目录

第一部分

数学基础

第一章　微积分

1.1　导数及偏导数

1.2　复合函数的链式法则

1.3　最优化问题,KTT 条件

1.3.1　拉格朗日函数

拉格朗日函数中涉及的概念梳理（乘子、算子、balabala）拉格朗日函数的作用、构造方法、求解方式原始问题与对偶问题的关系、

1.4　傅里叶变换

第二章　概率论

2.1　条件概率与贝叶斯方法

2.2　大数定律与中心极限定理

2.3　常见分布

常见分布：0-1 分布、二项分布、高斯分布等，高斯分布很重要数据 normalized 跟它有关，参数的初始化特跟它有关；

2.4　期望与方差与归一化

2.5　协方差

2.6　最大似然估计

在推导逻辑回归的损失函数时会用到。

第三章　线性代数

3.1　标量、向量、矩阵及张量的定义及运算

3.2　线性无关于线性相关

3.2.1　矩阵 x 矩阵

3.2.2　矩阵 x 向量

$$Ax = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = x_1 a_1 + x_2 a_2 + \cdots + x_n a_n \tag{3.1}$$

3.2.3 向量 x 向量

3.3 范数

3.4 线性空间与子空间

3.4.1

3.5 特征值与特征向量

3.6 矩阵分解

3.6.1 QR 分解

3.6.2 SVD 分解

3.6.3 奇异值分解

3.6.4 矩阵分解的几何意义

3.7 常见的距离计算方式：欧式距离、曼哈顿距离、余弦距离

第二部分

机器学习

如何解决监督学习问题？在给定足够的映射样本的情况下，学习将一个向量映射到另一个。本质是用数据和模型去为现有的问题 (existing problems) 提供解决方法 (solutions).

第四章　统计学习方法

4.1　统计学习方法概述

统计学习方法的三要素

4.1.1　模型（model）

4.1.2　策略（strategy）

4.1.3　算法（algorithm）

4.1.4　模型评估

用已知预测未知（泛化能力）-训练误差与测试误差，过拟合、欠拟合

奥卡姆剃刀

当假设空间含有不同复杂度（例如，不同的参数个数）的模型时，就要面临模型选择 (modelselection）的问题．我们希望选择或学习一个合适的模型．如果在假设空间中存在“真”模型，那么所选择的模型应该逼近真模型．具体地，所选择的模型要与真模型的参数个数相同，所选择的模型的参数向量与真模型的参数 l 向量相近．如果一味追求提高对训练数据的预测能力，所选模型的复杂度则往往会比真模型更高．这种现象称为过拟合（over-fiting)．过拟合是指学习时选择的模型所包含的参数过多，以致于出现这一模型对已知数据预测得很好，但对未知数据预测得很差的现象，可以说模型选择旨在避免过拟合并提高模型的预测能力.

4.1.5　有监督学习

有监督学习是一个函数，$f(x) = y$

4.1.5.1 分类

4.1.5.2 标注

4.1.5.3 回归

4.1.5.4 生成模型和判别模型

4.1.6 无监督学习

4.1.7 半监督学习

4.1.8 强化学习

4.2 感知机

用平面二分类，样本点在平面两侧

感知机的函数表示空间上的含义函数距离与几何距离函数与空间剖面的关系感知机的训练方法感知机的对偶形式及其训练方法拓展：感知机的收敛性拓展：感知机为什么不能表示异或

4.3 线性回归

用于拟合平面，样本点在平面附近

4.4 优化方法简述

感知机的优化方式

4.5 朴素贝叶斯

4.5.1 后验概率最大化的含义

$$
\begin{aligned}
R_{exp}(f) = E[L(Y, f(X))] &= \int_{X\times Y} L(y, f(x))P(x, y)dxdy \\
&= \iint_{X\times Y} L(y, f(x))P(y|x)P(x)dxdy \\
&= \int_X \left(\sum_{k=1}^{K} L\left(c_k, f(x)\right) P\left(y = c_k|x\right) \right) P(x)d(x) \\
&= E_X \sum_{k=1}^{K} L\left(c_k, f(X)\right) P\left(c_k|X\right)
\end{aligned}
\tag{4.1}
$$

因此对所有的 X=x 逐个极小化

$$\begin{aligned} f(x) &= \arg\min_{y \in Y} \sum_{k=1}^{K} L\left(c_k, y\right) P\left(c_k | X = x\right) \\ &= \arg\min_{y \in Y} \sum_{k=1}^{K} P\left(y \neq c_k | X = x\right) \\ &= \arg\min_{y \in Y} \left(1 - P\left(y = c_k | X = x\right)\right) \\ &= \arg\max_{y \in Y} P\left(y = c_k | X = x\right) \end{aligned} \tag{4.2}$$

对于不同的书，对于一些算法或算法行为有着不同的称谓，有的概念名称甚至连原书作者也拿捏不准，这也导致我们在初学翻阅各种资料时候发现一会儿又多了这个概念，一会儿又多了那个概念，及其痛苦。但是名称不重要，重要的是我们知道所指代的具体东西就行。下面就整理出笔者在学习中遇到过的各种“叫法”，仅供参考。

4.5.2 朴素贝叶斯的参数估计

4.5.3 贝叶斯估计

贝叶斯估计实际上就是在朴素贝叶斯的基础上添加了一个平滑性, 这种**估计**方法称为贝叶斯估计

$$P_{\lambda}\left(Y = c_k\right) = \frac{\sum_{i=1}^{m} I\left(y_i = c_k\right) + \lambda}{m + K\lambda} \tag{4.3}$$

$$P_{\lambda}\left(X^{(j)} = a_{jl} | Y = c_k\right) = \frac{\sum_{i=1}^{m} I\left(x_i^{(j)} = a_{jl}, y_i = c_k\right) + \lambda}{\sum_{i=1}^{m} I\left(y_i = c_k\right) + S_j \lambda} \tag{4.4}$$

4.5.4 贝叶斯决策

贝叶斯决策是一种统计决策理论，用于设计分类器，针对分类任务。朴素贝叶斯是基于贝叶斯决策理论, 假设条件独立性后的一种具体的分类器算法.

4.5.5 贝叶斯网络

如果不假设条件独立性, 而是认为条件之间存在概率依存关系, 那么模型就变成了贝叶斯网络

4.6 逻辑回归

4.6.1 线性回归

4.6.2 二项回归、多项回归

4.6.3 对分类条件概率建模

4.6.4 激活函数

4.7 广义线性模型

4.7.1 指数分布族

4.8 最大熵模型

4.8.1 熵、信息熵、条件熵...

条件熵 H(Y|X) 定义为 X 给定条件下 Y 的条件概率分布的熵对 X 的**数学期望**:

$$\begin{aligned} H(Y|X) &= \sum_x p(x) H(Y|X=x) \\ &= -\sum_x p(x) \sum_y p(y|x) \log p(y|x) \\ &= -\sum_x \sum_y p(x,y) \log p(y|x) \\ &= -\sum_{x,y} p(x,y) \log p(y|x) \end{aligned} \tag{4.5}$$

4.8.2 最大熵原理及其解释

4.8.2.1 证明均匀分布时熵最大

4.8.2.2 推导条件熵的公式

$$
\begin{aligned}
H(X,Y) &= -\sum_{x,y} p(x,y)\log p(x,y) \\
&= -\sum_{x,y} p(x,y)\log(p(y|x)p(x)) \\
&= -\sum_{x,y} p(x,y)\log p(y|x) - \sum_{x,y} p(x,y)\log p(x) \\
&= H(Y|X) - \sum_{x,y} p(x,y)\log p(x) \\
&= H(Y|X) - \sum_{x}\sum_{y} p(x,y)\log p(x) \\
&= H(Y|X) - \sum_{x}\log p(x)\sum_{y} p(x,y) \\
&= H(Y|X) - \sum_{x}(\log p(x))p(x) \\
&= H(Y|X) - \sum_{x} p(x)\log p(x) \\
&= H(Y|X) + H(X)
\end{aligned}
\tag{4.6}
$$

4.8.3 最大熵模型中的特征函数

最大熵模型中的每个特征会有一个权重，你可以把它理解成这个特征所描述的输入和输出有多么倾向于同时出现。由于特征函数是人为定义的，因此可以理解为，通过人的常识，认为添加的对模型的约束条件？

4.8.4 最大熵模型的优化

4.8.4.1 最优化问题的模型表示

引入算子，表示为拉格朗日函数，介绍凸函数，KKT 条件

4.8.4.2 对偶函数和极大似然估计形式的等价性

对偶函数 $\psi(w)$ 的推导

$$
\begin{aligned}
\psi(w) &= \min_{p \in C} L(P, w) \\
&= \sum_{x,y} \tilde{P}(\mathrm{x}) + \mathrm{P_w(y|x)} \log P_{\mathrm{w}}(y|x) + w_0 \left(1 - \sum_y P_w(y|x)\right) \\
&\quad + \sum_{i=1}^{n} w_i \left(\sum_{xy} \tilde{P}(x,y) f_i(x,y) - \sum_{xy} \mathrm{P(x)P}_w\mathrm{(y|x)} f_i(x,y) \right) \\
&= \sum_{x,y} \mathrm{P(x)P_w(y|x)} \log P_{\mathrm{w}}(y|x) \\
&\quad + \sum_{\mathrm{i}=1}^{n} w_i \left(\sum_{x,y} \tilde{P}(x,y) f_i(x,y) - \sum_{x,y} \mathrm{P(x)P}_w\mathrm{(y|x)} f_i(x,y) \right) \\
&= \sum_{x,y} \sum_{i=1}^{n} w_i \tilde{P}(x,y) f_i(x,y) \\
&\quad + \sum_{x,y} \tilde{P}(\mathrm{x}) \mathrm{P_w(y|x)} \left[\log P_{\mathrm{w}}(y|x) - \sum_{i=1}^{n} w_i f_i(x,y) \right] \\
&= \sum_{x,y} \sum_{i=1}^{n} w_i \tilde{P}(x,y) f_i(x,y) - \sum_{x,y} \tilde{P}(\mathrm{x}) \mathrm{P_w(y|x)} \log Z_w(x) \\
&= \sum_{x,y} \sum_{i=1}^{n} w_i \tilde{P}(x,y) f_i(x,y) - \sum_{x} \tilde{\mathrm{P}}(\mathrm{x}) \log Z_w(x)
\end{aligned}
\tag{4.7}
$$

4.8.5 最大熵模型和逻辑回归的等价性

4.8.6 实例

4.9 支持向量机

4.9.1 正则化-岭回归与 Lasso 回归

正则化项本质上是一种先验信息，整个最优化问题从贝叶斯观点来看是一种贝叶斯最大后验估计，其中正则化项对应后验估计中的先验信息，损失函数对应后验估计中的似然函数，两者的乘积即对应贝叶斯最大后验估计的形式，如果你将这个贝叶斯最大后验估计的形式取对数，即进行极大似然估计，你就会发现问题立马变成了损失函数 + 正则化项的最优化问题形式

支持向量机 (SVM) 是 90 年代中期发展起来的基于统计学习理论的一种机器学习方法，通过寻求结构化风险最小来提高学习机泛化能力，实现经验风险和置信范围的最小化，从而达到在统计样本量较少的情况下，亦能获得良好统计规律的目的。

核方法和核函数

4.10 序列建模

4.10.1 隐马尔科夫模型

4.10.2 最大熵马尔可夫模型

4.10.3 条件随机场

4.10.4 结构化感知机

4.10.5 总结

这四种模型的区别，各自的优缺点

4.11 K 近邻

4.12 决策树

4.13 优化方法

4.13.1 牛顿法和拟牛顿法

4.13.2 迭代尺度法

4.13.3 梯度下降法

第五章　特征工程与数据挖掘

5.1　数据审查

5.1.1　集中趋势

5.1.2　离散趋势

5.1.3　分布趋势

5.1.4　离群点

5.2　数据清洗

5.3　数据集成

5.3.1　实体识别

5.3.2　冗余属性识别

5.4　数据变换

5.5　数据规约

第六章　深度学习

6.1　多层感知机

6.2　卷积、池化

6.3　循环神经网络

6.3.1　RNN

6.3.2　LSTM

6.3.3　GRU

6.3.4　Bidirection

6.4　其他网络结构

6.4.1　归一化结构

6.4.2　Residual

6.4.3　Attention

6.4.4　卷积的变种

6.4.4.1　空洞卷积

等等等等

6.5 经典神经网络结构

6.5.1 DeepCNN

6.5.2 Seq2Seq

6.5.3 Transformer

第七章　机器学习经验

7.1　权重初始化

7.2　数据采样方式

7.3　BatchSize 的选择

7.4　损失函数的选择

7.4.1　交叉熵

7.4.2　均方差

7.4.3　Hinge

第八章　机器学习应用

8.1　知识图谱

8.2　自然语言处理

8.2.1　词向量

每一种的表示方法，生成方法，优缺点

8.2.1.1 N-gram

8.2.1.2 共现矩阵

8.2.1.3 Word2Vec

8.2.1.4 ELMo/BERT

8.2.2 分词与序列标注

8.3 图形学

8.3.1 图像识别

8.3.1.1 Resnet

8.3.1.2 U-net

8.3.1.3 WaveNet

8.3.2 超分辨率

8.3.3 图像分隔

8.3.4 语义识别

8.4 语音识别

8.4.1 去噪

8.4.2 提取特征

8.4.3 损失函数

8.4.4 模型结构

输入是语音，输出是音素或汉字/英语

8.5 推荐系统？

第九章　杂篇（暂时不清楚放到哪里的）

9.1　对几种概率模型的第二种理解

逻辑回归的建模需求 sigmoid 函数到 logit 函数到 logistic 函数广义线性模型与逻辑回归的关系从二项分布到多项分布到多项回归逻辑回归到最大熵模型

9.2　特征函数的用途

9.3　向量空间模型

9.4　流型

9.5　非关系型数据库

9.6　判别模型和生成模型

9.7　无监督学习

9.8　概率图模型

9.8.1　从概率图模型出发解释 HMM、CRF...

第十章　推荐阅读

10.1　数学篇

http://www.medicine.mcgill.ca/epidemiology/hanley/bios601/Mean-Quantile/intro-normal-distribution-2.pdf