$\boldsymbol{\Sigma}$的特征值，与定理 3.2.1的证明类似，优化问题的目标函数等于λ^*，为了使目标函数最大，要使λ^*尽量大，而λ^*最大可取所有特征值中第二大的，对应的特征向量就是优化问题的解，即**第二主成分**的方向向量。

后面的主成分算法同理类推。

Chapter 4

附录：数学基本方法

4.1 梯度

本节内容参考了Xu (2016)。

4.1.1 梯度与方向导数

函数在某点的

- **导数**表示函数曲线上的切线斜率，也表示函数在该点的变化率。
- **偏导数**是函数关于其某一个变量的导数，物理含义是函数沿着坐标轴正方向上的变化率。
- **方向导数**是函数在某点沿某个指定方向上的变化率。
- **梯度**是函数在该点沿所有方向变化率最大的那个方向，是一个向量。

4.1.2 梯度下降

设函数$u = u(x, y)$在点$P_0(x_0, y_0)$的某空间领域内$U \subset \mathbb{R}$有定义，l为从点P_0出发的射线，$P(x, y)$为l上且在U内的任一点, 令

$$
\begin{aligned}
x - x_0 =& \Delta x = t\cos\alpha \\
y - y_0 =& \Delta y = t\sin\alpha
\end{aligned}
$$

以$t=\sqrt{(\Delta x)^2+(\Delta y)^2}$ 表示P与P_0之间的距离，若极限

$$\lim_{t\to 0^+}\frac{u(P)-u(P_0)}{t}=\lim_{t\to 0^+}\frac{u(x_0+t\cos\alpha,y_0+t\sin\alpha)-u(P_0)}{t}$$

存在，则称此极限为函数$u=u(x,y)$在点$P_0(x_0,y_0)$沿着方向l的**方向导数**，记做$\frac{\partial u}{\partial l}\big|_{P_0}$。

Lemma 4.1.1. 假设函数$u=u(x,y)$在点$P_0(x_0,y_0)$可微，则$u=u(x,y)$在点$P_0(x_0,y_0)$沿着方向l的方向导数都存在，且

$$\frac{\partial u}{\partial l}\big|_{P_0}=u_x^{'}(P_0)\cos\alpha+u_y^{'}(P_0)\sin\alpha \tag{4.1}$$

假设函数$u=u(x,y)$在点$P_0(x_0,y_0)$存在一阶偏导数，则定义

$$\nabla u\big|_{P_0}=\Big(u_x^{'}(P_0),u_y^{'}(P_0)\Big)$$

为函数$u=u(x,y)$在点$P_0(x_0,y_0)$的**梯度**。

Theorem 4.1.1. 函数$u=u(x,y)$在点$P_0(x_0,y_0)$处的方向导数在其梯度方向上达到最大值，此最大值为梯度的模。

证明. 根据方向导数的定义

$$\begin{aligned}
\frac{\partial u}{\partial l}\big|_{P_0} =&u_x^{'}(P_0)\cos\alpha+u_y^{'}(P_0)\sin\alpha\\
=&\Big(u_x^{'}(P_0),u_y^{'}(P_0)\Big)\cdot(\cos\alpha,\sin\alpha)\\
=&\nabla u\big|_{P_0}\cdot l\\
=&\big|\nabla u\big|_{P_0}\big|\big|l\big|\cos\theta\\
=&\big|\nabla u\big|_{P_0}\big|\cos\theta
\end{aligned}$$

其中θ 为梯度向量$\nabla u\big|_{P_0}$与方向向量l的夹角。根据公式(公式 (4.2))可知，当夹角$\cos\theta=1$，即$\theta=0$时，方向导数$\frac{\partial u}{\partial l}\big|_{P_0}$取得最大值，此最大值为$\big|\nabla u\big|_{P_0}\big|$。 □

Chapter 5

附录：概率论中的重要结论

5.1 Hoeffding不等式

本节内容参考了Rigollet (2015)。

Lemma 5.1.1 (Hoeffding 引理)**.** 如果随机变量$Z \in [a, b] \quad a.s$，并且$\mathbb{E}Z = 0$，那么对于$\forall s \in \mathbb{R}$，都有

$$\mathbb{E}e^{sZ} \leq e^{\frac{s^2(b-a)^2}{8}} \tag{5.1}$$

证明. 令 $\psi(s) = \log \mathbb{E}e^{sZ}$，由于$Z \in [a, b] \quad a.s$，对于给定的$s$，$e^{sZ}$是有界函数。

根据勒贝格控制收敛定理的结论，可知e^{sZ}的积分和求导符号可互换，那么得到

$$\begin{aligned}
\psi^{'}(s) =& \frac{\mathbb{E}Ze^{sZ}}{\mathbb{E}e^{sZ}} \\
\psi^{''}(s) =& \frac{\mathbb{E}Z^2e^{sZ}\mathbb{E}e^{sZ} - (\mathbb{E}Ze^{sZ})^2}{(\mathbb{E}e^{sZ})^2} = \mathbb{E}Z^2\left(\frac{e^{sZ}}{\mathbb{E}e^{sZ}}\right) - \left(\mathbb{E}Z\frac{e^{sZ}}{\mathbb{E}e^{sZ}}\right)^2
\end{aligned}$$

因为$\int \frac{e^{sZ}}{\mathbb{E}e^{sZ}} d\mathbb{P} = 1$，可定义Radon-Nikodym导数

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = \frac{e^{sZ}}{\mathbb{E}e^{sZ}} \tag{5.2}$$

由此引入一个新的概率测度$\mathbb{Q}$。$\psi^{''}(s)$可以看做Z在概率测度$\mathbb{Q}$ 下的方差

$$
\begin{aligned}
\psi^{''}(s) =& \mathbb{E}^{\mathbb{Q}} Z^2 - \left(\mathbb{E}^{\mathbb{Q}} Z\right)^2 \\
=& \mathrm{Var}(Z) \\
=& \mathrm{Var}\left(Z - \frac{a+b}{2}\right) \quad (\text{随机变量加减常数不影响方差}) \\
\leq& \mathbb{E}\left(Z - \frac{a+b}{2}\right)^2 \quad (\text{根据方差的定义}) \\
\leq& \mathbb{E}\left(\frac{b-a}{2}\right)^2 \quad \left(\text{根据}Z \in [a,b]\text{得到}|Z - \frac{a+b}{2}| \leq \frac{b-a}{2}\right) \\
=& \frac{(b-a)^2}{4}
\end{aligned}
$$

所以

$$
\begin{aligned}
& \psi(s) = \int_0^s \int_0^y \psi^{''}(x) dx dy \leq \int_0^s \int_0^y \frac{(b-a)^2}{4} dx dy = \frac{s^2(b-a)}{8} \\
\Rightarrow \quad & \log \mathbb{E} e^{sZ} \leq \frac{s^2(b-a)}{8} \\
\Rightarrow \quad & \mathbb{E} e^{sZ} \leq e^{\frac{s^2(b-a)^2}{8}}
\end{aligned}
$$

□

Theorem 5.1.1 (Hoeffding定理)**.** 假设$X_1, ...X_n$为n个独立的随机变量，并且$X \in [0,1] \quad a.s$，那么对于 $\forall s > 0$，都有

$$
\begin{aligned}
P\left(\frac{1}{n}\sum_{i=1}^n (X_i - \mathbb{E}X_i) > t\right) \leq e^{-2nt^2} \\
P\left(\frac{1}{n}\sum_{i=1}^n (\mathbb{E}X_i - X_i) > t\right) \leq e^{-2nt^2} \\
P\left(\left|\frac{1}{n}\sum_{i=1}^n (\mathbb{E}X_i - X_i)\right| > t\right) \leq 2e^{-2nt^2}
\end{aligned}
$$

Chapter 6

附录：信息熵

本章参考了(匿名, 2010)和(李梅, 2012)。

假设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。

6.1 相关概念

6.1.1 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

$$H(Y|X) = \mathbb{E}_{\mathrm{x,y}} I(y_j|x_i) = -\sum_x \sum_y P(x,y)\log P(y|x) \tag{6.1}$$

Lemma 6.1.1. 与公式 (6.1)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

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

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**。

6.1.2 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

$$D_{\mathrm{KL}}(P\|Q) = \sum_x P(x)\log\frac{P(x)}{Q(x)} = \mathbb{E}_{\mathrm{x}}\log\frac{P(x)}{Q(x)} \tag{6.2}$$

相对熵可以用来度量两个随机变量的“距离”。

Lemma 6.1.2. 相对熵恒大于等于零。

证明. 对于任意分布P, Q，根据公式 (6.2)，可知

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

□

6.1.3 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

$$I(X,Y) = D_{\mathrm{KL}}(P(X,Y) \| P(X)P(Y)) \tag{6.3}$$

Lemma 6.1.3. 互信息与条件熵满足如下关系

$$H(X|Y) = H(X) - I(X,Y) \tag{6.4}$$

证明. 根据公式 (6.2)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

$$\begin{aligned}
H(X) - I(X,Y) &= -\sum_x P(x) \log P(x) - \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \\
&= -\sum_x \left(\sum_y P(x,y) \right) \log P(x) - \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \\
&= -\sum_{x,y} P(x,y) \log P(x) - \sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \\
&= -\sum_{x,y} P(x,y) \left(\log P(x) + \log \frac{P(x,y)}{P(x)P(y)} \right) \\
&= -\sum_{x,y} P(x,y) \log \frac{P(x,y)}{P(y)} \\
&= -\sum_{x,y} P(x,y) \log P(x \mid y) \\
&= H(X|Y) \quad \text{(根据公式 (6.1))}
\end{aligned}$$

□

6.2 熵的性质

X的熵具有如下几个性质

- 非负性：$H(X) \geq 0$。[1]
- 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

$$H(X) = -\sum_{i=1}^{n} p_i \log p_i \leq -\sum_{i=1}^{n} p_i \log q_i \tag{6.5}$$

  当X和Y的概率分布相同时，公式 (6.5)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

$$H(Z) \geq \alpha H(X) + (1 - \alpha) H(Y) \tag{6.6}$$

Theorem 6.2.1 (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

$$H(X) \leq \log n \tag{6.7}$$

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

$$G(p, \lambda) = -\sum_{i=1}^{n} p_i \log p_i + \lambda \left(\sum_{i=1}^{n} p_i - 1 \right) \tag{6.8}$$

[1]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 (6.8)中分别对p_i和λ求导，令其为零，得到

$$\begin{aligned}\frac{\partial G(p,\lambda)}{\partial p_i} &= -\log p_i - 1 + \lambda = 0 \\ \sum_{i=1}^{n} p_i - 1 &= 0\end{aligned} \tag{6.9}$$

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 6.2.1 (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

$$\begin{aligned}H(X,Y) =& H(Y) + H(X|Y) \\ H(X,Y) =& H(X) + H(Y|X)\end{aligned} \tag{6.10}$$

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 6.2.2 (熵的凸性)**.** 证明公式 (6.6)

证明.

$$
\begin{aligned}
H(Z) &= -\sum_{i=1}^{n} \gamma_i \log \gamma_i \\
&= -\sum_{i=1}^{n} \alpha p_i \log \gamma_i - \sum_{i=1}^{n} (1-\alpha) q_i \log \gamma_i \\
&= -\sum_{i=1}^{n} \alpha p_i \log \left(\gamma_i \frac{p_i}{p_i} \right) - \sum_{i=1}^{n} (1-\alpha) q_i \log \left(\gamma_i \frac{q_i}{q_i} \right) \\
&= -\alpha \sum_{i=1}^{n} p_i \log p_i - (1-\alpha) \sum_{i=1}^{n} q_i \log q_i - \alpha \sum_{i=1}^{n} p_i \log \frac{\gamma_i}{p_i} - (1-\alpha) \sum_{i=1}^{n} q_i \log \frac{\gamma_i}{q_i} \\
&= \alpha H(X) + (1-\alpha) H(Y) - \alpha \sum_{i=1}^{n} p_i \log \frac{\gamma_i}{p_i} - (1-\alpha) \sum_{i=1}^{n} q_i \log \frac{\gamma_i}{q_i}
\end{aligned}
\tag{6.11}
$$

其中公式 (6.11)的倒数第二项

$$
\begin{aligned}
-\alpha \sum_{i=1}^{n} p_i \log \frac{\gamma_i}{p_i} &= \alpha \left(-\sum_{i=1}^{n} p_i \log \gamma_i + \sum_{i=1}^{n} p_i \log p_i \right) \\
&\geq 0 \quad (\text{根据公式 (6.5)})
\end{aligned}
$$

同理可知公式 (6.11)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 6.2.2. 条件熵小于无条件熵，即$H(X|Y) \leq H(X)$

证明.

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

根据熵的极值性，可知

$$
\begin{aligned}
& -\sum_{x} P(x|y)\log P(x|y) \le -\sum_{x} P(x|y)\log P(x) \\
\Rightarrow \quad & \sum_{x} P(x|y)\log P(x|y) - \sum_{x} P(x|y)\log P(x) \ge 0 \\
\Rightarrow \quad & H(X|Y) - H(X) \le 0
\end{aligned}
$$

所以

□

6.2.1 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

$$H(X|Y) = H(X,Y) - H(Y) \tag{6.12}$$

- 根据互信息定义展开可得

$$H(X|Y) = H(X) - I(X,Y) \tag{6.13}$$

- 根据公式 (6.12)和公式 (6.13)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

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

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

Chapter 7

附录：贝叶斯决策论

7.1 先验概率与后验概率

此部分参考了夏飞 (2017)。

7.1.1 概念

$$P(\boldsymbol{\theta}|X) = \frac{P(X|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(X)} \tag{7.1}$$

公式 (7.1)中$P(\boldsymbol{\theta}|X)$ 称为**后验概率**，$P(X|\boldsymbol{\theta})$称为**条件概率**(也是似然估计中的**似然函数**)，$P(\boldsymbol{\theta})$称为**先验概率**，$P(X)$是随机变量X 自身的概率，$P(X)$也被称为“证据”(evidence)。

7.1.2 频率学派与贝叶斯学派

对于概率的认知有频率学派和贝叶斯学派两种。

- 频率学派：模型参数是未知的定值，观测是随机变量。估计的方法是**极大似然估计**(Maximum Likelihood)，不依赖于先验概率。
- 贝叶斯学派：模型参数是随机变量，观测是定值。估计的方法是**最大后验估计**(Maximum a posteriori)：根据已有的经验和知识推断一个先验概率，然后在新证据不断积累的情况下调整这个概率。

7.1.3 参数估计：极大似然与最大后验

极大似然估计(ML)与最大后验估计(MAP)的数学方法的区别如下。

- 极大似然估计(ML)把似然函数$P(X|\boldsymbol{\theta})$取得最大值时的参数作为估计值。(对数)极大似然函数的估计参数可以写成

$$\hat{\boldsymbol{\theta}}_{ML} = \arg\max_{\boldsymbol{\theta}} P(X|\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \sum_{x\in X} \log P(x|\boldsymbol{\theta})$$

- 最大后验估计(MAP)与极大似然估计的区别在于加入了先验概率$P(\boldsymbol{\theta})$，

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_{MAP} &= \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{\theta}|X) \\
&= \arg\max_{\boldsymbol{\theta}} \frac{P(X|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(X)} \\
&= \arg\max_{\boldsymbol{\theta}} P(X|\boldsymbol{\theta})P(\boldsymbol{\theta}) \\
&= \arg\max_{\boldsymbol{\theta}} \left(\sum_{x\in X} \log P(x|\boldsymbol{\theta}) + \log P(\boldsymbol{\theta}) \right)
\end{aligned}$$

7.2 共轭分布

在贝叶斯理论中，如果后验概率$P(\boldsymbol{\theta}|X)$和先验概率$P(\boldsymbol{\theta})$满足同样的分布律，那么先验概率分布和后验概率分布就叫做**共轭分布**，同时先验分布叫做似然函数的**共轭先验分布**。

Theorem 7.2.1. 二项分布的共轭先验是Beta分布。

证明. 假设

- 先验分布为Beta分布，$\boldsymbol{\theta} \sim \boldsymbol{\beta}(\alpha, \beta)$，那么$P(\boldsymbol{\theta}|\alpha,\beta) = \frac{1}{\boldsymbol{\beta}(\alpha,\beta)}\boldsymbol{\theta}^{\alpha-1}(1-\boldsymbol{\theta})^{\beta-1}$，其中$\boldsymbol{\beta}(\alpha,\beta) = \int_0^1 \boldsymbol{\theta}^{\alpha-1}(1-\theta)^{\beta-1}d\boldsymbol{\theta} = \frac{G(\alpha)G(\beta)}{G(\alpha+\beta)}$
- 似然概率服从二项分布，$P(X=k|\boldsymbol{\theta}) = C_n^k\boldsymbol{\theta}^k(1-\boldsymbol{\theta})^{n-k}$

后验概率

$$
\begin{aligned}
P(\boldsymbol{\theta}|X=k) =&\frac{P(X=k|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(X=k)}\\
=&\frac{P(X=k|\boldsymbol{\theta})P(\boldsymbol{\theta})}{\int_0^1 P(\boldsymbol{\theta},X=k)d\boldsymbol{\theta}}\\
=&\frac{P(X=k|\boldsymbol{\theta})P(\boldsymbol{\theta})}{\int_0^1 P(X=k|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}\\
=&\frac{C_n^k\boldsymbol{\theta}^k(1-\boldsymbol{\theta})^{n-k}\frac{1}{\beta(\alpha,\beta)}\boldsymbol{\theta}^{\alpha-1}(1-\boldsymbol{\theta})^{\beta-1}}{\int_0^1 C_n^k\boldsymbol{\theta}^k(1-\boldsymbol{\theta})^{n-k}\frac{1}{\beta(\alpha,\beta)}\boldsymbol{\theta}^{\alpha-1}(1-\boldsymbol{\theta})^{\beta-1}d\boldsymbol{\theta}}\\
=&\frac{\boldsymbol{\theta}^{\alpha+k-1}(1-\boldsymbol{\theta})^{\beta+n-k-1}}{\int_0^1\boldsymbol{\theta}^{\alpha+k-1}(1-\boldsymbol{\theta})^{\beta+n-k-1}d\boldsymbol{\theta}}\\
=&\frac{\boldsymbol{\theta}^{\alpha+k-1}(1-\boldsymbol{\theta})^{\beta+n-k-1}}{B(\alpha+k,\beta+n-k)}\quad\left(\text{根据定义}B(\alpha,\beta)=\int_0^1\boldsymbol{\theta}^{\alpha-1}(1-\boldsymbol{\theta})^{\beta-1}d\boldsymbol{\theta}\right)
\end{aligned}
$$

后验概率也服从Beta分布$\boldsymbol{\theta}|X=k\sim\boldsymbol{\beta}(\alpha+k,\beta+n-k)$，故二项分布的共轭先验是Beta分布。

□

参考文献

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

Rigollet, P. (2015). Lecture note, mathematics of machine learning, mit open course. 1, 14

Xu, L.-C. (2016). 理解梯度下降. 12

匿名 (2010). 信息论与编码. 16

夏飞 (2017). 聊一聊机器学习的mle和map. 24

李梅 (2012). 信息论基础. 16

李航 (2012). 统计学习方法. 清华大学出版社. 1

索引