



子午线

学习笔记

作者: leekarry

组织: 果壳

时间: June 1, 2019

版本: 0.1



我很快,快到时间都会变慢。而我一慢,时间就会过得飞快。

目 录

I 机器学习	1
1 机器学习概论	2
1.1 基本概念	2
1.2 统计学习三要素	3
1.3 模型评估与模型选择	3
1.4 正则化与交叉验证	3
1.5 生成模型与判别模型	4
1.6 频率学派和贝叶斯学派的参数估计	4
2 概率分布	6
2.1 二元分布	6
2.2 多项式变量	8
2.3 高斯分布	10
2.4 指数族分布	17
2.5 非参数化方法	21
3 变分法	24
4 矩阵的性质	25
5 回归的线性模型	26
5.1 线性基函数模型	26
5.2 偏置-方差分解	29
5.3 贝叶斯线性回归	30
5.4 贝叶斯模型比较	33
5.5 证据近似	34
5.6 固定基函数的局限性	35
6 分类的线性模型	36
6.1 判别函数	36
6.2 概率生成式模型	40
6.3 概率判别式模型	40
6.4 拉普拉斯近似	41
6.5 贝叶斯 logistic 回归	41

7 神经网络	42
7.1 前馈神经网络	42
7.2 网络训练	43
7.3 误差反向传播	43
7.4 Hessian 矩阵	43
7.5 神经网络的正则化	43
7.6 混合密度网络	43
7.7 贝叶斯神经网络	43
8 组合模型	44
8.1 贝叶斯模型平均	44
8.2 委员会	44
8.3 提升方法	44
8.4 基于树的模型	46
8.5 条件混合模型	46
8.6 logistic 模型的混合	46
9 支持向量机	47
9.1 间隔与支持向量	47
9.2 对偶问题	48
9.3 序列最小最优算法	50
9.4 核函数	53
9.5 软间隔与正则化	55
9.6 支持向量回归	55
9.7 核方法	55
10 混合模型和 EM 算法	56
10.1 一般形式的 EM 算法	56
10.2 EM 的另一种观点	59
10.3 K 均值聚类	59
10.4 混合高斯	59
11 概率图模型	63
11.1 贝叶斯网络	63
11.2 条件独立	63
11.3 马尔科夫随机场	63
11.4 图模型中的推断	63
11.5 隐马尔可夫模型	63
11.6 条件随机场	70

12 采样方法	71
12.1 基本采样算法	71
12.2 马尔科夫链蒙特卡罗	71
12.3 吉布斯采样	71
12.4 切片采样	71
12.5 混合蒙特卡罗算法	71
12.6 估计划分函数	71
13 近似推断	72
13.1 变分推断	72
13.2 变分线性回归	72
13.3 指数族分布	72
13.4 局部变分方法	72
13.5 变分 logistic 回归	72
13.6 期望传播	72
14 连续潜在变量	73
14.1 主成分分析	73
14.2 概率 PCA	75
14.3 核 PCA	76
14.4 非线性隐变量模型	76

第 I 部分 I

机器学习

第1章 机器学习概论

1.1 基本概念

统计学习的特点

统计学习 (statistical learning) 是关于计算机基于数据构建概率统计模型并运用模型对数据进行预测与分析的一门学科。

- 统计学习以计算机及网络为平台；
- 统计学习以数据为研究对象；
- 统计学习的目的是对数据进行预测与分析；
- 统计学习以方法为中心；
- 统计学习是概率论、统计学、信息论、计算理论、最优化理论及计算机科学等多个领域的交叉学科。

统计学习的对象

统计学习的对象是数据 (data)。它从数据出发, 提取数据的特征, 抽象出数据的模型, 发现数据中的知识, 又回到对数据的分析与预测中去。**统计学习关于数据的基本假设是同类数据具有一定的统计规律性, 这是统计学习的前提。**

统计学习的目的

统计学习用于对数据进行预测与分析, 特别是对未知新数据进行预测与分析。对数据的预测与分析是通过构建概率统计模型实现的。统计学习总的目标就是考虑学习什么样的模型和如何学习模型, 以使模型能对数据进行准确的预测与分析, 同时也要考虑尽可能地提高学习效率。

统计学习的方法

统计学习的方法是基于数据构建统计模型从而对数据进行预测与分析。

- 监督学习 (supervised learning)
- 非监督学习 (unsupervised learning)
- 半监督学习 (semi-supervised learning)
- 强化学习 (reinforcement learning)

实现统计学习方法的步骤如下：

- (1) 得到一个有限的训练数据集
- (2) 确定包含所有可能的模型的假设空间, 即学习模型的集合
- (3) 确定模型选择的准则, 即学习的策略

- (4) 实现求解最优化模型的算法,即学习的算法
- (5) 通过学习方法选择最优模型
- (6) 利用学习的最优模型对新数据进行预测或分析

统计学习的研究

统计学习方法 (statistical learning method), 旨在开发新的学习方法。

统计学习理论 (statistical learning theory), 旨在探求统计学习方法的有效性与效率。

统计学习应用 (application of statistical learning), 旨在将统计学习方法应用到实际问题中去, 解决实际问题。

统计学习的重要性

统计学习是处理海量数据的有效方法; 统计学习是计算机智能化的有效手段; 统计学习是计算机科学学习发展的一个重要组成部分。

1.2 统计学习三要素

- (1) 模型: 统计学习首先要考虑的问题是学习什么样的模型
- (2) 策略: 有了模型的假设空间, 统计学习接着需要考虑的是按照什么样的准则学习或选择最优的模型
- (3) 算法: 算法是指学习模型的具体计算方法

1.3 模型评估与模型选择

统计学习的目的是使学到的模型不仅对已知数据而且对未知数据都能有很好的预测能力。不同的学习方法会给出不同的模型。当损失函数给定时, 基于损失函数的模型的训练误差 (training error) 和模型的测试误差 (test error) 就自然成为学习方法评估的标准。测试误差反映了学习方法对未知的测试数据集的预测能力——泛化能力。

1.4 正则化与交叉验证

模型选择的典型方法是正则化 (regularization)。正则化是结构风险最小化策略的实现, 是在经验风险上加一个正则化项 (regularizer) 和罚项 (penalty term)

$$\min_{f \in F} \frac{1}{2} \sum_{i=1}^N L(y_i, f(x_i)) + \lambda J(f) \quad (1.1)$$

另一种常用的模型选择方法是交叉验证 (cross validation) 训练集用来训练模型验证集用来选择模型测试集用来对模型进行评估

1. 简单交叉验证

2. S 折交叉验证
3. 留一交叉验证

1.5 生成模型与判别模型

监督学习方法又可以分为生成方法 (generative approach) 和判别方法 (discriminative approach)。所学到的模型分别称为生成模型和判别模型。

生成方法由数据学习联合分布 $P(X, Y)$, 然后求出条件概率分布 $P(Y|X)$ 作为预测的模型, 即生成模型:

$$P(Y|X) = \frac{P(X, Y)}{P(X)} \quad (1.2)$$

这样的方法之所以称为生成方法, 是因为模型表示了给定输入 X 产生输出 Y 的生成关系。

判别方法由数据直接学习决策函数 $f(X)$ 或者条件概率分布 $P(Y|X)$ 作为预测的模型, 即判模型。判别方法关心的是对给定的输入 X , 应该预测什么样的输出 Y 。

生成方法的特点: 生成方法可以还原出联合概率分布 $P(X|Y)$, 而判别方法则不能; 生成方法的学习收敛速度更快, 即当样本容量增加的时候, 学到的模型可以更快地收敛于真实模型; 当存在隐变量时, 仍可以用生成方法学习, 此时判别方法就不能用。

判别方法的特点: 判别方法直接学习的是条件概率 $P(Y|X)$ 或决策函数 $f(X)$, 直接面对预测, 往往学习的准确率更高; 由于直接学习 $P(Y|X)$ 或 $f(X)$, 可以对数据进行各种程度上的抽象、定义特征并使用特征, 因此可以简化学习问题。

1.6 频率学派和贝叶斯学派的参数估计

频率学派与贝叶斯学派的区别

简单地说, 频率学派与贝叶斯学派探讨**不确定性**这件事的出发点与立足点同。频率学派从"自然"角度出发, 试图直接为"事件"本身建模, 即事件 A 在独立重复试验中发生的频率趋于极限 p , 那么这个极限就是该事件发生的概率。贝叶斯学派并不从试图刻画"事件"本身, 而从"观察者"角度出发。贝叶斯学派并不试图说"事件本身是随机的", 或者"世界的本体带有某种随机性", 而只是从"观察者知识不完备"这一出发点开始, 构造一套在贝叶斯概率论的框架下可以对不确定知识做出推断的方法。体现在参数估计中, 频率学派认为参数是客观存在, 不会改变, 虽然未知, 但却是固定值; 贝叶斯学派则认为参数是随机值, 因此参数也可以有分布。

频率学派的参数估计

极大似然估计 (Maximum Likelihood Estimate, MLE), 也叫最大似然估计。若总体 X 属离散型 (连续型与此类似), 其分布律 $P\{X = x\} = p(x; \theta)$, $\theta \in \Theta$ 的形式为已知, θ 为待估参数, Θ 是 θ 的取值范围, 设 X_1, X_2, \dots, X_n 是来自 X 的样本, 则 X_1, X_2, \dots, X_n 的联合概率分

布为

$$\prod_{i=1}^n p(x; \theta) \quad (1.3)$$

设 x_1, x_2, \dots, x_n 是相应的样本值, 则

$$L(\theta) = L(x_1, x_2, \dots, x_n; \hat{\theta}) = \underset{\theta \in \Theta}{\operatorname{argmax}} \prod_{i=1}^n p(x; \theta) \quad (1.4)$$

贝叶斯学派的参数估计

最大后验估计 (Maximum a Posteriori estimation, MAP), 它与极大似然估计最大的区别就是, 它考虑了参数本身的分布, 也就是先验分布。最大后验估计是根据经验数据获得对难以观察的量的点估计。可以看作规则化的最大似然估计。假设 x 为独立同分布的采样, θ 为模型参数, p 为我们所使用的模型。那么最大似然估计可以表示为

$$\hat{\theta}_{MLE}(x) = \underset{\theta}{\operatorname{argmax}} p(x|\theta) \quad (1.5)$$

现在, 假设 θ 的先验分布为 g 。通过贝叶斯理论, 对于 θ 的后验分布如下式所示:

$$p(\theta|x) = \frac{p(x|\theta)g(\theta)}{\int_{\theta \in \Theta} p(x|\theta')g(\theta')d\theta'} \quad (1.6)$$

分母为 x 的边缘概率与 θ 无关, 因此最大后验等价于使分子最大, 故目标函数为

$$\hat{\theta}_{MAP}(x) = \underset{\theta}{\operatorname{argmax}} p(x|\theta)g(\theta) \quad (1.7)$$

第2章 概率分布

概率论在解决模式识别问题时起着重要作用。现在探究一下某些特殊的概率分布的例子以及它们的性质。概率分布的一个作用是在给定有限次观测 x_1, \dots, x_N 的前提下, 对随机变量 \vec{x} 的概率分布 $p(\vec{x})$ 建模。这个问题被称为密度估计 (density estimation)。本章中, 我们假设数据点是独立同分布的。

首先, 我们考虑离散随机变量的二项分布和多项式分布, 以及连续随机变量的高斯分布。这是参数分布 (parametric distribution) 的具体例子。之所以被称为参数分布, 是因为少量可调节的参数控制了整个概率分布。为了把这种模型应用到密度估计问题中, 我们需要一个步骤, 能够在给定观察数据集的条件下, 确定参数的合适的值。在频率学家的观点中, 我们通过最优化某些准则 (例如似然函数) 来确定参数的具体值。相反, 在贝叶斯观点中, 给定观察数据, 我们引入参数的先验分布, 然后使用贝叶斯定理来计算对应后验概率分布。

我们会看到, 共轭先验 (conjugate prior) 有着很重要的作用, 它使得后验概率分布的函数形式与先验概率相同, 因此使得贝叶斯分析得到了极大的简化。指数族分布有很多重要的性质, 将在本章详细讨论。

参数方法的一个限制是它假定分布有一个具体的函数形式, 这对于一个具体应用来说是不合适的。另一种替代的方法是非参数 (nonparametric) 密度估计方法。这种方法中分布的形式通常依赖于数据集的规模。这些模型仍然具有参数, 但是这些参数控制的是模型的复杂度而不是分布的形式。本章最后, 我们会考虑三种非参数化方法, 分布依赖于直方图、最近邻以及核函数。

2.1 二元分布

考虑一个二元随机变量 $x \in \{0, 1\}$ 。 $x = 1$ 的概率被记作参数 μ , 因此

$$p(x = 1|\mu) = \mu \quad (2.1)$$

其中 $0 \leq \mu \leq 1$ 。 x 的概率分布因此可以写成

$$\text{Bern}(x|\mu) = \mu^x(1 - \mu)^{1-x} \quad (2.2)$$

这被叫做伯努利分布 (Bernoulli distribution)。均值和方差为

$$\mathbb{E}[x] = \mu \quad (2.3)$$

$$\text{var}[x] = \mu(1 - \mu) \quad (2.4)$$

用最大似然估计方法求得

$$\mu_{ML} = \frac{1}{N} \sum_{n=1}^N x_n = \frac{m}{N} \quad (2.5)$$

m 为数据集里 $x = 1$ (正面朝上) 的观测数量。

我们也可以求解给定数据集规模 N 的条件下, $x = 1$ 的观测出现的数量 m 的概率分布。这被称谓二项分布 (binomial distribution)。

$$Bin(m|N, \mu) = \binom{N}{m} \mu^m (1 - \mu)^{N-m} \quad (2.6)$$

其中

$$\binom{N}{m} \equiv \frac{N!}{(N-m)!m!} \quad (2.7)$$

均值和方差为

$$\mathbb{E}[m] \equiv \sum_{m=0}^N m Bin(m|N, \mu) = N\mu \quad (2.8)$$

$$var[m] \equiv \sum_{m=0}^N (m - \mathbb{E}[m])^2 Bin(m|N, \mu) = N\mu(1 - \mu) \quad (2.9)$$

Beta 分布

我们已经看到伯努利分布的参数 μ 的最大似然解。对于小规模的数据集会给出严重的过拟合结果。为了用贝叶斯的观点看待这个问题, 我们需要引入一个关于 μ 的先验分布 $p(\mu)$ 。

在贝叶斯统计中, 如果后验分布与先验分布属于同类, 则先验分布与后验分布被称为**共轭分布**, 而先验分布被称为似然函数的共轭先验。具体地说, 就是给定贝叶斯公式, 假定似然函数 $p(x|\theta)$ 是已知的, 问题就是选取什么样的先验分布 $p(\theta)$ 会让后验分布与先验分布具有相同的数学形式。共轭先验的好处主要在于代数上的方便性, 可以直接给出后验分布的封闭形式, 否则的话只能数值计算。共轭先验也有助于获得关于似然函数如何更新先验分布的直观印象。所有指数家族的分布都有共轭先验。

因此, 我们把先验分布选择为 Beta 分布, 定义为

$$Beta(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1-\mu)^{b-1} \quad (2.10)$$

其中, $\Gamma(x)$ 定义为

$$\Gamma(x) \equiv \int_0^\infty u^{x-1} e^{-u} du \quad (2.11)$$

有以下性质

$$\Gamma(x+1) = x\Gamma(x) \quad (2.12)$$

$$\Gamma(x+1) = x! \quad (2.13)$$

Beta 分布的均值和方差为

$$\mathbb{E}[\mu] = \frac{a}{a+b} \quad (2.14)$$

$$\text{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)} \quad (2.15)$$

参数 a 和 b 被称为超参数 (hyperparameter), 因为它们控制了参数 μ 的概率分布。

μ 的后验概率分布现在可以这样得到: 把 Beta 先验与二项似然函数相乘, 然后归一化。我们看到后验概率分布的形式为

$$p(\mu|m, l, a, b) = \frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)} \mu^{m+a+1} (1-\mu)^{l+b+1} \quad (2.16)$$

其中 $l = N - m$, 即对应于硬币“反面朝上”的样本数量。我们看到, 如果一个数据集里有 m 次观测为 $x = 1$, 有 l 次观测为 $x = 0$, 那么从先验概率到后验概率, a 的值变大了 m , b 的值变大了 l 。这让我们可以简单地把先验概率中的超参数 a 和 b 分别看成 $x = 1$ 和 $x = 0$ 的有效观测数。

另外, 如果接下来观测到更多的数据, 那么后验概率分布可以扮演先验概率的角色。学习过程中的顺序 (sequential) 方法可以自然而然地得出。它与先验和似然函数的选择无关, 只取决于数据独立同分布的假设。

给定数据集 D 的情况下, x 的预测分布为

$$p(x=1|D) = \int_0^1 p(x=1|\mu)p(\mu|D)d\mu = \int_0^1 \mu p(\mu|D)d\mu = \mathbb{E}[\mu|D] \quad (2.17)$$

2.2 多项式变量

二元变量可以用来描述只能取两种可能值中的某一种这样的量。然而, 经常会遇到可以取 K 个互斥状态中的某一种的离散变量。一种比较方便的表示方法是“1-of- K ”表示法。例如, 如果我们有一个能够取 $K=6$ 种状态的变量, 这个变量的某次特定的观测恰好对应于 $x_3 = 1$ 的状态, 那么 x 就可以表示为

$$\mathbf{x} = (0, 0, 1, 0, 0, 0)^T \quad (2.18)$$

这样的向量满足 $\sum_{k=1}^K x_k = 1$ 。如果我们用参数 μ_k 表示 $x_k = 1$ 的概率,那么 x 的分布就是

$$p(\mathbf{x}|\boldsymbol{\mu}) = p(x_1, x_2, \dots, x_K | \mu_1, \dots, \mu_K) = \prod_{k=1}^K \mu_k^{x_k} \quad (2.19)$$

参数 μ_k 要满足 $\mu_k \geq 0, \sum_k \mu_k = 1$ 。易知

$$\mathbb{E}[\mathbf{x}|\boldsymbol{\mu}] = \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu}) \mathbf{x} = (\mu_1, \mu_2, \dots, \mu_K)^T = \boldsymbol{\mu} \quad (2.20)$$

现在考虑一个有 N 个独立观测值 x_1, x_2, \dots, x_N 的数据集 D 。对应的似然函数的形式为

$$p(D|\boldsymbol{\mu}) = \prod_{n=1}^N \prod_{k=1}^K \mu_k^{x_{nk}} = \prod_{k=1}^K \mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^K \mu_k^{m_k} \quad (2.21)$$

其中

$$m_k = \sum_n x_{nk} \quad (2.22)$$

表示观测到 $x_k = 1$ 的次数。这被称为这个分布的充分统计量。为了找到 $\boldsymbol{\mu}$ 的最大似然解,需要关于 μ_k 最大化 $\ln p(D|\boldsymbol{\mu})$,通过拉格朗日乘数 λ 实现。

$$\mu_k^{ML} = \frac{m_k}{N} \quad (2.23)$$

我们可以考虑 m_1, m_2, \dots, m_K 在参数 $\boldsymbol{\mu}$ 和观测总数 N 条件下的联合分布。这个分布的形式为

$$Mult(m_1, m_2, \dots, m_K | \boldsymbol{\mu}, N) = \binom{N}{m_1 m_2 \dots m_K} \prod_{k=1}^K \mu_k^{m_k} \quad (2.24)$$

这被称为多项式分布 (multinomial distribution)。归一化系数是把 N 个物体分成大小为 m_1, m_2, \dots, m_K 的 K 组的方案总数,定义为

$$\binom{N}{m_1 m_2 \dots m_K} = \frac{N!}{m_1! m_2! \dots m_K!} \quad (2.25)$$

注意, m_k 满足下面的限制

$$\sum_{k=1}^K m_k = N \quad (2.26)$$

狄利克雷分布

现在介绍多项式分布的参数 $\{\mu_k\}$ 的一组先验分布。通过观察多项式分布的形式,我们看到,共轭先验为

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^K \mu_k^{\alpha_k - 1} \quad (2.27)$$

其中 $0 \leq \mu_k \leq 1$ 且 $\sum_k \mu_k = 1$ 。这里, $\alpha_1, \alpha_2, \dots, \alpha_K$ 是分布的参数, α 表示 $(\alpha_1, \alpha_2, \dots, \alpha_K)^T$ 。概率的归一化形式为

$$Dir(\mu|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_K)} \prod_{k=1}^K \mu_k^{\alpha_k-1} \quad (2.28)$$

这被称为狄利克雷分布 (Dirichlet distribution)。

$$\alpha_0 = \sum_{k=1}^K \alpha_k \quad (2.29)$$

用似然函数乘以先验, 我们得到了参数 $\{\mu_k\}$ 的后验分布, 形式为

$$p(\mu|D, \alpha) \propto p(D|\mu)p(\mu|\alpha) \propto \prod_{k=1}^K \mu_k^{\alpha_k+m_k-1} \quad (2.30)$$

我们看到后验分布的形式又变成了狄利克雷分布。确定归一化系数后变成

$$p(\mu|D, \alpha) = Dir(\mu|\alpha + m) \quad (2.31)$$

2.3 高斯分布

高斯分布, 也被称为正态分布, 广泛应用于连续型随机变量分布的模型中。对于一元变量 x 的情形,

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \quad (2.32)$$

其中 μ 是均值, σ^2 是方差。

$$\mu_{ML} = \frac{1}{N} \sum_{i=1}^N x_i \quad \text{无偏} \quad (2.33)$$

$$\sigma_{ML}^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu_{ML})^2 \quad \text{有偏} \quad (2.34)$$

$$\mathbb{E}[\sigma_{ML}^2] = \frac{N-1}{N} \sigma^2 \quad (2.35)$$

$$\hat{\sigma}^2 = \frac{1}{N-1} \sum_{i=1}^N (x_i - \mu_{ML})^2 \quad \text{无偏} \quad (2.36)$$

对于 D 维向量 \mathbf{x} , 多元高斯分布的形式为

$$\mathcal{N}(\mathbf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)\right\} \quad (2.37)$$

其中, μ 是一个 D 维均值向量, Σ 是一个 $D \times D$ 的协方差矩阵, $|\Sigma|$ 是 Σ 的行列式。对 Σ 进

行正交分解。

$$\Sigma = U\Lambda U^T \quad (2.38)$$

$$UU^T = U^T U = I \quad (2.39)$$

$$\Lambda = \text{diag}(\lambda_i) \quad (2.40)$$

$$U = (u_1, u_2, \dots, u_p)_{p \times p} \quad (2.41)$$

考虑高斯分布的几何形式。高斯对于 \mathbf{x} 的依赖是通过下面形式的二次型

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \quad (2.42)$$

Δ 被叫做 $\boldsymbol{\mu}$ 和 \mathbf{x} 之间的马氏距离。协方差矩阵 Σ 可以表示成特征向量的展开的形式

$$\begin{aligned} \Sigma &= U\Lambda U^T \\ &= (u_1 \ u_2 \ \dots \ u_p) \begin{pmatrix} \lambda_1 & \dots & 0 \\ \vdots & \lambda_i & \vdots \\ 0 & \dots & \lambda_p \end{pmatrix} \begin{pmatrix} u_1^T \\ \vdots \\ u_p^T \end{pmatrix} \\ &= (u_1 \lambda_1 \ \dots \ u_p \lambda_p) \begin{pmatrix} u_1^T \\ \vdots \\ u_p^T \end{pmatrix} \\ &= \sum_{i=1}^P u_i \lambda_i u_i^T \end{aligned} \quad (2.43)$$

于是

$$\begin{aligned} \Sigma^{-1} &= (U\Lambda U^T)^{-1} = U\Lambda^{-1}U^T \\ &= \sum_{i=1}^P u_i \frac{1}{\lambda_i} u_i^T \end{aligned} \quad (2.44)$$

马氏距离就可以表示为

$$\begin{aligned} \Delta^2 &= (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \\ &= \sum_{i=1}^P \underbrace{(\mathbf{x} - \boldsymbol{\mu})^T u_i}_{y_i} \frac{1}{\lambda_i} \underbrace{u_i^T (\mathbf{x} - \boldsymbol{\mu})}_{y_i^T} = \sum_{i=1}^P \frac{y_i^2}{\lambda_i} \end{aligned} \quad (2.45)$$

其中

$$y_i = (\mathbf{x} - \boldsymbol{\mu})^T u_i \quad (2.46)$$

可以把 $\{y_i\}$ 表示成单位正交向量 u_i 关于原始的 x_i 坐标经过平移和旋转后形成的新的坐标系。



高斯分布的均值和方差为

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu} \quad (2.47)$$

$$\text{var}[\mathbf{x}] = \boldsymbol{\Sigma} \quad (2.48)$$

虽然高斯分布被广泛用作概率密度模型,但是它有着一些巨大的局限性.

1. 自由参数的数量。可以进一步地把协方差矩阵限制成正比于单位矩阵, $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$, 被称为各向同性 isotropic 的协方差。尽管这样的方法限制了概率分布的自由度的数量, 并且使得求协方差矩阵的逆矩阵可以更快地完成, 但是这样做也极大地限制了概率密度的形式, 限制了它描述模型中有趣的相关性的能力。
2. 单峰。因此不能够很好地近似多峰分布。

高斯分布一方面相当灵活, 因为它有很多参数。另一方面, 它又有很大的局限性, 因为它不能够近似很多分布。引入潜在变量 (latent variable) 也被称为隐藏变量 (hidden variable) 或者未观察变量 (unobserved variable), 会让这两个问题都得到解决。

条件高斯分布

多元高斯分布的一个重要性质是, 如果两组变量是联合高斯分布, 那么以一组变量为条件, 另一组变量同样是高斯分布, 类似地, 任何一个变量的边缘分布也是高斯分布。

首先考虑条件概率的情形, 假设 \mathbf{x} 是一个服从高斯分布 $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 的 D 维向量。我们把 \mathbf{x} 划分成两个不相交的子集 x_a, x_b 。不失一般性, 我们可令 x_a 为 \mathbf{x} 的前 M 个分量, 令

x_b 为剩余的 $D - M$ 个分量, 因此

$$\mathbf{x} = \begin{pmatrix} x_a \\ x_b \end{pmatrix} \quad (2.49)$$

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} \quad (2.50)$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \quad (2.51)$$

$$(2.52)$$

注意, 协方差矩阵的对称性 $\boldsymbol{\Sigma}^T = \boldsymbol{\Sigma}$ 表明 Σ_{aa} 和 Σ_{bb} 也是对称的, 而 $\Sigma_{ab}^T = \Sigma_{ba}$ 。

定理 2.1

已知 $x \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $Y = AX + B$ 有如下结论:

1. $\mathbb{E}[Y] = A\boldsymbol{\mu} + B$
2. $\text{var}[Y] = A \cdot \boldsymbol{\Sigma} \cdot A^T$



x_a 可以表示为

$$x_a = \underbrace{(I_m \ o)}_A \underbrace{\begin{pmatrix} x_a \\ x_b \end{pmatrix}}_x \quad (2.53)$$

$$\mathbb{E}[x_a] = (I_m \ o)(\mu_a \ \mu_b)^T = \mu_a \quad (2.54)$$

$$\text{var}[x_a] = (I_m \ o) \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \begin{pmatrix} I_m \\ o \end{pmatrix} = \Sigma_{aa} \quad (2.55)$$

$$x_a \sim \mathcal{N}(\mu_a, \Sigma_{aa}) \quad (2.56)$$

同理

$$\mathbb{E}[x_b] = \mu_b \quad (2.57)$$

$$\text{var}[x_b] = \Sigma_{bb} \quad (2.58)$$

$$x_b \sim \mathcal{N}(\mu_b, \Sigma_{bb}) \quad (2.59)$$

构造

$$x_{b|a} = x_b - \Sigma_{ba} \Sigma_{aa}^{-1} x_a \quad (2.60)$$

$$\mu_{b|a} = \mu_b - \Sigma_{ba} \Sigma_{aa}^{-1} \mu_a \quad (2.61)$$

$$\Sigma_{bb|a} = \Sigma_{bb} - \Sigma_{ba} \Sigma_{aa}^{-1} \Sigma_{ab} \quad (2.62)$$

$$x_{b|a} = \underbrace{(-\Sigma_{ba}\Sigma_{aa}^{-1}I_n)}_A \underbrace{\begin{pmatrix} x_a \\ x_b \end{pmatrix}}_x \quad (2.63)$$

$x_{b|a}$ 的分布为

$$\mathbb{E}[x_{b|a}] = (-\Sigma_{ba}\Sigma_{aa}^{-1}I_n) \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} = \mu_{b|a} \quad (2.64)$$

$$\text{var}[x_{b|a}] = (-\Sigma_{ba}\Sigma_{aa}^{-1}I_n) \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \begin{pmatrix} -\Sigma_{ba}\Sigma_{aa}^{-1}I_n \\ I \end{pmatrix} = \Sigma_{bb|a} \quad (2.65)$$

$$p(x_{b|a}) \sim \mathcal{N}(\mu_{b|a}, \Sigma_{bb|a}) \quad (2.66)$$

因为

$$x_b = x_{b|a} + \Sigma_{ba}\Sigma_{aa}^{-1}x_a \quad (2.67)$$

所以

$$\mathbb{E}[x_b|x_a] = \mu_{b|a} + \Sigma_{ba}\Sigma_{aa}^{-1}x_a \quad (2.68)$$

$$\text{var}[x_b|x_a] = \text{var}[x_{b|a}] = \Sigma_{bb|a} \quad (2.69)$$

$$p(x_b|x_a) \sim \mathcal{N}(\mu_{b|a}, \Sigma_{bb|a}) \quad (2.70)$$

高斯变量的贝叶斯定理

令边缘概率分布和条件概率分布的形式如下

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\mu, \Lambda^{-1}) \quad (2.71)$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{Ax} + \mathbf{b}, \mathbf{L}^{-1}) \quad (2.72)$$

其中, μ , A 和 b 是控制均值的参数, Λ 和 L 是精度矩阵。如果 x 的维度为 M , y 的维度为 D , 那么矩阵 A 的大小为 $D \times M$ 。

1. 求 $p(y)$ 。

$$y = Ax + b + \epsilon, \quad \epsilon \sim \mathcal{N}(0, L^{-1}) \quad (2.73)$$

$$\mathbb{E}[y] = \mathbb{E}(Ax + b) + \mathbb{E}(\epsilon) = A\mu + b \quad (2.74)$$

$$\text{var}[y] = \text{var}[Ax + b + \epsilon] = A\Lambda^{-1}A^T + L^{-1} \quad (2.75)$$

$$y \sim \mathcal{N}(A\mu + b, A\Lambda^{-1}A^T + L^{-1}) \quad (2.76)$$

2. 求 $p(x|y)$ 。

定义

$$z = \begin{pmatrix} x \\ y \end{pmatrix} \quad (2.77)$$

$$z \sim \mathcal{N} \left(\begin{bmatrix} \mu \\ A\mu + b \end{bmatrix}, \begin{bmatrix} \Lambda^{-1} & \Delta \\ \Delta^T & A\Lambda^{-1}A^T + L^{-1} \end{bmatrix} \right) \quad (2.78)$$

其中

$$\begin{aligned} \Delta &= \text{Cov}(x, y) \\ &= E[(x - E(x)) \cdot (y - E(y))^T] \\ &= E[(x - \mu)(y - A\mu - b)^T] \\ &= E[(x - \mu)(Ax + b + \epsilon - A\mu - b)^T] \\ &= E[(x - \mu)(Ax - A\mu)^T + (x - \mu)\epsilon^T] \\ &= E[(x - \mu)(Ax - A\mu)^T] + E[(x - \mu)\epsilon^T] \\ &= E[(x - \mu)(Ax - A\mu)^T] \\ &= E[(x - \mu)(x - \mu)^T] \cdot A^T \\ &= \text{var}[x] \cdot A^T \\ &= \Lambda^{-1}A^T \end{aligned} \quad (2.79)$$

所以

$$\mathbb{E}[x|y] = (\Lambda + A^T L A)^{-1} \{A^T L(y - b) + \Lambda \mu\} \quad (2.80)$$

$$\text{cov}[x|y] = (\Lambda + A^T L A)^{-1} \quad (2.81)$$

高斯分布的最大似然估计

给定一个数据集 $X = (x_1, x_2, \dots, x_N)^T$, 其中观测 $\{x_n\}$ 假定是独立地从多元高斯分布中抽取的。我们可以使用最大似然法估计分布的参数。

$$\mu_{ML} = \frac{1}{N} \sum_{n=1}^N x_n \quad (2.82)$$

$$\Sigma_{ML} = \frac{1}{N} \sum_{n=1}^N (x_n - \mu_{ML})(x_n - \mu_{ML})^T \quad (2.83)$$

$$\mathbb{E}[\mu_{ML}] = \mu \quad (2.84)$$

$$\mathbb{E}[\Sigma_{ML}] = \frac{N-1}{N} \Sigma \quad (2.85)$$

顺序估计

顺序的方法允许每次处理一个数据点,然后丢弃这个点。这对于在线应用很重要。并且当数据集相当大以至于一次处理所有数据点不可行的情况下,顺序方法也很重要。对于高斯分布

$$\begin{aligned}\mu_{ML}^{(N)} &= \frac{1}{N} \sum_{n=1}^N x_n \\ &= \frac{1}{N} x_N + \frac{1}{N} \sum_{n=1}^{N-1} x_n \\ &= \mu_{ML}^{(N-1)} + \frac{1}{N} (x_N - \mu_{ML}^{(N-1)})\end{aligned}\tag{2.86}$$

随着 N 的增加,后续数据点的贡献也会逐渐变小。我们不能总是能够使用这种方法推导出一个顺序的算法。因此我们要寻找一个更加通用的顺序学习的方法,这就引出了 **Robbins-Monro** 算法。考虑一对随机变量 θ 和 z , 它们由一个联合概率分布 $p(z, \theta)$ 所控制。已知 θ 的条件下, z 的条件期望定义了一个确定的函数 $f(\theta)$, 形式如下

$$f(\theta) \equiv \mathbb{E}[z|\theta] = \int z p(z|\theta) dz \tag{2.87}$$

通过这种方式定义的函数被称为回归函数 (regression function)。我们的目标是寻找根 θ^* 使得 $f(\theta^*) = 0$ 。如果我们有观测 z 和 θ 的一个大数据集,那么我们可以直接对回归函数建模,得到根的一个估计。但是假设我们每次观测到一个 z 的值,我们想找到一个对应的顺序估计方法来找到 θ^* 。下面的解决这种问题的通用步骤由 **Robbins and Monro** 给出。我们假设 z 的条件方差是有穷的,因此

$$\mathbb{E}[(z - f)^2|\theta] < \infty \tag{2.88}$$

不失一般性,假设当 $\theta > \theta^*$ 时 $f(\theta) > 0$, 当 $\theta < \theta^*$ 时 $f(\theta) < 0$ 。下式定义了一个根 θ^* 的顺序估计的序列

$$\theta^{(N)} = \theta^{(N-1)} - \alpha_{N-1} z(\theta^{(N-1)}) \tag{2.89}$$

$z(\theta^{(N)})$ 是当 θ 的取值为 $\theta^{(N)}$ 时 z 的观测值。系数 $\{\alpha_N\}$ 表示一个满足下列条件的正数序列

$$\lim_{N \rightarrow \infty} \alpha_N = 0 \tag{2.90}$$

$$\sum_{N=1}^{\infty} \alpha_N = \infty \tag{2.91}$$

$$\sum_{N=1}^{\infty} \alpha_N^2 < \infty \tag{2.92}$$

可以证明由公式给出的顺序估计确实以概率 1 收敛于根。第一个条件确保了后续的修正幅度会逐渐变小,从而这个过程可以收敛于一个极限值。第二个条件用来确保算法不会收

敛不到根的值。第三个条件保证了累计的噪声具有一个有限的方差,因此不会导致收敛失败。

高斯分布的贝叶斯推断

现在我们通过引入高斯分布中的参数的先验分布,介绍一种贝叶斯的方法。

1. σ^2 已知,估计 μ 。
2. μ 已知,估计 σ^2 。
3. μ 和 σ^2 都未知。

学生 t 分布

周期变量

混合高斯模型

虽然高斯分布有一些重要的分析性质,但是当它遇到实际数据集时,也会有巨大的局限性。通过将更基本的概率分布进行线性组合的这样的叠加方法,可以被形式化为概率模型,被称为混合模型 (mixture distributions)。通过使用足够多的高斯分布,并且调节它们的均值和方差以及线性组合的系数,几乎所有的连续概率密度都能够以任意的精度近似。考虑 K 个高斯概率密度的叠加,形式为

$$p(x) = \sum_{k=1}^K \pi_k \mathcal{N}(x|\mu_k, \Sigma_k) \quad (2.93)$$

这被称为混合高斯 (mixture of Gaussians)。其中

$$\sum_{k=1}^K \pi_k = 1, 0 \leq \pi_k \leq 1 \quad (2.94)$$

2.4 指数族分布

指数族分布的成员有许多共同的重要性质,并且以某种程度的一般性下讨论这些性质是很有启发性的。参数为 η 的变量 x 的指数族分布定义为具有下面形式的概率分布的集合

$$p(x|\eta) = h(x)g(\eta)\exp\{\eta^T u(x)\} \quad (2.95)$$

其中 x 可能是标量或者向量,可能是离散的或者是连续的。这里 η 被称为概率分布的自然参数 (natural parameters), $u(x)$ 是 x 的某个函数。函数 $g(\eta)$ 可以被看成系数,它确保了概率分布是归一化的,因此满足

$$\int p(x|\eta)dx = 1 \quad (2.96)$$

下面看一些例子

1. 伯努利分布

$$\begin{aligned}
p(x|\mu) &= \text{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x} \\
&= \exp\{x \ln \mu + (1-x) \ln(1-\mu)\} \\
&= (1-\mu) \exp\left\{\ln\left(\frac{\mu}{1-\mu}\right)x\right\}
\end{aligned} \tag{2.97}$$

与公式 2.95 比较,可以看出

$$\eta = \ln\left(\frac{\mu}{1-\mu}\right) \tag{2.98}$$

从中可以解出 η , 得到 $\mu = \sigma(\eta)$, 其中

$$\sigma(\eta) = \frac{1}{1 + \exp(-\eta)} \tag{2.99}$$

被称为 logistic sigmoid 函数。因此可以使用公式 2.95 给出的标准形式把伯努利分布写成下面的形式

$$p(x|\mu) = \sigma(-\eta) \exp(\eta x) \tag{2.100}$$

2. 单一观测 x 的多项式分布

$$\begin{aligned}
p(x|\mu) &= \prod_{k=1}^M \mu_k^{x_k} = \exp\left\{\sum_{k=1}^M x_k \ln \mu_k\right\} \\
\mu(x) &= x \\
h(x) &= 1 \\
g(\eta) &= 1
\end{aligned} \tag{2.101}$$

注意参数 η_k 不是相互独立的, 因为参数 μ_k 要满足下面的限制

$$\sum_{k=1}^M \mu_k = 1 \tag{2.102}$$

因此给定任意 $M-1$ 个参数 μ_k , 剩下的参数就固定了。使用这个限制, 这种表达方式下多项式分布变成了

$$\begin{aligned}
&\exp\left\{\sum_{k=1}^M x_k \ln \mu_k\right\} \\
&= \exp\left\{\sum_{k=1}^{M-1} x_k \ln \mu_k + \left(1 - \sum_{k=1}^{M-1} x_k\right) \ln\left(1 - \sum_{k=1}^{M-1} \mu_k\right)\right\} \\
&= \exp\left\{\sum_{k=1}^{M-1} x_k \ln\left(\frac{\mu_k}{1 - \sum_{j=1}^{M-1} \mu_j}\right) + \ln\left(1 - \sum_{k=1}^{M-1} \mu_k\right)\right\}
\end{aligned} \tag{2.103}$$

令

$$\ln \left(\frac{\mu_k}{1 - \sum_j \mu_j} \right) = \eta_k \quad (2.104)$$

从中我们可以解出 μ_k 。首先两侧对 k 求和, 然后整理, 回带, 可得

$$\mu_k = \frac{\exp(\eta_k)}{1 + \sum_j \exp(\eta_j)} \quad (2.105)$$

这被称为 softmax 函数, 或者归一化指数 (normalized exponential)。在这个表达方式的形式下, 多项式分布的形式为

$$p(x|\eta) = \left(1 + \sum_{k=1}^{M-1} \exp(\mu_k) \right)^{-1} \exp(\mu^T x) \quad (2.106)$$

这是指数族分布的标准形式, 其中参数向量 $\eta = (\eta_1, \dots, \eta_{M-1}, 0)^T$ 。在这个指数族分布中

$$\mu(x) = x \quad (2.107)$$

$$h(x) = 1 \quad (2.108)$$

$$g(\eta) = \left(1 + \sum_{k=1}^{M-1} \exp(\eta_k) \right)^{-1} \quad (2.109)$$

3. 高斯分布

$$\begin{aligned} p(x|\mu, \sigma^2) &= \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\} \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} \exp \left\{ -\frac{1}{2\sigma^2} x^2 + \frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} \mu^2 \right\} \end{aligned} \quad (2.110)$$

经过简单的推导后, 它可以转化为公式 2.95 给出的标准指数族分布的形式, 其中

$$\eta = \begin{pmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{pmatrix} \quad (2.111)$$

$$\mu(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \quad (2.112)$$

$$h(x) = (2\pi)^{\frac{1}{2}} \quad (2.113)$$

$$g(\eta) = (-2\eta_2)^{\frac{1}{2}} \exp \left(\frac{\eta_1^2}{4\eta_2^2} \right) \quad (2.114)$$

最大似然与充分统计量

用最大似然法估计公式 2.95 给出的一般形式的指数族分布的参数向量 μ 的问题。

$$\int h(x)g(\eta)\exp\{\eta^T u(x)\} = 1 \quad (2.115)$$

对上式的两侧关于 μ 取梯度,我们有

$$\begin{aligned} \nabla g(\eta) \underbrace{\int h(x)\exp\{\eta^T u(x)\} dx}_{1/g(\eta)} + g(\eta) \int h(x)\exp\{\eta^T u(x)\} u(x) dx &= 0 \\ \Rightarrow -\frac{1}{g(\eta)} \nabla g(\eta) &= \int \underbrace{g(\eta)h(x)\exp\{\eta^T u(x)\}}_{p(x|\eta)} u(x) dx = \mathbb{E}[u(x)] \\ \Rightarrow -\nabla \ln g(\eta) &= \mathbb{E}[u(x)] \end{aligned} \quad (2.116)$$

同理, $u(x)$ 的协方差,可以根据 $g(\eta)$ 的二阶导数表达,对于高阶矩的情形也类似。因此,如果我们能够对一个来自指数族分布的概率分布进行归一化,那么我们总能够通过简单的求微分的方式找到它的矩。现在考虑一组独立同分布的数据 $X = \{x_1, \dots, x_N\}$ 。对于这个数据集,似然函数为

$$p(X|\eta) = \left(\prod_{n=1}^N h(x_n) \right) g(\eta)^N \exp \left\{ \eta^T \sum_{n=1}^N u(x_n) \right\} \quad (2.117)$$

令 $\ln p(X|\eta)$ 关于 η 的导数等于零,我们可以得到最大似然估计 μ_{ML} 满足的条件

$$-\nabla \ln g(\eta) = \frac{1}{N} \sum_{n=1}^N u(x_n) \quad (2.118)$$

原则上可以通过这个方程来得到 μ_{ML} 。我们看到最大似然估计的解只通过 $\sum_n u(x_n)$ 对数据产生依赖,因此这个量被称为分布的充分统计量 (sufficient statistic)。我们不需要存储整个数据集本身,只需要存储充分统计量的值即可。

共轭先验

对于指数族分布的任何成员,都存在一个共轭先验,可以写成下面的形式

$$p(\eta|\varkappa, \nu) = f(\varkappa, \nu) g(\eta)^\nu \exp\{\nu \eta^T \varkappa\} \quad (2.119)$$

无信息先验

在许多情形下,我们可能对分布应该具有的形式几乎完全不知道。这时,我们可以寻找一种形式的先验分布,被称为无信息先验 (noninformative prior)。这种先验分布的目的是尽量对后验分布产生尽可能小的影响。有时被称为“让数据自己说话”。

1. 位置参数
2. 缩放参数

2.5 非参数化方法

使用一些非参数化方法进行概率密度估计。这种方法对概率分布的形式进行了很少的假设。

直方图法

标准的直方图简单地把 x 划分成不同的宽度为 Δ_i 的箱子, 然后对落在第 i 个箱子中的 x 的观测数量 n_i 进行计数。为了把这种计数转换成归一化的概率密度, 我们简单地把观测数量除以观测的总数 N , 再除以箱子的宽度 Δ_i , 得到每个箱子的概率的值

$$p_i = \frac{n_i}{N\Delta_i} \quad (2.120)$$

在实际应用中, 直方图方法对于快速地将一维或者二维的数据可视化很有用, 但是并不适用于大多数概率密度估计的应用。一个明显的问题是估计的概率密度具有不连续性, 这种不连续性是因为箱子的边缘造成的, 而不是因为生成数据的概率分布本身的性质造成。直方图的另一个主要的局限性是维数放大。但是, 概率密度估计的直方图方法确实告诉了我们两个重要的事情。

1. 为了估计在某个特定位置的概率密度, 我们应该考虑位于那个点的某个领域内的数据点。
2. 为了获得好的结果, 平滑参数的值既不能太大也不能太小。

假设观测服从 D 维空间的某个未知的概率密度分布 $p(x)$ 。把这个 D 维空间选择成欧几里德空间, 并且我们想估计 $p(x)$ 的值。根据以前对于局部性的讨论, 让我们考虑包含 x 的某个小区域 \mathcal{R} 。这个区域的概率质量为

$$P = \int_{\mathcal{R}} p(x) dx \quad (2.121)$$

现在我们假设收集了服从 $p(x)$ 分布的 N 次观测。由于每个数据点都有一个落在区域 \mathcal{R} 中的概率 P , 因此位于区域 \mathcal{R} 内部的数据点的总数 K 将服从二项分布

$$\text{Bin}(K|N, P) = \frac{N!}{K!(N-K)!} P^K (1-P)^{N-K} \quad (2.122)$$

落在区域内部的数据点的平均比例为 $\mathbb{E}[\frac{K}{N}] = P$ 。类似地, 以此为均值的概率分布的方差为 $\text{var}[\frac{K}{N}] = \frac{P(1-P)}{N}$ 。对于大的 N 值, 这个分布将会在均值附近产生尖峰, 并且

$$K \simeq NP \quad (2.123)$$

$$P \simeq p(x)V \quad (2.124)$$

$$p(x) = \frac{K}{NV} \quad (2.125)$$

上式的成立依赖于两个相互矛盾的假设, 即区域 \mathcal{R} 要足够小, 使得这个区域内的概率密度近似为常数, 但是也要足够大, 使得落在这个区域内的数据点的数量 K 能够足够让二项分布达到尖峰。

我们有两种方式利用这个结果。

1. 固定 K 然后从数据中确定 V 的值, 这就是 K 近邻方法。
2. 固定 V 然后从数据中确定 K 的值, 这就是核方法。

在极限 $N \rightarrow \infty$ 的情况下, 如果 V 随着 N 而合适地收缩, 并且 K 随着 N 增大, 那么可以证明 K 近邻概率密度估计和核方法概率密度估计都会收敛到真实的概率密度。

核密度估计

把区域 \mathcal{R} 取成以 x 为中心的小超立方体, 为了统计落在这个区域内的数据点的数量 K , 定义下面的函数比较方便

$$k(u) = \begin{cases} 1, & |u_i| \leq \frac{1}{2}, i = 1, \dots, D \\ 0, & \text{其它情况} \end{cases} \quad (2.126)$$

这表示一个以原点为中心的单位立方体。函数 $k(u)$ 是核函数的一个例子, 在这个问题中也被称为 Parzen 窗 (parzen window)。根据公式 2.126, 如果数据点 x_n 位于以 x 为中心的边长为 h 的立方体中, 那么量 $k(\frac{x-x_n}{h})$ 的值等于 1, 否则它的值为 0。于是

$$p(x) = \frac{1}{N} \sum_{n=1}^N \frac{1}{h^D} k\left(\frac{x-x_n}{h}\right) \quad (2.127)$$

这个函数表述为以 N 个数据点 x_n 为中心的 N 个立方体。

核密度估计有人为带来的非连续性的问题。如果我们选择一个平滑的核函数, 那么我们就可以得到一个更加光滑的模型。

$$p(x) = \frac{1}{N} \sum_{n=1}^N \frac{1}{(2\pi h^2)^{\frac{D}{2}}} \exp\left\{-\frac{\|x-x_n\|^2}{2h^2}\right\} \quad (2.128)$$

h 表示高斯分布的标准差。因此我们概率密度模型可以通过这种方式获得: 令每个数据点都服从高斯分布, 然后把数据集里的每个数据点的贡献相加, 之后除以 N , 使得概率密度正确地归一化。

这种估计方法有一个很大的优点,即不需要进行“训练”阶段的计算,因为“训练”阶段只需要存储训练集即可。然而,这也是一个巨大的缺点,因为估计概率密度的计算代价随着数据集的规模线性增长。

近邻方法

核方法进行概率密度估计的一个困难之处是控制核宽度的参数 h 对于所有的核都是固定的。与之不同,考虑固定 K 的值然后使用数据来确定合适的 V 值。为了完成这一点,我们考虑一个以 x 为中心的小球体,然后我们想估计概率密度 $p(x)$ 。并且,我们允许球体的半径可以自由增长,直到它精确地包含 K 个数据点。这样,概率密度 $p(x)$ 的估计就由公式 2.126 给出,其中 V 等于最终球体的体积。这种方法被称为 K 近邻方法。

第 3 章 变分法



第 4 章 矩阵的性质



第5章 回归的线性模型

5.1 线性基函数模型

回归问题的最简单模型是输入变量的线性组合

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \dots + w_D x_D$$

其中 $\mathbf{x} = (x_1, \dots, x_D)^T$ 。这通常被简单地称为**线性回归 (linear regression)**。这个模型的关键性质是它是参数 w_0, \dots, w_D 的一个线性函数。但是, 它也是输入变量 x_i 的一个线性函数, 这给模型带来了极大的局限性。因此, 我们扩展模型的类别: 将输入变量的固定的非线性函数进行线性组合, 形式为

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x})$$

其中 $\phi_j(\mathbf{x})$ 被称为基函数 (basis function)。

1. 高斯基函数

$$\phi_j(x) = \exp \left\{ -\frac{(x - \mu_j)^2}{2s^2} \right\}$$

2. sigmoid 基函数

$$\phi_j(x) = \frac{1}{1 + \exp \left(\frac{x - \mu_j}{s} \right)}$$

3. 傅里叶基函数

最大似然与最小平方

最小化平方和误差函数可以看成高斯噪声模型的假设下的最大似然解。假设目标变量 t 由确定的函数 $y(\mathbf{x}, \mathbf{w})$ 给出, 这个函数被附加了高斯噪声, 即

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon \quad (5.1)$$

其中 ϵ 是一个零均值的高斯随机变量, 精度为 β 。因此我们有

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1}) \quad (5.2)$$

最优的预测由目标变量的条件给出

$$\mathbb{E}[t|\mathbf{x}] = \int t p(t|\mathbf{x}) dt = y(\mathbf{x}, \mathbf{w}) \quad (5.3)$$

现在考虑一个输入数据集 $X = \{x_1, \dots, x_N\}$, 对应的目标值为 t_1, \dots, t_N 。把目标向量 $\{t_n\}$ 组成一个列向量, 记作 \mathbf{t} 。这个变量的字体与多元目标值的一次观测 (记作 t) 不同。假设这些数据点是独立同分布的。那么可以得到下面的似然函数

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^N \mathcal{N}(t_n | \mathbf{w}^T \phi(x_n), \beta^{-1}) \quad (5.4)$$

注意, 在有监督学习问题中 (例如回归问题和分类问题), 我们不是在寻找模型来对输入变量的概率分布建模。因此 x 总会出现在条件变量的位置上。为了保持记号的简洁性, 不显式地写出 x 。取似然函数的对数, 我们有

$$\begin{aligned} \ln p(\mathbf{t}|\mathbf{w}, \beta) &= \sum_{n=1}^N \ln \mathcal{N}(t_n | \mathbf{w}^T \phi(x_n), \beta^{-1}) \\ &= \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi) - \beta E_D(\mathbf{w}) \end{aligned} \quad (5.5)$$

其中平方和误差函数的定义为

$$\begin{aligned} E_D(\mathbf{w}) &= \frac{1}{2} \sum_{n=1}^N \{t_n - \mathbf{w}^T \phi(x_n)\}^2 \\ &= \frac{1}{2} (\mathbf{w}^T \phi(x_1) - t_1 \quad \dots \quad \mathbf{w}^T \phi(x_N) - t_N) \begin{pmatrix} \mathbf{w}^T \phi(x_1) - t_1 \\ \vdots \\ \mathbf{w}^T \phi(x_N) - t_N \end{pmatrix} \\ &= \frac{1}{2} (\mathbf{w}^T (\phi(x_1) \quad \dots \quad \phi(x_N)) - (t_1 \quad \dots \quad t_N)) \left(\begin{pmatrix} \mathbf{w}^T \phi(x_1) \\ \vdots \\ \mathbf{w}^T \phi(x_N) \end{pmatrix} - \begin{pmatrix} t_1 \\ \vdots \\ t_N \end{pmatrix} \right) \\ &= \frac{1}{2} (\mathbf{w}^T \Phi^T - \mathbf{t}^T) (\Phi \mathbf{w} - \mathbf{t}) \\ &= \frac{1}{2} (\mathbf{w}^T \Phi^T \Phi \mathbf{w} - 2 \mathbf{w}^T \Phi^T \mathbf{t} + \mathbf{t}^T \mathbf{t}) \end{aligned} \quad (5.6)$$

使用最大似然方法确定 \mathbf{w} 和 β 。首先关于 \mathbf{w} 求最大值。

$$\begin{aligned} \nabla \ln p(\mathbf{t}|\mathbf{w}, \beta) &= \beta \sum_{n=1}^N \{t_n - \mathbf{w}^T \phi(x_n)\} \phi(x_n)^T \\ &= \Phi^T \Phi \mathbf{w} - \Phi^T \mathbf{t} \end{aligned} \quad (5.7)$$

令梯度为零, 可得

$$\mathbf{w} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{t} \quad (5.8)$$

这被称为最小平方问题的规范方程 (normal equation)。这里 Φ 是一个 $N \times M$ 的矩阵, 被称为设计矩阵 (design matrix), 它的元素为 $\Phi_{nj} = \phi_j(x_n)$, 即

$$\begin{pmatrix} \Phi_0(x_1) & \Phi_0(x_1) & \dots & \Phi_{M-1}(x_1) \\ \Phi_0(x_2) & \Phi_0(x_2) & \dots & \Phi_{M-1}(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ \Phi_0(x_N) & \Phi_0(x_N) & \dots & \Phi_{M-1}(x_N) \end{pmatrix} \quad (5.9)$$

关于噪声精度参数 β 最大化似然函数, 结果为

$$\frac{1}{\beta_{ML}} = \frac{1}{N} \sum_{n=1}^N \{t_n - \mathbf{w}_{ML}^T \phi(x_n)\}^2 \quad (5.10)$$

因此我们看到噪声精度的倒数由目标值在回归函数周围的残留方差 (residual variance) 给出。

最小平方的几何描述

顺序学习

最大似然解的求解过程涉及到一次处理整个数据集。这种批处理技术对于大规模数据集来说计算量相当大。如果数据集充分大, 那么使用顺序算法 (也被称为在线算法) 可能更有价值。顺序算法中, 每次只考虑一个数据点, 模型的参数在每观测到一个数据点之后进行更新。顺序学习也适用于实时的应用。在实时应用中, 数据观测以一个连续的流的方式持续到达, 我们必须在观测到所有数据之前就做出预测。

我们可以获得一个顺序学习的算法通过考虑随机梯度下降 (stochastic gradient descent) 也被称为顺序梯度下降 (sequential gradient descent) 的方法。

正则化最小平方

为误差函数添加正则化项的思想来控制过拟合, 因此需要最小化的总的误差函数的形式为

$$E_D(\mathbf{w}) + \lambda E_W(\mathbf{w}) \quad (5.11)$$

L1 正则化引起稀疏解的多种解释。

1. 用图解释:

L2 正则相当于用圆去逼近目标, 而 L1 正则相当于用菱形去逼近目标, 所以更容易引起交点在坐标轴上即得到稀疏解。

2. 从导数角度解释:

L2 正则无法将目标函数的极值点拉拢到稀疏解上, 而 L1 正则因为 L1 导函数的特殊性从而可以在一定范围内将极值点直接拉拢到稀疏解上。

3. 从先验概率分布角度解释:

L2 正则相当于假设参数是服从高斯分布的, 而 L1 正则相当于假设了参数是服从拉普拉斯分布的, 自然拉普拉斯分布比高斯分布更集中在 0 这个点上。

多个输出

在某些应用中, 我们可能想预测 $K > 1$ 个目标变量。我们把这些目标变量聚焦起来, 记作目标向量 \mathbf{t} 。这个问题可以这样解决: 对于 \mathbf{t} 的每个分量, 引入一个不同的基函数集合, 从而变成了多个独立的回归问题。但是, 一个更有趣的并且更常用的方法是对目标向量的所有分量使用一组相同的基函数来建模, 即

$$\mathbf{y}(\mathbf{x}, \mathbf{w}) = \mathbf{W}^T \boldsymbol{\phi}(\mathbf{x}) \quad (5.12)$$

其中 \mathbf{y} 是一个 K 维列向量, \mathbf{W} 是一个 $M \times K$ 的参数矩阵, $\boldsymbol{\phi}(\mathbf{x})$ 是一个 M 维列向量, 每个元素为 $\phi_j(\mathbf{x})$ 。

5.2 偏置-方差分解

目前为止, 我们对于回归的线性模型的讨论中, 我们假定了基函数的形式和数量都是固定的。如果使用有限规模的数据集来训练复杂的模型, 那么使用最大似然法, 或者等价地使用最小平方方法, 会导致严重的过拟合问题。正如前面所说, 过拟合现象确实是最大似然方法的一个不好的性质。但是当我们在使用贝叶斯方法对参数进行求和或者积分时, 过拟合现象不会出现。从贝叶斯观点讨论模型的复杂度之前, 从频率学家的观点考虑一下模型的复杂度的问题——偏置-方差折中 (bias-variance trade-off)。

最优的预测 (变分法可求出) 由条件期望 (记作 $h(\mathbf{x})$) 给出, 即

$$h(\mathbf{x}) = \mathbb{E}[\mathbf{t}|\mathbf{x}] = \int \mathbf{t} p(\mathbf{t}|\mathbf{x}) d\mathbf{t} \quad (5.13)$$

平方损失函数的期望可以写成

$$\begin{aligned} \mathbb{E}[L] &= \iint \{y(\mathbf{x}) - h(\mathbf{x})\}^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t} \\ &= \int \{y(\mathbf{x}) - h(\mathbf{x})\}^2 p(\mathbf{x}) d\mathbf{x} + \iint \{h(\mathbf{x}) - \mathbf{t}\}^2 p(\mathbf{x}, \mathbf{t}) d\mathbf{x} d\mathbf{t} \end{aligned} \quad (5.14)$$

考虑第一项的被积函数, 对于一个特定的数据集 D , 它的形式为

$$\{y(\mathbf{x}; D) - h(\mathbf{x})\}^2 \quad (5.15)$$

由于这个量与特定的数据集 D 相关, 因此我们对所有的数据集取平均。如果我们在括号

内加上然后减去 $\mathbb{E}_D[y(x; D)]$, 然后展开, 我们有

$$\begin{aligned}
 & \{y(x; D) - \mathbb{E}_D[y(x; D)] + \mathbb{E}_D[y(x; D)] - h(x)\}^2 \\
 &= \{y(x; D) - \mathbb{E}_D[y(x; D)]\}^2 \\
 &+ \{\mathbb{E}_D[y(x; D)] - h(x)\}^2 \\
 &+ 2\{y(x; D) - \mathbb{E}_D[y(x; D)]\}\{\mathbb{E}_D[y(x; D)] - h(x)\}
 \end{aligned} \tag{5.16}$$

现在关于 D 求期望, 然后注意到**最后一项等于零**, 可得

$$\begin{aligned}
 & \mathbb{E}_D[\{y(x; D) - h(x)\}^2] \\
 &= \underbrace{\mathbb{E}_D[\{y(x; D) - \mathbb{E}_D[y(x; D)]\}^2]}_{\text{(偏置)}^2} + \underbrace{\mathbb{E}_D[\{\mathbb{E}_D[y(x; D)] - h(x)\}^2]}_{\text{方差}}
 \end{aligned} \tag{5.17}$$

将式 5.17 代入式 5.14 中, 就得到了对于期望平方损失的分解

$$\text{期望损失} = \text{偏置}^2 + \text{方差} + \text{噪声} \tag{5.18}$$

其中

$$\text{偏置}^2 = \int \{\mathbb{E}_D[y(x; D)] - h(x)\}^2 p(x) dx \tag{5.19}$$

$$\text{方差} = \int \mathbb{E}_D[\{y(x; D) - \mathbb{E}_D[y(x; D)]\}^2] p(x) dx \tag{5.20}$$

$$\text{噪声} = \iint \{h(x) - t\}^2 p(x, t) dx dt \tag{5.21}$$

我们的目标是最小化期望损失, 它可以分解为 (平方) 偏置、方差和一个常数噪声项的和。对于非常灵活的模型来说, 偏置较小, 方差较大。对于相对固定的模型来说, 偏置较大, 方差较小。有着最优预测能力的模型是在偏置和方差之前取得最优的平衡的模型。

5.3 贝叶斯线性回归

使用最大似然方法设置线性回归模型的参数时, 由基函数的数量控制的模型的复杂度需要根据数据集的规模进行调整。为对数似然函数增加一个正则化项意味着模型的复杂度可以通过正则化系数的值进行控制, 虽然基函数的数量和形式的选择仍然对于确定模型的整体行为十分重要。

这就产生了对于特定的应用确定合适的模型复杂度的问题。这个问题不能简单地通过最大化似然函数来确定, 因为这总会产生过于复杂的模型和过拟合现象。独立的额外数据能够用来确定模型的复杂度, 但这需要较大的计算量, 并且浪费了有价值的数据。因此我们转而考虑线性回归的贝叶斯方法, 这会避免最大似然的过拟合问题, 也会引出使用训练数据本身确定模型复杂度的自动化方法。

参数分布

关于线性拟合的贝叶斯方法的讨论, 我们首先引入模型参数 w 的先验概率分布。把噪声精度参数 β 当做已知常数。对应的共轭先验是高斯分布。

$$p(w) \equiv \mathcal{N}(w | m_0, S_0) \quad (5.22)$$

均值为 m_0 协方差为 S_0 后验分布

$$p(w | t) = \mathcal{N}(w | m_N, S_N) \quad (5.23)$$

其中,

$$m_N = S_N(S_0^{-1} + \beta \Phi^T t) \quad (5.24)$$

$$S_N^{-1} = S_0^{-1} + \beta \Phi^T \Phi \quad (5.25)$$

为了简化起见, 考虑高斯先验的一个特定的形式——零均值各向同性高斯分布。这个分布由一个精度参数 α 控制, 即

$$p(w | \alpha) = \mathcal{N}(w | 0, \alpha^{-1} I) \quad (5.26)$$

后验概率分布的对数由对数似然函数与先验的对数求和的方式得到。它是 w 的函数, 形式为

$$\ln p(w | t) = -\frac{\beta}{2} \sum_{n=1}^N \{t_n - w^T \phi(x_n)\}^2 - \frac{\alpha}{2} w^T w + \text{常数} \quad (5.27)$$

于是, 后验分布关于 w 的最大化等价于对平方和误差函数加上一个二次正则项进行最小化。

预测分布

在实际应用中, 我们通常感兴趣的不是 w 本身的值, 而是对于新的 x 值预测出 t 的值。这需要计算出预测分布 (predictive distribution), 定义为

$$p(t | t, \alpha, \beta) = \int p(t | w, \beta) p(w | t, \alpha, \beta) dw \quad (5.28)$$

其中 t 是训练数据的目标变量的值组成的向量。并且, 为了简化记号, 我们在右侧省略了条件概率中出现的输入向量。预测分布的形式为

$$p(t | x, t, \alpha, \beta) = \mathcal{N}(t | m_N^T \phi(x), \sigma_N^2(x)) \quad (5.29)$$

其中预测分布的方差 $\sigma_N^2(\mathbf{x})$ 为

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \phi(\mathbf{x})^T \mathbf{S}_N \phi(\mathbf{x}) \quad (5.30)$$

公式 5.30 的第一项表示数据中的噪声, 而第二项反映了与参数 \mathbf{w} 关联的不确定性。由于噪声和 \mathbf{w} 的分布是相互独立的高斯分布, 因此它们的值是可以相加的。注意, 当额外的数据点被观测到的时候, 后验概率分布会变窄。从而可以证明出 $\sigma_{N+1}^2(\mathbf{x}) \leq \sigma_N^2(\mathbf{x})$ 在极限 $N \rightarrow \infty$ 的情况下, 公式 5.30 的第二项趋于零, 从而预测分布的方差只与参数 β 控制的具有可加性的噪声有关。

等价核

把公式 5.24 代入线性基函数模型中, 可以写成下面的形式

$$\begin{aligned} y(\mathbf{x}, \mathbf{m}_N) &= \mathbf{m}_N^T \phi(\mathbf{x}) \\ &= \beta \phi(\mathbf{x})^T \mathbf{S}_N \Phi^T \mathbf{t} \\ &= \sum_{n=1}^N \frac{\beta \phi(\mathbf{x})^T \mathbf{S}_N \phi(\mathbf{x}_n) t_n}{1} \\ &= \sum_{n=1}^N k(\mathbf{x}, \mathbf{x}_n) t_n \end{aligned} \quad (5.31)$$

可以看成在点 \mathbf{x} 处的预测均值由训练集目标变量 t_n 的线性组合给出。函数 $k(\mathbf{x}, \mathbf{x}_n)$ 被称为平滑矩阵 (smoother matrix) 或者等价核 (equivalent kernel)。像这样的回归函数, 通过对训练集里目标值进行线性组合做预测, 被称为线性平滑 (linear smoother)。核函数 $k(\mathbf{x}, \mathbf{x}')$ 给出了 \mathbf{x} 与 \mathbf{x}' 的函数关系。可以看到, \mathbf{x} 处的预测分布的均值 $y(\mathbf{x}, \mathbf{m}_N)$ 可以通过对目标值加权组合的方式获得。距离 \mathbf{x} 较近的数据点可以赋一个较高的权值, 而距离 \mathbf{x} 较远的数据点可以赋一个较低的权值。这种局部性不仅对于局部的高斯基函数成立, 对于非局部的多项式基函数和 sigmoid 基函数也成立。

考虑 $y(\mathbf{x})$ 和 $y(\mathbf{x}')$ 的协方差

$$\begin{aligned} \text{cov}[y(\mathbf{x}), y(\mathbf{x}')] &= \text{cov}[\phi(\mathbf{x})^T \mathbf{w}, \mathbf{w}^T \phi(\mathbf{x}')] \\ &= \phi(\mathbf{x})^T \mathbf{S}_N \phi(\mathbf{x}') \\ &= \beta^{-1} k(\mathbf{x}, \mathbf{x}') \end{aligned} \quad (5.32)$$

根据等价核的形式, 我们可以看到在附近的点处的预测均值相关性较高, 而对于距离较远的点处, 相关性就较低。

用核函数表示线性回归给出了解决回归问题的另一种方法。我们不引入一组基函数 (它隐式地定义了一个等价的核), 而是直接定义一个局部的核函数, 然后在给定观测数据集的条件下, 使用这个核函数对新的输入变量 \mathbf{x} 做预测。这就引入了用于回归问题 (以及分类问题) 的一个很实用的框架, 被称为高斯过程 (Gaussian process)。

5.4 贝叶斯模型比较

前面介绍了使用交叉验证的方法,来设置正则化参数的值,或者从多个模型中选择一个合适的。这里,从贝叶斯的角度考虑模型选择的问题。模型比较的贝叶斯观点仅仅涉及到使用概率来表示模型选择的不确定性,以及恰当地使用概率的加和规则和乘积规则。假设我们想比较 L 个模型 $\{M_i\}, i = 1, 2, \dots, L$ 。这里,一个模型指的是观测数据 D 上的概率分布。我们会假设数据是由这些模型中的一个生成的,但是我们不知道究竟是哪一个。我们的不确定性通过先验概率分布 $p(M_i)$ 表示。给定一个数据集 D , 我们想估计后验分布

$$p(M_i|D) \propto p(M_i)p(D|M_i) \quad (5.33)$$

先验分布让我们能够表达不同模型之间的优先级。**我们简单的假设所有的模型都有相同的先验概率。**模型证据 (model evidence) $p(D|M_i)$, 表达了数据展现出的不同模型的优先级。模型证据有时也被称为边缘似然 (marginal likelihood), 因为它可以被看做在模型空间中的似然函数, 在这个空间中参数已经被求和或者积分。两个模型的模型证据的比值 $\frac{p(D|M_i)}{p(D|M_j)}$ 被称为贝叶斯因子 (Bayes factor)。

一旦我们知道了模型上的后验概率分布, 那么根据概率的加和规则与乘积规则, 预测分布为

$$p(t|\mathbf{x}, D) = \sum_{i=1}^L p(t|\mathbf{x}, M_i, D)p(M_i|D) \quad (5.34)$$

这是混合分布 (mixture distribution) 的一个例子。这个公式中, 整体的预测分布由下面的方式获得: 对各个模型的预测分布 $p(t|\mathbf{x}, M_i, D)$ 求加权平均, 权值为这些模型的后验概率 $p(M_i|D)$ 。

对模型求平均的一个简单的近似是使用最可能的一个模型做预测。这被称为模型选择 (model selection)。

对于一个由参数 \mathbf{w} 控制的模型, 根据概率的加和规则和乘积规则, 模型证据为

$$p(D|M_i) = \int p(D|\mathbf{w}, M_i)p(\mathbf{w}|M_i)d\mathbf{w} \quad (5.35)$$

从取样的角度来看, 边缘似然函数可以被看成从一个模型中生成数据集 D 的概率, 这个模型的参数是从先验分布中随机取样的。同时, 注意到模型证据恰好就是在估计参数的后验分布时出现在贝叶斯定理的分母中的归一化项, 因为

$$p(\mathbf{w}|D, M_i) = \frac{p(D|\mathbf{w}, M_i)p(\mathbf{w}|M_i)}{p(D|M_i)} \quad (5.36)$$

对参数的积分进行一个简单的近似。假设后验分布在最大似然值 \mathbf{w}_{MAP} 附近是一个尖峰, 宽度为 $\Delta\mathbf{w}_{\text{后验}}$, 那么可以用被积函数的值乘以尖峰的宽度来近似这个积分。进一步假设

先验分布是平的, 宽度为 $\Delta w_{\text{先验}}$, 即 $p(w) = \frac{1}{\Delta w_{\text{先验}}}$, 那么我们有

$$p(D) = \int p(D|w)p(w)dw \simeq p(D|w_{MAP}) \frac{\Delta w_{\text{后验}}}{\Delta w_{\text{先验}}} \quad (5.37)$$

取对数可得

$$\ln p(D) \simeq \ln p(D|w_{MAP}) + \ln \left(\frac{\Delta w_{\text{后验}}}{\Delta w_{\text{先验}}} \right) \quad (5.38)$$

第一项表示拟合由最可能参数给出的数据。对于平的先验分布来说, 这对应于对数似然。第二项用于根据模型的复杂度来惩罚模型。

对于一个有 M 个参数的模型, 我们可以对每个参数进行类似的近似。假设所有的参数的 $\frac{\Delta w_{\text{后验}}}{\Delta w_{\text{先验}}}$ 都相同, 我们有

$$\ln p(D) \simeq \ln p(D|w_{MAP}) + M \ln \left(\frac{\Delta w_{\text{后验}}}{\Delta w_{\text{先验}}} \right) \quad (5.39)$$

因此, 在这种非常简单的近似下, 复杂度惩罚项的大小随着模型中可调节参数 M 的数量线性增加。随着我们增加模型的复杂度, 第一项通常会增大, 因为一个更加复杂的模型能够更好地拟合数据, 而第二项会减小, 因为它依赖于 M 。由最大模型证据确定的最优的模型复杂度需要在这两个相互竞争的项之间进行折中。

5.5 证据近似

在处理线性基函数模型的纯粹的贝叶斯方法中, 我们会引入超参数 α 和 β 的先验分布, 然后通过对超参数以及参数 w 求积分的方式做预测。但是, 虽然我们可以解析地求出对 w 的积分或者求出对超参数的积分, 但是对所有这些变量完整地求积分是没有解析解的。这里讨论一种近似方法。这种方法中, 首先对参数 w 求积分, 得到边缘似然函数 (marginal likelihood function), 然后通过最大化边缘似然函数, 确定超参数的值。这个框架在统计学的文献中被称为经验贝叶斯, 或者被称为第二类最大似然, 或者被称为推广的最大似然。在机器学习的文献中, 这种方法也被称为证据近似 (evidence approximation)。

如果引入 α 和 β 上的超先验分布, 那么预测分布可以通过对 w, α, β 求积分的方法得到

$$p(t|\mathbf{t}) = \iiint p(t|w, \beta) p(w|\mathbf{t}, \alpha, \beta) p(\alpha, \beta|\mathbf{t}) dw d\alpha d\beta \quad (5.40)$$

如果后验分布 $p(\alpha, \beta|\mathbf{t})$ 在 $\hat{\alpha}$ 和 $\hat{\beta}$ 附近有尖峰, 那么预测分布可以通过对 w 积分的方式简单地得到, 其中 α 和 β 被固定为 $\hat{\alpha}$ 和 $\hat{\beta}$

$$p(t|\mathbf{t}) \simeq p(t|\mathbf{t}, \hat{\alpha}, \hat{\beta}) = \int p(t|w, \hat{\beta}) p(w|\mathbf{t}, \hat{\alpha}, \hat{\beta}) dw \quad (5.41)$$

根据贝叶斯定理, α 和 β 的后验分布为

$$p(\alpha, \beta|\mathbf{t}) \propto p(\mathbf{t}|\alpha, \beta) p(\alpha, \beta) \quad (5.42)$$

如果先验分布相对比较平,那么在证据框架中, $\hat{\alpha}$ 和 $\hat{\beta}$ 可以通过最大化边缘似然函数 $p(t|\alpha, \beta)$ 来获得。

5.6 固定基函数的局限性

线性模型有一些重要的局限性,这使得我们要转而关注更加复杂的模型,例如支持向量机和神经网络。困难的产生主要是因为我们假设了基函数在观测到任何数据之前就被固定下来,而这正是维度灾难问题的一个表现形式。结果,基函数的数量随着输入空间的维度 D 迅速增长,通常是指数方式的增长。

幸运的是,真实数据集有两个性质,可以帮助我们缓解这个问题。

1. 数据向量 $\{x_n\}$ 通常位于一个非线性流形内部。由于输入变量之间的相关性,这个流形本身的维度小于输入空间的维度。如果我们使用局部基函数,那么我们可以让基函数只分布在输入空间中包含数据的区域。这种方法被用在径向基函数网络中,也被用在支持向量机和相关向量机当中。神经网络模型使用可调节的基函数,这些基函数有着 sigmoid 非线性的性质。神经网络可以通过调节参数,使得在输入空间的区域中基函数会按照数据流形发生变化。
2. 目标变量可能只依赖于数据流形中的少量可能的方向。利用这个性质,神经网络可以通过选择输入空间中基函数产生响应的方向。

第6章 分类的线性模型

分类的目标是将输入变量 x 分到 K 个离散的类别 C_k 中的某一类。所谓分类线性模型,是指决策面是输入向量 x 的线性函数,因此被定义为 D 维输入空间中的 $(D-1)$ 维超平面。

6.1 判别函数

判别函数是一个以向量 x 为输入,把它分配到 K 个类别中的某一个类别(记作 C_k) 的函数。

二分类

线性判别函数的最简单的形式是输入向量的线性函数,即

$$y(x) = \mathbf{w}^T \mathbf{x} + w_0 \quad (6.1)$$

其中 \mathbf{w} 被称为权向量 (weight vector), w_0 被称为偏置 (bias)。对应的决策边界由 $y(x) = 0$ 确定。对于一个输入向量 \mathbf{x} ,如果 $y(x) \geq 0$,那么它被分到 C_1 中,否则被分到 C_2 中。向量 \mathbf{w} 与决策面内的任何向量都正交,从而 \mathbf{w} 确定了决策面的方向。任意一点 x 到决策面的距离为

$$r = \frac{y(x)}{\|\mathbf{w}\|} \quad (6.2)$$

多分类

考虑把线性判别函数推广到 $K > 2$ 个类别。

- 1 对其他。使用 $K-1$ 个分类器,每个分类器用来解决一个二分类问题,把属于类别 C_k 和不属于那个类别的点分开。
2. 1 对 1。引入 $\frac{K(K-1)}{2}$ 个二元判别函数,对一对类别都设置一个判别函数。

上述两种方法都会产生输入空间无法分类的区域。通过引入 K 类判别函数可以避免这些问题。 K 类判别函数由 K 个线性函数组成,形式为

$$y_k(x) = \mathbf{w}_k^T \mathbf{x} + w_{k0} \quad (6.3)$$

对于点 \mathbf{x} ,如果所有的 $j \neq k$ 都有 $y_k(x) > y_j(x)$,那么就把它分到 C_k 。于是类别 C_k 和 C_j 之间的决策面为 $y_k(x) = y_j(x)$,并且对应于一个 $(D-1)$ 维超平面,形式为

$$(\mathbf{w}_k - \mathbf{w}_j)^T \mathbf{x} + (w_{k0} - w_{j0}) = 0 \quad (6.4)$$

这样的判别函数的决策区域总是单连通的,并且是凸的。现在介绍三种学习线性判别函数的参数的方法,即基于最小平方的方法、Fisher 线性判别函数,以及感知器算法。

用于分类的最小平方方法

最小平方误差函数的最小化产生了参数值的简单的解析解。使用最小平方方法的一个理由是它在给定输入向量的情况下,近似了目标值的条件期望 $\mathbb{E}[t|\mathbf{x}]$ 。对于二元表示方法,条件期望由后验类概率向量给出。但不幸的是,这些概率通常很难近似。事实上,近似的过程有可能产生位于区间 $(0, 1)$ 之外的值,这是因为线性模型的灵活性很受限。

使用向量记号,我们可以很容易地把这量聚焦在一起表示,即

$$\mathbf{y}(\mathbf{x}) = \tilde{\mathbf{W}}^T \tilde{\mathbf{x}} = \begin{pmatrix} w_{1,0} & w_{2,0} & \cdots & w_{K,0} \\ w_{1,1} & w_{2,1} & \cdots & w_{K,1} \\ \vdots & \vdots & \ddots & \vdots \\ w_{1,D} & w_{2,D} & \cdots & w_{K,D} \end{pmatrix}^T \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_D \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_K \end{pmatrix} \quad (6.5)$$

其中 $\tilde{\mathbf{W}}$ 是一个矩阵,第 k 列由 $D+1$ 维向量 $\tilde{\mathbf{w}}_k = (w_{k0}, \mathbf{w}_k^T)^T$ 组成, $\tilde{\mathbf{x}}$ 是对应的增广输入向量 $(1, \mathbf{x}^T)^T$,它带有一个虚输入 $x_0 = 1$ 。

通过最小化平方和误差函数来确定参数矩阵 $\tilde{\mathbf{W}}$,考虑一个训练数据集 $\mathbf{x}_n, \mathbf{t}_n$, 其中 $n = 1, \dots, N$, 然后定义一个矩阵 \mathbf{T} , 它的第 n 行是向量 \mathbf{t}_n^T 。还定义了一个矩阵 $\tilde{\mathbf{x}}_n^T$ 。这样,平方和误差函数可以写成

$$E_D(\tilde{\mathbf{W}}) = \frac{1}{2} \text{Tr}\{(\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \tilde{\mathbf{T}})^T(\tilde{\mathbf{X}}\tilde{\mathbf{W}} - \tilde{\mathbf{T}})\} \quad (6.6)$$

令上式关于 $\tilde{\mathbf{W}}$ 的导数等于零,整理,可以得到 $\tilde{\mathbf{W}}$ 的解,形式为

$$\tilde{\mathbf{W}} = (\tilde{\mathbf{X}}^T \tilde{\mathbf{X}})^{-1} \tilde{\mathbf{X}}^T \tilde{\mathbf{T}} = \tilde{\mathbf{X}}^\dagger \tilde{\mathbf{T}} \quad (6.7)$$

这样,我们得到了判别函数,形式为

$$\mathbf{y}(\mathbf{x}) = \tilde{\mathbf{W}}^T \tilde{\mathbf{x}} = \mathbf{T}^T (\tilde{\mathbf{X}}^\dagger)^T \tilde{\mathbf{x}} \quad (6.8)$$

最小平方方法对于判别函数的参数给出了精确的解析解。但是,最小平方解对于离群点缺少鲁棒性。最小平方方法对应于高斯条件分布假设下的最大似然法,而二值目标向量的概率分布显然不是高斯分布。通过使用更恰当的概率模型,我们会得到性质比最小平方方法更好的分类方法。

Fisher 线性判别函数

Fisher 提出的思想是最大化一个函数,这个函数能够让类均值的投影分开得较大,同时让每个类别内部的方差较小,从而最小化了类别的重叠。

首先考虑一个二分类问题,这个问题中有 C_1 类的 N_1 个点,以及 C_2 类的 N_2 个点。因此两类的均值向量为

$$\mathbf{m}_1 = \frac{1}{N_1} \sum_{n \in C_1} \mathbf{x}_n, \quad \mathbf{m}_2 = \frac{1}{N_2} \sum_{n \in C_2} \mathbf{x}_n \quad (6.9)$$

如果投影到 \mathbf{w} 上,那么最简单的度量类别之间分开程度的方式就是类别均值投影之后的距离。这说明,我们可以选择 \mathbf{w} 使得下式取得最大值

$$m_2 - m_1 = \mathbf{w}^T (\mathbf{m}_2 - \mathbf{m}_1) \quad (6.10)$$

其中

$$m_k = \mathbf{w}^T \mathbf{m}_k \quad (6.11)$$

是来自类别 C_k 的投影数据的均值。但是,通过增大 \mathbf{w} ,这个表达式可以任意大。为了解决这个问题,我们可以将 \mathbf{w} 限制为单位长度,即 $\sum_i w_i^2 = 1$ 。使用拉格朗日乘数法来进行有限制条件的最大化问题的求解,我们可以发现 $\mathbf{w} \propto (\mathbf{m}_2 - \mathbf{m}_1)$ 。

来自类别 C_k 的数据经过变换后的类内方差为

$$s_k^2 = \sum_{n \in C_k} (\mathbf{w}^T \mathbf{x}_n - m_k)^2 \quad (6.12)$$

可以把整个数据集的总的方差定义为 $s_1^2 + s_2^2$ 。Fisher 准则根据类间方差和类内方差的比值定义,即

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2} \quad (6.13)$$

利用公式 6.9, 6.11

$$\begin{aligned} (m_2 - m_1)^2 &= \left(\frac{1}{N_2} \sum_{n \in C_2} \mathbf{w}^T \mathbf{x}_n - \frac{1}{N_1} \sum_{n \in C_1} \mathbf{w}^T \mathbf{x}_n \right)^2 \\ &= \left(\mathbf{w}^T (\bar{\mathbf{x}}_{C_1} - \bar{\mathbf{x}}_{C_2}) \right)^2 \\ &= \mathbf{w}^T (\bar{\mathbf{x}}_{C_1} - \bar{\mathbf{x}}_{C_2}) (\bar{\mathbf{x}}_{C_1} - \bar{\mathbf{x}}_{C_2})^T \mathbf{w} \\ &= \mathbf{w}^T S_B \mathbf{w} \\ s_1^2 &= \frac{1}{N} \sum_{n \in C_1} (\mathbf{w}^T \mathbf{x}_n - \frac{1}{N} \sum_{n \in C_1} \mathbf{w}^T \mathbf{x}_n) (\mathbf{w}^T \mathbf{x}_n - \frac{1}{N} \sum_{n \in C_1} \mathbf{w}^T \mathbf{x}_n)^T \quad (6.14) \\ &= \frac{1}{N} \sum_{n \in C_1} \mathbf{w}^T (\mathbf{x}_n - \bar{\mathbf{x}}_{C_1}) (\mathbf{x}_n - \bar{\mathbf{x}}_{C_1})^T \mathbf{w} \\ &= \mathbf{w}^T S_1 \mathbf{w} \\ s_1^2 + s_2^2 &= \mathbf{w}^T (S_1 + S_2) \mathbf{w} \\ &= \mathbf{w}^T S_W \mathbf{w} \end{aligned}$$

显式地表达出 $J(\mathbf{w})$ 对 \mathbf{w} 的依赖

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}} \quad (6.15)$$

其中 S_B 是类间 (between-class) 协方差矩阵, S_W 被称为类内 (within-class) 协方差矩阵。对该式关于 \mathbf{w} 求导, 并令其为零, 有

$$\begin{aligned} \frac{\partial J(\mathbf{w})}{\partial \mathbf{w}} &= \frac{\partial [\mathbf{w}^T S_B \mathbf{w} \cdot (\mathbf{w}^T S_W \mathbf{w})^{-1}]}{\partial \mathbf{w}} \\ &= 2S_B \mathbf{w} (\mathbf{w}^T S_W \mathbf{w})^{-1} - 2\mathbf{w}^T S_B \mathbf{w} \cdot (\mathbf{w}^T S_W \mathbf{w})^{-2} \cdot S_W \mathbf{w} \\ &= 0 \\ &\Rightarrow S_B \mathbf{w} \cdot \mathbf{w}^T S_W \mathbf{w} = \mathbf{w}^T S_B \mathbf{w} \cdot S_W \mathbf{w} \\ &\Rightarrow S_W \mathbf{w} = \frac{\mathbf{w}^T S_W \mathbf{w}}{\mathbf{w}^T S_B \mathbf{w}} \cdot S_B \mathbf{w} \\ &\Rightarrow \mathbf{w} \propto S_W^{-1} S_B \mathbf{w} = S_W^{-1} (\mathbf{m}_2 - \mathbf{m}_1)(\mathbf{m}_2 - \mathbf{m}_1)^T \mathbf{w} \end{aligned} \quad (6.16)$$

我们看到 $S_B \mathbf{w}$ 总是在 $\mathbf{m}_2 - \mathbf{m}_1$ 的方向上。更重要的是, 我们不关心 \mathbf{w} 的大小, 只关心它的方向, 因此

$$\mathbf{w} \propto S_W^{-1} (\mathbf{m}_2 - \mathbf{m}_1) \quad (6.17)$$

公式 6.17 的结果被称为 Fisher 线性判别函数 (Fisher linear discriminant)。如果类内协方差是各向同性的, 从而 S_W 正比于单位矩阵, 那么我们看到 \mathbf{w} 正比于类均值的差。

与最小平方的关系

最小平方方法确定线性判别函数的目标是使模型的预测尽可能地与目标值接近。相反, Fisher 判别准则的目标是使输出空间的类别有最大的区分度。对于二分类问题, Fisher 准则可以看成最小平方的一个特例。

多分类的 Fisher 判别函数

感知器算法

线性判别模型的另一个例子是 Rosenblatt 提出的感知器算法。它对应于一个二分类的模型, 这个模型中, 输入向量 \mathbf{x} 首先使用一个固定的非线性变换得到一个特征向量 $\phi(\mathbf{x})$, 这个特征向量然后被用于构造一个一般的线性模型, 形式为

$$y(\mathbf{x}) = f(\mathbf{w}^T \phi(\mathbf{x})) \quad (6.18)$$

其中非线性激活函数 $f(\cdot)$ 是一个阶梯函数, 形式为

$$f(a) = \begin{cases} +1, & a \geq 0 \\ -1, & a < 0 \end{cases} \quad (6.19)$$

向量 $\phi(\mathbf{x})$ 通常包含一个偏置分量 $\phi_0(\mathbf{x}) = 1$ 。用来确定感知器的参数 \mathbf{w} 的算法可以很容易地从误差函数最小化的思想中得到。一个自然的选择是误分类的模型的总数。但是,这样做会使得学习算法不会很简单,因为这样做会使误差函数变为 \mathbf{w} 的分段常函数,从而当 \mathbf{w} 的变化使得决策边界移过某个数据点时,这个函数会不连续变化。这样会不连续变化。这样做还使得使用误差函数改变 \mathbf{w} 的方法无法使用,因为在几乎所有的地方梯度都等于零。

因此考虑一个另外的误差函数,被称为感知器准则 (perceptron criterion)。

$$E_P(\mathbf{w}) = - \sum_{n \in \mathcal{M}} \mathbf{w}^T \phi_n t_n \quad (6.20)$$

其中 $\phi_n = \phi(\mathbf{x}_n)$ 和 \mathcal{M} 表示所有误分类模型的集合。某个特定的误分类模型对于误差函数的贡献是 \mathbf{w} 空间中模式被误分类的区域中 \mathbf{w} 的线性函数,而在正确分类的区域,误差函数等于零。总的误差函数因此是分段线性的。我们现在对这个误差函数使用随机梯度下降算法。这样,权向量 \mathbf{w} 的变化为

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E_P(\mathbf{w}) = \mathbf{w}^{(\tau)} + \eta \phi_n t_n \quad (6.21)$$

感知器收敛定理 (perceptron convergence theorem) 表明,如果存在一个精确的解 (即,如果训练数据线性可分),那么感知器算法可以保证在有限步骤内找到一个精确解。

这里要画一个图

6.2 概率生成式模型

连续输入

最大似然解

离散特征

指数族分布

6.3 概率判别式模型

固定基函数

logistic 回归

迭代重加权最小平方

多类 logistic 回归

probit 回归

标准链接函数

6.4 拉普拉斯近似

模型比较和 BIC

6.5 贝叶斯 logistic 回归

拉普拉斯近似

预测分布

第7章 神经网络

在上两章,我们考虑了由固定基函数的线性组合构成的回归模型和分类模型。我们看到,这些模型具有一些有用的分析性质和计算性质,但是它们的实际应用被维数灾难问题限制了。为了将这些模型应用于大规模的问题,有必要根据数据调节基函数。

支持向量机是这样解决这个问题:首先定义以训练数据点为中心的基函数,然后在训练过程中选择一个子集。支持向量机的一个优点是,虽然训练阶段涉及到非线性优化,但是目标函数是凸函数,因此最优化问题的解相对很直接,并且通常随着数据规模的增加而增多。相关向量机也选择固定基函数集合的一个子集,通常会生成一个相当稀疏的模型。与支持向量机不同,相关向量机也产生概率形式的输出,嘎然这种输出的产生会以训练阶段的非凸优化为代价。

另一种方法是事先固定基函数的数量,但是允许基函数可调节。换名话,就是使用参数形式的基函数,这些参数可以在训练阶段调节。在模式识别中,这种类型的最成功的模型是有前馈神经网络,也被称为多层感知器 (multilayer perceptron)。与具有同样泛化能力的支持向量机相比,最终的模型会相当简洁,因此计算的速度更快。这种简洁性带来的代价就是,与相关向量机一样,构成了网络训练根基的似然函数不再是模型参数的凸函数。然而,在实际应用中,考察模型在训练阶段消耗的计算资源是很有价值的,这样做会得到一个简洁的模型,它可以快速地处理新数据。

首先,我们考虑神经网络的函数形式,包括基函数的具体参数,然后我们讨论使用最大似然框架确定神经网络参数的问题,这涉及到非线性最优化问题的解。这种方法需要计算对数似然函数关于神经网络参数的导数,我们会看到这些导数可以使用误差反向传播 (error backpropagation) 的方法高效地获得。我们还会说明误差反向传播的框架如何推广到计算其他的导数,例如 Jacobian 矩阵和 Hessian 矩阵。接下来,我们讨论神经网络训练的正则化和各种方法,以及方法之间的关系。我们还会考虑神经网络模型的一些扩展。特别地,我们会描述一个通用的框架,用来对条件概率密度建模。这个框架被称为混合密度网络 (mixture density network)。最后,我们讨论神经网络的贝叶斯观点。

7.1 前馈神经网络

深度前馈网络 (deep feedforward network) 也叫做前馈神经网络 (feedforward neural network) 或者多层感知机 (multilayer perceptron, MLP), 是典型的深度学习模型。前馈神经网络的目标是近似某个函数 f^* 。前馈网络定义了一个映射 $y = f(x; \theta)$, 并且学习参数 θ 的值, 使它能够得到最佳的函数近似。

前馈神经网络之所以被称作网络, 是因为它们通常用许多不同函数复合在一起来表示。该模型与一个有向无环图相关联, 而图描述了函数是如何复合在一起的。例如, 我们三个函数 $f^{(1)}, f^{(2)}, f^{(3)}$ 连接在一个链上以形成 $f^{(3)}(f^{(2)}(f^{(1)}(x)))$ 。

在神经网络训练过程中,我们让 $f(x)$ 去匹配 $f^*(x)$ 的值。训练数据为我们提供了在不同训练点上取值的、含有噪声的 $f^*(x)$ 近似实例。每个样本 x 都伴随着一个标签 $y \approx f^*(x)$ 。训练样本直接指明了输出层在每一点 x 上必须做什么;它必须产生一个接近 y 的值。但是训练数据并没有直接指明其他层应该怎么做。学习算法必须决定如何使用这些层来诞生想要的输出,但是训练数据并没有说每个单独的层应该做什么。相反,学习算法必须决定如何使用这些层来最好地实现 f^* 的近似。因为训练数据并没有给出这些层中的每一层所需的输出,所以这些层被称为隐藏层 (hidden layer)。

7.2 网络训练

7.3 误差反向传播

7.4 Hessian 矩阵

7.5 神经网络的正则化

7.6 混合密度网络

7.7 贝叶斯神经网络

第 8 章 组合模型

8.1 贝叶斯模型平均

8.2 委员会

8.3 提升方法

提升 (boosting) 方法是一种常用的统计学习方法,应用广泛且有效。在分类问题中,它通过改变训练样本的权重,学习多个分类器,并将这些分类器进行线性组合,提高分类的性能。

提升方法基于这样一种思路:对于一个复杂任务来说,将多个专家的判断进行适当的综合所得出的判断,要比其中任何一个专家单独的判断好。提升方法就是从弱学习算法出发,反复学习,得到一系列弱分类器(又称基本分类器),然后组合这些弱分类器,构成一个强分类器。对提升方法来说,有两个问题需要回答:

1. 在每一轮如何改变训练数据的权值或概率分布;
2. 如果将弱分类器组合成一个强分类器。

AdaBoost 的做法是,提高那些被前一轮弱分类器错误分类样本的权值,而降低那些被正确分类样本的权值。这样一来,那些没有得到正确分类的数据,由于其权值的加大而受到后一轮的弱分类器的更大关注。于是,分类问题被一系列的弱分类器“分而治之”。至于第 2 个问题,即弱分类器的组合,AdaBoost 采取加权表决的方法。具体地,加大分类误差率小的弱分类器的权值,使其在表决中起较大作用,减小分类误差率大的弱分类器的权值,使其在表决中起较小的作用。AdaBoost 的巧妙之处就在于它将这些想法自然且有效地实现在一种算法里。

AdaBoost 算法

输入:训练数据集 $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$; 弱学习算法;

输出:最终分类器 $G(x)$

- (1) 初始化训练数据的权值分布

$$D_1 = (w_{11}, \dots, w_{1i}, \dots, w_{1N}), w_{1i} = \frac{1}{N}, i = 1, 2, \dots, N \quad (8.1)$$

假设训练数据集具有均匀的权值分布,即每个训练样本在基本分类器的学习中作用相同,这一假设保证第 1 步能够在原始数据上学习基本分类器 $G_1(x)$ 。

- (2) 对 $m = 1, 2, \dots, M$

- a. 使用具有权值分布 D_m 的训练数据集学习, 得到基本分类器

$$G_m(x)X \rightarrow \{-1, +1\} \quad (8.2)$$

- b. 计算 $G_m(x)$ 在训练数据集上的分类误差率

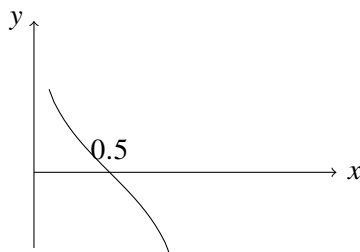
$$e_m = \sum_{i=1}^N P(G_m(x_i) \neq y_i) = \sum_{i=1}^N w_{mi} I(G_m(x_i) \neq y_i) \quad (8.3)$$

这表明, $G_m(x)$ 在加权的训练数据集上的分类误差率是被 $G_m(x)$ 误分类样本的权值之和。

- c. 计算 $G_m(x)$ 的系数

$$a_m = \frac{1}{2} \log \frac{1 - e_m}{e_m} \quad (8.4)$$

a_m 表示 $G_m(x)$ 在最终分类器中的重要性。



由上图可知, 当 $e_m \leq \frac{1}{2}$ 时, $a_m \geq 0$, 并且 a_m 随着 e_m 的减小而增大, 所以**分类误差率越小的基本分类器在最终分类器中的作用越大**。

- d. 更新训练数据集的权值分布

$$\begin{aligned} D_{m+1} &= (w_{m+1,1}, \dots, w_{m+1,i}, \dots, w_{m+1,N}) \\ w_{m+1,i} &= \frac{w_{mi}}{Z_m} \exp(-a_m y_i G_m(x_i)), i = 1, 2, \dots, N \end{aligned} \quad (8.5)$$

这里, Z_m 是规范化因子

$$Z_m = \sum_{i=1}^N w_{mi} \exp(-a_m y_i G_m(x_i)) \quad (8.6)$$

它使 D_{m+1} 成为一个概率分布。式 8.5 可以写成

$$w_{m+1,i} = \begin{cases} \frac{w_{mi}}{Z_m} e^{-a_m}, & G_m(x_i) = y_i \\ \frac{w_{mi}}{Z_m} e^{a_m}, & G_m(x_i) \neq y_i \end{cases} \quad (8.7)$$

由此可知, **被基本分类器 $G_m(x)$ 误分类样本的权值得以扩大, 而被正确分类样本的权值却得以缩小**。因此, 误分类样本在下一轮学习中起更大的作用。不改变所给的训练数据, 而不断改变训练数据权值的分布, 使得训练数据在基本分

类器的学习中起不同的作用,这是 AdaBoost 的一个特点。

(3) 构建基本分类器的线性组合

$$f(x) = \sum_{m=1}^M a_m G_m(x) \quad (8.8)$$

得到最终分类器

$$G(x) = \text{sign}(f(x)) = \text{sign}\left(\sum_{m=1}^M a_m G_m(x)\right) \quad (8.9)$$

线性组合 $f(x)$ 实现 M 个基本分类器的加权表决。系数 a_m 表示了基本分类器 $G_m(x)$ 的重要性,这里,所有 a_m 之和并不为 1。 $f(x)$ 的符号决定实例 x 的类, $f(x)$ 的绝对值表示分类的确信度。利用基本分类器的线性组合构建最终分类器是 AdaBoost 的另一个特点。

AdaBoost 算法的训练误差分析

AdaBoost 最基本的性质是它能在学习过程中不断减少训练误差,即在训练数据集上的分类误差率。

AdaBoost 算法的解释

AdaBoost 算法还有另一个解释,即可以认为 AdaBoost 算法是模型为加法模型、损失函数为指数函数、学习算法为前向分步算法时的二类分类学习方法

8.4 基于树的模型

8.5 条件混合模型

8.6 logistic 模型的混合

第9章 支持向量机

支持向量机(support vector machines, SVM)是一种二类分类模型。它的基本模型是定义在特征空间上的间隔最大的线性分类器,间隔最大使它有别于感知机;支持向量机还包括核技巧,这使它成为实质上的非线性分类器。支持向量机的学习策略就是间隔最大化,可形式化为一个求解凸二次规划(convex quadratic programming)的问题,也等价于正则化的合页损失函数的最小化问题。支持向量机的学习算法是求解凸二次规划的最优化算法。

支持向量机学习方法包含构建由简至繁的模型:线性可分支持向量机(linear support vector machine in linearly separable case)、线性支持向量机(linear support vector machine)及非线性支持向量机(non-linear support vector machine)。简单模型是复杂模型的基础,也是复杂模型的特殊情况。当训练数据线性可分时,通过软件间隔最大化(hard margin maximization),学习一个线性的分类器,即线性可分支持向量机,又称为硬间隔支持向量机;当训练数据近似线性可分时,通过软件间隔最大化(soft margin maximization),也学习一个线性的分类器,即线性支持向量机,又称谓软间隔支持向量机;当训练数据线性不可分时,通过使用核技巧(kernel trick)及软间隔最大化,学习非线性支持向量机。

当输入空间为欧氏空间或离散集合、特征空间为希尔伯特空间时,核函数(kernel function)表示将输入从输入空间映射到特征空间得到的特征向量之间的内积。通过使用核函数可以学习非线性支持向量机,等价于隐式地在高维的特征空间中学习线性支持向量机。这样的方法称为核技巧。核方法(kernel method)是比支持向量机更为一般的机器学习方法。

Cortes 与 Vapnik 提出线性支持向量机, Boser, Guyon 与 Vapnik 又引入核技巧, 提出非线性支持向量机。

9.1 间隔与支持向量

给定训练样本集 $D = \{(x_1, y_1), \dots, (x_m, y_m)\}$, $y_i \in \{-1, +1\}$, 分类学习最基本的想法是基于训练集 D 在样本空间中找到一个划分超平面, 将不同类别的样本分开。如图 9.1

样本空间中任意点 x 到超平面 (w, b) 的距离可写为

$$\gamma = \frac{y(w^T x + b)}{\|w\|} = \frac{yf(x)}{\|w\|}$$

注: $yf(x)$ 相当于 $|f(x)|$ 。

假设超平面 (w, b) 能将训练样本正确分类, 即对于 $(x_i, y_i) \in D$, 若 $y_i = +1$, 则有 $w^T x_i + b > 0$; 若 $y_i = -1$, 则有 $w^T x_i + b < 0$ 。令

$$\begin{cases} w^T x_i + b \geq +1, & y_i = +1; \\ w^T x_i + b \leq -1, & y_i = -1. \end{cases} \quad (9.1)$$

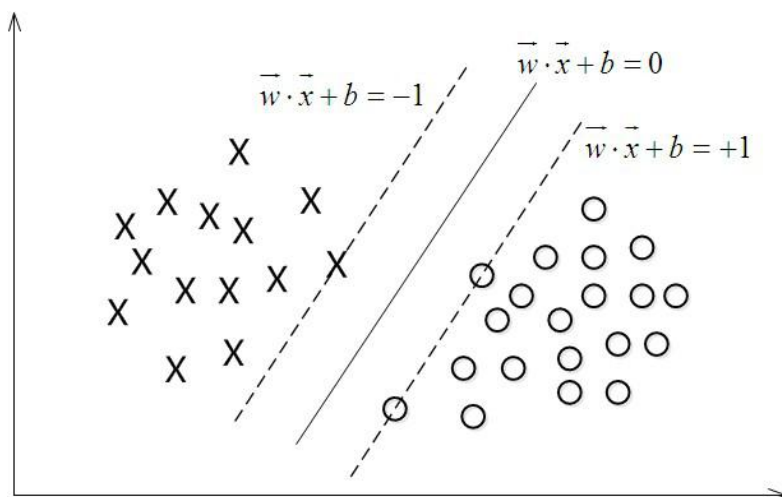


图 9.1

距离超平面最近的这几个训练样本点使式 9.1 的等号成立, 它们被称为“支持向量 (support vector)”, 两个异类支持向量到超平面的距离之和为

$$\gamma = \frac{2}{\|\mathbf{w}\|}, \quad (9.2)$$

它被称为“间隔”(margin)。

欲找到具有“最大间隔”的划分超平面, 也就是要找到能满足式 9.1 中约束的参数 \mathbf{w} 和 b , 使得 γ 最大, 即

$$\begin{aligned} \min \quad & \frac{1}{2} \|\mathbf{w}\|^2, \\ \text{s.t.} \quad & y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1, \quad i = 1, \dots, m. \end{aligned} \quad (9.3)$$

这就是支持向量机的基本型。

9.2 对偶问题

注意到式 9.3 本身是一个凸二次规划问题, 能直接用现成的优化计算包求解, 但我们可以有更高效率的办法。由于这个问题的特殊结构, 还可以通过拉格朗日对偶性变换到对偶变量的优化问题, 即通过求解与原问题等价的对偶问题得到原始问题的最优解, 这就是线性可分条件下支持向量机的对偶算法, 这样做的优点在于:

1. 对偶问题往往更容易求解;
2. 可以自然的引入核函数, 进而推广到非线性分类问题。

该问题的拉格朗日函数可写为

$$L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^n a_i (y_i(\mathbf{w}^T \mathbf{x}_i + b) - 1) \quad (9.4)$$

然后令

$$\theta(\mathbf{w}) = \max_{a_i \geq 0} L(\mathbf{w}, b, \mathbf{a}) \quad (9.5)$$

具体写出来,目标函数变成了

$$\min_{\mathbf{w}, b} \theta(\mathbf{w}) = \min_{\mathbf{w}, b} \max_{a_i \geq 0} L(\mathbf{w}, b, a) = p^* \quad (9.6)$$

这里用 p^* 表示这个问题的最优值,且和最初的问题是等价的。如果直接求解,那么一上来便得面对 w 和 b 两个参数,而 a_i 以是不等式约束,这个求解过程不好做。考虑对偶问题

$$\min_{\mathbf{w}, b} \theta(\mathbf{w}) = \max_{a_i \geq 0} \min_{\mathbf{w}, b} L(\mathbf{w}, b, a) = d^* \quad (9.7)$$

原始问题通过满足 KKT 条件,已经转化成了对偶问题。而求解这个对偶问题,分为 3 个步骤

1. 让 $L(\mathbf{w}, b, a)$ 关于 \mathbf{w} 和 b 最小化

首先固定 a , 要让 L 关于 w 和 b 最小化, 分别对 w 和 b 求偏导, 令其等于 0;

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{w}} &= \|\mathbf{w}\| - \sum_{i=1}^n a_i y_i x_i \mathbf{w}^T = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^n a_i y_i x_i \\ \frac{\partial L}{\partial b} &= \sum_{i=1}^n a_i y_i = 0 \Rightarrow \sum_{i=1}^n a_i y_i = 0 \end{aligned} \quad (9.8)$$

将以上结果代入之前的 L , 得到

$$\begin{aligned} L(\mathbf{w}, b, a) &= \frac{1}{2} \sum_{i,j=1}^n a_i a_j y_i y_j x_i^T x_j - \sum_{i,j=1}^n a_i a_j y_i y_j x_i^T x_j - b \sum_{i=1}^n a_i y_i + \sum_{i=1}^n a_i \\ &= \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i,j=1}^n a_i a_j y_i y_j x_i^T x_j \end{aligned} \quad (9.9)$$

2. 求对 a 的极大

求对 a 的极大, 即是关于对偶问题的最优化问题。经过上一个步骤的求解, 得到的拉格朗日函数式子已经没有了变量 w 和 b , 只有 a 。从上面的式子得到

$$\begin{aligned} \max_a \quad & \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i,j=1}^n a_i a_j y_i y_j x_i^T x_j \\ s.t. \quad & a_i \geq 0, i = 1, \dots, n \\ & \sum_{i=1}^n a_i y_i = 0 \end{aligned} \quad (9.10)$$

这样, 求出了 a_i , 从而根据

$$\begin{aligned} \mathbf{w}^* &= \sum_{i=1}^n a_i y_i x_i \\ b^* &= -\frac{\max \mathbf{w}^{*T} x_i + \min \mathbf{w}^{*T} x_i}{2} \end{aligned} \quad (9.11)$$

即可求出 w, b , 最终得出分离超平面和分类决策函数。

3. 利用 SMO 算法求解对偶问题中的拉格朗日乘子

在求得 $L(w, b, a)$ 关于 w 和 b 最小化和对 a 的极大之后, 最后一步便是利用 SMO 算法求解对偶问题中的拉格朗日乘子

9.3 序列最小最优算法

接上一小节, 讨论支持向量机学习的实现问题。讲述其中的序列最小最优优化 (sequential minimal optimization, SMO) 算法。

SMO 算法是一种启发式算法, 其基本思路是: **如果所有变量的解都满足此最优化问题的 KKT 条件, 那么这个最优化问题的解就得到了。**因为 KKT 条件是该最优化问题的充分必要条件。否则, 选择两个变量, 固定其他变量, 针对这两个变量构建一个二次规划问题。这个二次规划问题关于这两个变量的解应该更接近原始二次规划问题的解, 因为这会使得原始二次规划问题的目标函数值变得更小。重要的是, 这时子问题可以通过解析方法求解, 这样就可以大大提高整个算法的计算计算速度。子问题有两个变量, 一个是违反 KKT 条件最严重的那一个, 另一个由约束条件自动确定。如此, SMO 算法将原问题不断分解为子问题并对子问题求解, 进而达到求解原问题的目的。

SMO 算法要解如下凸二次规划的对偶问题:

$$\min_a \quad \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N a_i a_j y_i y_j K(x_i, x_j) - \sum_{i=1}^N a_i \quad (9.12)$$

$$s.t. \quad C \geq a_i \geq 0, \quad i = 1, \dots, n \quad (9.13)$$

$$\sum_{i=1}^N a_i y_i = 0 \quad (9.14)$$

注意, 子问题的两个变量中只有一个是自由变量。假设 a_1, a_2 为两个变量, a_3, a_4, \dots, a_N 固定, 那么由等式约束 9.14 可知

$$a_1 = -y_1 \sum_{i=2}^N a_i y_i \quad (9.15)$$

如果 a_2 确定, 那么 a_1 也随之确定。所以子问题中同时更新两个变量。

整个 SMO 算法包括两个部分: **求解两个变量二次规划的解析方法和选择变量的启发式方法。**

于是 SMO 的最优化问题 9.12 的子问题可以写成

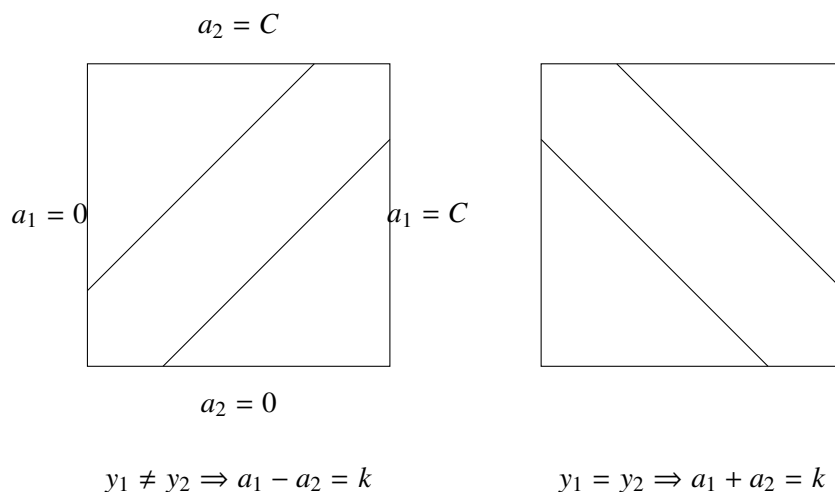
$$\begin{aligned} \min_{a_1, a_2} W(a_1, a_2) = & \frac{1}{2}K_{11}a_1^2 + \frac{1}{2}K_{22}a_2^2 + y_1y_2K_{12}a_1a_2 \\ & - (a_1 + a_2) + y_1a_1 \sum_{i=3}^N y_i a_i K_{i1} + y_2a_2 \sum_{i=3}^N y_i a_i K_{i2} \end{aligned} \quad (9.16)$$

$$s.t \ a_1y_1 + a_2y_2 = - \sum_{i=3}^N y_i a_i = \varsigma \quad (9.17)$$

$$0 \leq a_i \leq C, \ i = 1, 2 \quad (9.18)$$

其中, $K_{ij} = K(x_i, x_j), i, j = 1, 2, \dots, N, \varsigma$ 是常数。

为了求解两个变量的二次规划问题 9.16, 首先分析约束条件, 然后在此约束条件下求极小。由于只有两个变量 (a_1, a_2) , 约束可以用二维空间中的图形表示, 如图所示



不等式约束使得 (a_1, a_2) 在盒子 $[0, C] \times [0, C]$ 内, 等式约束使 (a_1, a_2) 在平行盒子的对角线的直线上。因此要求的是目标函数在一条平行于对角线线的线段上的最优值。这使得两个变量的最优化问题成为实质上的单变量的最优化问题, 不妨考虑为变量 a_2 的最优化问题。

假设问题 9.16 的初始可行解为 a_1^{old}, a_2^{old} , 最优解为 a_1^{new}, a_2^{new} , 并且假设在沿着约束方向未经剪辑时 a_2 的最优解为 $a_2^{new, unc}$ 。

引进记号

$$g(x) = \sum_{i=1}^N a_i y_i K(x_i, x) + b \quad (9.19)$$

$$v_i = \sum_{j=3}^N y_j a_j K(x_i, x_j) = g(x_i) - \sum_{j=1}^2 a_j y_j K(x_i, x_j) - b, \ i = 1, 2 \quad (9.20)$$

目标函数可写成

$$W(a_1, a_2) = \frac{1}{2}K_{11}a_1^2 + \frac{1}{2}K_{22}a_2^2 + y_1y_2K_{12}a_1a_2 - (a_1 + a_2) + y_1a_1v_1 + y_2a_2v_2 \quad (9.21)$$

令

$$E_i = g(x_i) - y_i = \left(\sum_{j=1}^N a_j y_j K(x_j, x_i) + b \right) - y_i, \quad i = 1, 2 \quad (9.22)$$

当 $i = 1, 2$ 时, E_i 为函数 $g(x)$ 对输入 x_i 的预测值与真实输出 y_i 之差。由 $a_1y_1 = \varsigma - a_2y_2$ 及 $y_i^2 = 1$, 可将 a_1 表示为

$$a_1 = (\varsigma - y_2a_2)y_1 \quad (9.23)$$

代入式 9.21, 得到只是 a_2 的函数的目标函数, 对 a_2 求导数

$$\frac{\partial W}{\partial a_2} = K_{11}a_2 + K_{22}a_2 - 2K_{12}a_2 - K_{11}\varsigma y_2 + K_{12}\varsigma y_2 + y_1y_2 - 1 - v_1v_2 + y_2v_2 \quad (9.24)$$

令其为 0, 得到

$$\begin{aligned} (K_{11} + K_{22} - 2K_{12})a_2 &= y_2(y_2 - y_1 + \varsigma K_{11} - \varsigma K_{12} + v_1 - v_2) \\ &= y_2 \left[y_2 - y_1 + \varsigma K_{11} - \varsigma K_{12} + \left(g(x_1) - \sum_{j=1}^2 y_j a_j K_{1j} - b \right) \right. \\ &\quad \left. \left(g(x_2) - \sum_{j=1}^2 y_j a_j K_{2j} - b \right) \right] \end{aligned} \quad (9.25)$$

将 $\varsigma = a_1^{old}y_1 + a_2^{old}y_2$ 代入, 得到

$$\begin{aligned} (K_{11} + K_{22} - 2K_{12})a_2^{new,unc} &= y_2((K_{11} + K_{22} - 2K_{12})a_2^{old}y_2 + y_2 - y_1 + g(x_1) - g(x_2)) \\ &= (K_{11} + K_{22} - 2K_{12})a_2^{old} + y_2(E_1 - E_2) \end{aligned} \quad (9.26)$$

将 $\eta = K_{11} + K_{22} - 2K_{12}$ 代入, 于是得到

$$a_2^{new,unc} = a_2^{old} + \frac{y_2(E_1 - E_2)}{\eta} \quad (9.27)$$

要使其满足不等式的约束必须将其限制在区间 $[L, H]$ 内, 从而得到 a_2^{new} 的表达式

$$a_2^{new} = \begin{cases} H, & a_2^{new,unc} > H \\ a_2^{new,unc}, & L \leq a_2^{new,unc} \leq H \\ L, & a_2^{new,unc} < L \end{cases} \quad (9.28)$$

如果 $y_1 \neq y_2$

$$L = \max(0, a_2^{old} - a_1^{old}), \quad H = \min(C, C + a_2^{old} - a_1^{old})$$

如果 $y_1 = y_2$

$$L = \max(0, a_2^{old} + a_1^{old} - C), \quad H = \min(C, a_2^{old} + a_1^{old})$$

变量的选择方法

SMO 算法在每个子问题中选择两个变量优化, 其中至少一个变量是违反 KKT 条件的。

1. 第 1 个变量的选择

SMO 称选择第 1 个变量的过程为外层循环。外层循环在训练样本中选取违反 KKT 条件最严重的样本点, 并将其对应的变量作为第 1 个变量。具体地, 检验训练样本点 (x_i, y_i) 是否满足 KKT 条件, 即

$$a_i = 0 \Leftrightarrow y_i g(x_i) \geq 1 \quad (9.29)$$

$$C > a_i > 0 \Leftrightarrow y_i g(x_i) = 1 \quad (9.30)$$

$$a_i = C \Leftrightarrow y_i g(x_i) \leq 1 \quad (9.31)$$

2. 第 2 个变量的选择

SMO 称选择第 2 个变量的过程为内层循环, 假设在外层循环中已经找到第 1 个变量 a_1 , 现在要在内层循环中找到第 2 个变量 a_2 。第 2 个变量选择的标准是希望能使 a_2 有足够大的变化。由式 9.27 知, a_2^{new} 是依赖于 $|E_1 - E_2|$ 的, 加了加快计算速度, 一种简单的做法是选择 a_2 , 使其对应的 $|E_1 - E_2|$ 最大。

3. 计算阈值 b 和差值 E_i

9.4 核函数

SVM 处理线性可分的情况, 而对于非线性的情况, SVM 的处理方法是选择一个核函数 $K < \cdot, \cdot >$, 通过将数据映射到高维空间, 来解决在原始空间中线性不可分的问题。

此外, 用对偶形式表示学习器的优势在于该表示中可调参数的个数不依赖输入属性的个数, 通过使用恰当的核函数来替代内积, 可以隐式得将非线性的训练数据映射到高维空间, 而不增加可调参数的个数。

在线性不可分的情况下, 支持向量机首先在低维空间中完成计算, 然后通过核函数将输入空间映射到高维性空间, 最终在高维特征空间中构造出最优分离超平面, 从而把平面上本身不好分的非线性数据分开。

而在我们遇到核函数之前, 如果用原始的方法, 那么在用线性学习器学习一个非线性关系, 需要选择一个非线性特征集, 并且将数据写成新的表达形式, 这等价于应用一个固定的非线性映射, 将数据映射到特征空间, 在特征空间中使用线性学习器, 因此考虑的假设集是这种类型的函数:

$$f(x) = \sum_{i=1}^N w_i \phi_i(x) + b \quad (9.32)$$

这里 $\phi: X \rightarrow F$ 是从输入空间到某个特征空间的映射, 这意味着建立非线性学习器分为

两步：

1. 首先使用一个非线性映射将数据变换到一个特征空间 F
2. 然后在特征空间使用线性学习器分类

而由于对偶形式就是线性学习器的一个重要性质，这意味着假设可以表达为训练点的线性组合，因此决策规则可以用测试点和训练点的内积来表示：

$$f(x) = \sum_{i=1}^l a_i y_i < \phi(x_i), \phi(x) > + b \quad (9.33)$$

如果有一种方式可以在特征空间中直接计算内积 $< \phi(x_i), \phi(x) >$ ，就像在原始输入点的函数中一样，就有可能将两个步骤融合到一起建立一个非线性的学习器，这样直接计算的方法称为**核函数方法**

定义 9.1. Kernel

核是一个函数 K ，对所有 $x, z \in X$ ，满足 $K(x, z) = < \phi(x), \phi(z) >$ ，这里 ϕ 是从 X 到内积特征空间 F 的映射。

下面举一个核函数把低维空间映射到高维空间的例子：

我们考虑核函数 $K(v_1, v_2) = < v_1, v_2 >^2$ ，即“内积平方”，这里 $v_1 = (x_1, y_1)$ ， $v_2 = (x_2, y_2)$ 是二维空间中的两个点。这个核函数对应着一个二维空间到三维空间的映射，它的表达式是：

$$P(x, y) = (x^2, \sqrt{2}xy, y^2)$$

可以验证，

$$\begin{aligned} < P(v_1), P(v_2) > &= < (x_1^2, \sqrt{2}x_1y_1, y_1^2), (x_2^2, \sqrt{2}x_2y_2, y_2^2) > \\ &= x_1^2x_2^2 + 2x_1x_2y_1y_2 + y_1^2y_2^2 \\ &= (x_1x_2 + y_1y_2)^2 \\ &= < v_1, v_2 >^2 \\ &= K(v_1, v_2) \end{aligned}$$

上面的例子所说，核函数的作用就是隐含着一个从低维空间到高维空间的映射，而这个映射可以把低维空间中线性不可分的两类点变成线性可分的。

核函数的本质

1. 实际中，我们会经常遇到线性不可分的样例，此时，我们的常用做法是把特征映射到高维空间中去
2. 但进一步，如果凡是遇到线性不可分的样例，一律映射到高维空间，那么这个维度大小会高到可怕的。那咋办呢？
3. 此时，核函数就隆重登场了，核函数的价值在于它虽然也是讲特征进行从低维到高维的转换，但核函数绝就绝在它事先在低维上进行计算，而将实质上的分类效果表在了高维上，也就避免了直接在高维空间中的复杂计算。

几个核函数：

- 多项式核 $K(x_1, x_2) = (< x_1, x_2 > + R)^d$
- 高斯核 $K(x_1, x_2) = \exp(-\|x_1 - x_2\|^2 / 2\sigma^2)$
- 线性核 $K(x_1, x_2) = < x_1, x_2 >$

9.5 软间隔与正则化

9.6 支持向量回归

9.7 核方法

第 10 章 混合模型和 EM 算法

10.1 一般形式的 EM 算法

EM 算法是一种迭代算法, 1977 年由 Dempster 等人总结提出的, 用于含有隐变量 (hidden variable) 的概率模型参数的极大似然估计, 或极大后验概率估计。EM 算法的每次迭代由两步组成: E 步, 求期望 (expectation); M 步, 求极大 (maximization)。所以这一算法称为期望极大算法 (expectation maximization), 简称 EM 算法。

概率模型有时既含有观测变量 (observable variable), 又含有隐变量或潜在变量 (latent variable)。如果概率模型的变量都是观测变量, 那么给定数据, 可以直接用极大似然估计法, 或贝叶斯估计法估计模型参数。但是, 当模型含有隐变量时, 就不能简单地使用这些估计方法。EM 算法就是含有隐变量的概率模型参数的极大似然估计法, 或极大后验概率估计法。仅讨论极大似然估计, 极大后验概率估计与其类似。

EM 算法的导出

为什么 EM 算法能近似实验对观测数据的极大似然估计呢? 下面通过近似求解观测数据的对数似然函数的极大化问题来导出 EM 算法。

我们面对一个含有隐变量的概率模型, 目标是极大化观测数据 (不完全数据) Y 关于参数 θ 的对数似然函数, 即极大化

$$\begin{aligned} L(\theta) &= \log P(Y|\theta) = \log \sum_Z P(Y, Z|\theta) \\ &= \log \left(\sum_Z P(Y|Z, \theta) P(Z|\theta) \right) \end{aligned} \quad (10.1)$$

注意到这一极大化的主要困难是式 10.1 中有未观测数据并有包含和 (或积分) 的对数。

事实上, EM 算法是通过迭代逐步极大化 $L(\theta)$ 的。假设在第 i 次迭代后 θ 的估计值是 $\theta^{(i)}$ 。我们希望新估计值 θ 能使 $L(\theta)$ 增加, 即 $L(\theta) > L(\theta^{(i)})$, 并逐步达到极大值。为此, 考虑两者的差:

$$L(\theta) - L(\theta^{(i)}) = \log \left(\sum_Z P(Y|Z, \theta) P(Z|\theta) \right) - \log P(Y|\theta^{(i)}) \quad (10.2)$$

利用 Jensen 不等式¹(Jensen inequality) 得到其下界:

$$\begin{aligned}
 L(\theta) - L(\theta^{(i)}) &= \log \left(\sum_Z P(Y|Z, \theta^{(i)}) \frac{P(Y|Z, \theta)P(Z|\theta)}{P(Y|Z, \theta^{(i)})} \right) - \log P(Y|\theta^{(i)}) \\
 &\geq \sum_Z P(Y|Z, \theta^{(i)}) \log \frac{P(Y|Z, \theta)P(Z|\theta)}{P(Y|Z, \theta^{(i)})} - \log P(Y|\theta^{(i)}) \\
 &= \sum_Z P(Y|Z, \theta^{(i)}) \log \frac{P(Y|Z, \theta)P(Z|\theta)}{P(Y|Z, \theta^{(i)})P(Y|\theta^{(i)})}
 \end{aligned} \tag{10.3}$$

令

$$B(\theta, \theta^{(i)}) = L(\theta^{(i)}) + \sum_Z P(Y|Z, \theta^{(i)}) \log \frac{P(Y|Z, \theta)P(Z|\theta)}{P(Y|Z, \theta^{(i)})P(Y|\theta^{(i)})} \tag{10.4}$$

则

$$L(\theta) \geq B(\theta, \theta^{(i)}) \tag{10.5}$$

即函数 $B(\theta, \theta^{(i)})$ 是 $L(\theta)$ 的一个下界, 因此任何可以使 $B(\theta, \theta^{(i)})$ 增大的 θ , 也可以使 $L(\theta)$ 增大。为了使 $L(\theta)$ 尽可能大的增大, 选择 $\theta^{(i+1)}$ 使 $B(\theta, \theta^{(i)})$ 达到极大, 即

$$\theta^{(i+1)} = \arg \max_{\theta} B(\theta, \theta^{(i)}) \tag{10.6}$$

现在求 $\theta^{(i+1)}$ 的表达式。省去对 θ 的极大化而言是常数的项, 有

$$\begin{aligned}
 \theta^{(i+1)} &= \arg \max_{\theta} \left(L(\theta^{(i)}) + \sum_Z P(Y|Z, \theta^{(i)}) \log \frac{P(Y|Z, \theta)P(Z|\theta)}{P(Y|Z, \theta^{(i)})P(Y|\theta^{(i)})} \right) \\
 &= \arg \max_{\theta} \left(\sum_Z P(Y|Z, \theta^{(i)}) \log (P(Y|Z, \theta)P(Z|\theta)) \right) \\
 &= \arg \max_{\theta} \left(\sum_Z P(Y|Z, \theta^{(i)}) \log (P(Y, Z|\theta)) \right) \\
 &= \arg \max_{\theta} Q(\theta, \theta^{(i)})
 \end{aligned} \tag{10.7}$$

式 10.7 等价于 EM 算法的一次迭代, 即求 Q 函数及其极大化。EM 算法是通过不断求解下界的极大化逼近求解对数似然函数极大化的算法。

定义 10.1. Q 函数

完全数据的对数似然函数 $\log P(Y, Z|\theta)$ 关于在给定观测数据 Y 和当前参数 $\theta^{(i)}$ 下对未观测数据 Z 的条件概率分布 $P(Z|Y, \theta^{(i)})$ 的期望称为 Q 函数。即

$$Q(\theta, \theta^{(i)}) = E_Z[\log P(Y, Z|\theta)|Y, \theta^{(i)}] \tag{10.8}$$

¹这里用到的是 $\log \sum_j \lambda_j y_j \geq \sum_j \lambda_j \log y_j$

EM 算法

输入: 观测变量数据 Y , 隐变量数据 Z , 联合分布 $P(Y, Z|\theta)$, 条件分布 $P(Z|Y, \theta)$;

输出: 模型参数 θ

1. 选择参数的初值 $\theta^{(0)}$, 开始迭代;
2. E 步: 记 $\theta^{(i)}$ 为第 i 次迭代参数 θ 的估计值, 在第 $i+1$ 次迭代的 E 步, 计算

$$\begin{aligned} Q(\theta, \theta^{(i)}) &= E_Z[\log P(Y, Z|\theta)|Y, \theta^{(i)}] \\ &= \sum_Z \log P(Y, Z|\theta) P(Z|Y, \theta^{(i)}) \end{aligned} \quad (10.9)$$

这里, $P(Z|Y, \theta^{(i)})$ 是在给定观测数据 Y 和当前的参数估计 $\theta^{(i)}$ 下隐变量数据 Z 的条件概率分布;

3. M 步: 求使 $Q(\theta, \theta^{(i)})$ 极大化的 θ , 确定第 $i+1$ 次迭代的参数的估计值 $\theta^{(i+1)}$

$$\theta^{(i+1)} = \arg \max_{\theta} Q(\theta, \theta^{(i)}) \quad (10.10)$$

4. 重复第 (2)、(3) 步, 直到收敛。

EM 算法的收敛性

EM 算法提供一种近似计算含有隐变量概率模型的极大似然估计的方法。EM 算法的最大优点是简单性和普适性。我们自然地要问: EM 算法得到的估计序列是否收敛? 如果收敛, 是否收敛到全局最大值或局部极大值?

定理 10.1

设 $P(Y|\theta)$ 为观测数据的似然函数, $\theta^{(i)} (i = 1, 2, \dots)$ 为 EM 算法得到的参数估计序列, $P(Y|\theta^{(i)}) (i = 1, 2, \dots)$ 为对应的似然函数序列, 则 $P(Y|\theta^{(i)})$ 是单调递增的, 即

$$P(Y|\theta^{(i+1)}) \geq P(Y|\theta^{(i)}) \quad (10.11)$$

设 $L(\theta) = \log P(Y|\theta)$ 为观测数据的对数似然函数, $L(\theta^{(i)}) (i = 1, 2, \dots)$ 为对应的对数似然函数序列。

1. 如果 $P(Y|\theta)$ 有上界, 则 $L(\theta^{(i)}) = \log P(Y|\theta^{(i)})$ 收敛到某一值 L^*
2. 在函数 $Q(\theta, \theta')$ 与 $L(\theta)$ 满足一定条件下, 由 EM 算法得到的参数估计序列 $\theta^{(i)}$ 的收敛值 θ^* 是 $L(\theta)$ 的稳定点



10.2 EM 的另一种观点

10.3 K 均值聚类

10.4 混合高斯

EM 算法的一个重要应用是高斯混合模型的参数估计。高斯混合模型应用广泛,在许多情况下,EM 算法是学习高斯混合模型 (Gaussian mixture model) 的有效方法。

定义 10.2. 高斯混合模型

高斯混合模型是指具有如下形式的概率分布模型:

$$P(y|\theta) = \sum_{k=1}^K \alpha_k \phi(y|\theta_k) \quad (10.12)$$

其中, α_k 是系数, $\alpha_k \geq 0$, $\sum_{k=1}^K \alpha_k = 1$; $\phi(y|\theta_k)$ 是高斯分布密度, $\theta_k = (\mu_k, \sigma_k^2)$,

$$\phi(y|\theta_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(y - \mu_k)^2}{2\sigma_k^2}\right) \quad (10.13)$$

称为第 k 个模型。



高斯混合模型参数估计的 EM 算法

假设观测数据 y_1, y_2, \dots, y_N 由高斯混合模型生成,

$$P(y|\theta) = \sum_{k=1}^K \alpha_k \phi(y|\theta_k) \quad (10.14)$$

其中, $\theta = (\alpha_1, \alpha_2, \dots, \alpha_K; \theta_1, \theta_2, \dots, \theta_K)$ 。我们用 EM 算法估计高斯混合模型的参数 θ

1. 明确隐变量, 写出完全数据的对数似然函数

隐变量由 γ_{jk} 表示, 其定义如下:

$$\gamma_{jk} = \begin{cases} 1, & \text{第 } j \text{ 个观测来自第 } k \text{ 个模型} \\ 0, & \text{否则} \end{cases} \quad (10.15)$$

γ_{jk} 是 0-1 随机变量。那么完全数据是

$$(y_j, \gamma_{j1}, \gamma_{j2}, \dots, \gamma_{jK}), \quad j = 1, 2, \dots, N$$

于是可以写出完全数据的似然函数

$$\begin{aligned}
 P(y, \gamma | \theta) &= \prod_{j=1}^N P(y_j, \gamma_{j1}, \gamma_{j2}, \dots, \gamma_{jK} | \theta) \\
 &= \prod_{k=1}^K \prod_{j=1}^N [\alpha_k \phi(y | \theta_k)]^{\gamma_{jk}} \\
 &= \prod_{k=1}^K \alpha_k^{n_k} \prod_{j=1}^N [\phi(y | \theta_k)]^{\gamma_{jk}} \\
 &= \prod_{k=1}^K \alpha_k^{n_k} \prod_{j=1}^N \left[\frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(y - \mu_k)^2}{2\sigma_k^2}\right) \right]^{\gamma_{jk}}
 \end{aligned}$$

式中, $n_k = \sum_{j=1}^N \gamma_{jk}$, $\sum_{k=1}^K n_k = N$ 。(意思是选择第 k 个高斯模型的次数)

那么, 完全数据的对数似然函数为

$$\log P(y, \gamma | \theta) = \sum_{k=1}^K \left\{ n_k \log \alpha_k + \sum_{j=1}^N \gamma_{jk} \left[\log \left(\frac{1}{\sqrt{2\pi}} \right) - \log \sigma_k - \frac{1}{2\sigma_k^2} (y_j - \mu_k)^2 \right] \right\}$$

2. EM 算法的 E 步: 确定 Q 函数

$$\begin{aligned}
 Q(\theta, \theta^{(i)}) &= E[\log P(y, \gamma | \theta) | y, \theta^{(i)}] \\
 &= E \left\{ \sum_{k=1}^K \left\{ n_k \log \alpha_k + \sum_{j=1}^N \gamma_{jk} \left[\log \left(\frac{1}{\sqrt{2\pi}} \right) - \log \sigma_k - \frac{1}{2\sigma_k^2} (y_j - \mu_k)^2 \right] \right\} \right\} \\
 &= \sum_{k=1}^K \left\{ \overbrace{n_k \log \alpha_k}^{\text{常数}} + \sum_{j=1}^N E(\gamma_{jk}) \left[\log \left(\frac{1}{\sqrt{2\pi}} \right) - \log \sigma_k - \frac{1}{2\sigma_k^2} (y_j - \mu_k)^2 \right] \right\}
 \end{aligned}$$

这里需要计算 $E(\gamma_{jk} | y, \theta)$, 记为 $\hat{\gamma}_{jk}$

$$\begin{aligned}
 \hat{\gamma}_{jk} &= E(\gamma_{jk} | y, \theta) = P(\gamma_{jk} = 1 | y, \theta) \text{ (二项分布)} \\
 &= \frac{P(\gamma_{jk} = 1 | y, \theta)}{\sum_{k=1}^K P(\gamma_{jk} = 1 | y, \theta)} \\
 &= \frac{P(y_j | \gamma_{jk} = 1, \theta) P(\gamma_{jk} = 1 | \theta)}{\sum_{k=1}^K P(y_j | \gamma_{jk} = 1, \theta) P(\gamma_{jk} = 1 | \theta)} \\
 &= \frac{\alpha_k \phi(y_j | \theta_k)}{\sum_{k=1}^K \alpha_k \phi(y_j | \theta_k)}, \quad j = 1, 2, \dots, N; k = 1, 2, \dots, K
 \end{aligned}$$

$\hat{\gamma}_{jk}$ 是在当前模型参数下第 j 个观测数据来自第 k 个分模型的概率, 称为分模型 k 对观测数据 y_j 的响应度。代入式中得

$$Q(\theta, \theta^{(i)}) = \sum_{k=1}^K \left\{ n_k \log \alpha_k + \sum_{j=1}^N \hat{\gamma}_{jk} \left[\log \left(\frac{1}{\sqrt{2\pi}} \right) - \log \sigma_k - \frac{1}{2\sigma_k^2} (y_j - \mu_k)^2 \right] \right\} \quad (10.16)$$

3. 确定 EM 算法的 M 步迭代的 M 步是求函数 $Q(\theta, \theta^{(i)})$ 对 θ 的极大值, 即求新一轮迭代的模型参数:

$$\theta^{(i+1)} = \arg \max_{\theta} Q(\theta, \theta^{(i)})$$

分别对 μ_k, σ_k^2 求偏导并令其为 0, 即可得到 $\hat{\mu}_k, \hat{\sigma}_k$ 。求 $\hat{\alpha}_k$ 是在 $\sum_{k=1}^K = 1$ 条件下求偏导并令其为 0 得到。

$$\hat{\mu}_k = \frac{\sum_{j=1}^N \hat{\gamma}_{jk} y_j}{\sum_{j=1}^N \hat{\gamma}_{jk}}, \quad k = 1, 2, \dots, K \quad (10.17)$$

$$\hat{\sigma}_k^2 = \frac{\sum_{j=1}^N \hat{\gamma}_{jk} (y_j - \mu_k)^2}{\sum_{j=1}^N \hat{\gamma}_{jk}}, \quad k = 1, 2, \dots, K \quad (10.18)$$

$$\hat{\alpha}_k = \frac{n_k}{N} = \frac{\sum_{j=1}^N \hat{\gamma}_{jk}}{N}, \quad k = 1, 2, \dots, K \quad (10.19)$$

$$(10.20)$$

4. 重复以上计算, 直到对数似然函数值不再有明显的变化为止。

现将估计高斯混合模型参数的 EM 算法总结如下

输入: 观测数据 y_1, y_2, \dots, y_N , 高斯混合模型;

输出: 高斯混合模型参数。

- (1) 取参数的初始值开始迭代
- (2) E 步: 依据当前模型参数, 计算分模型 k 对观测数据 y_j 的响应度

$$\hat{\gamma}_{jk} = \frac{\alpha_k \phi(y_j | \theta_k)}{\sum_{k=1}^K \alpha_k \phi(y_j | \theta_k)}, \quad j = 1, 2, \dots, N; k = 1, 2, \dots, K$$

(3) M 步: 计算新一轮迭代的模型参数

$$\hat{\mu}_k = \frac{\sum_{j=1}^N \hat{\gamma}_{jk} y_j}{\sum_{j=1}^N \hat{\gamma}_{jk}}, \quad k = 1, 2, \dots, K$$

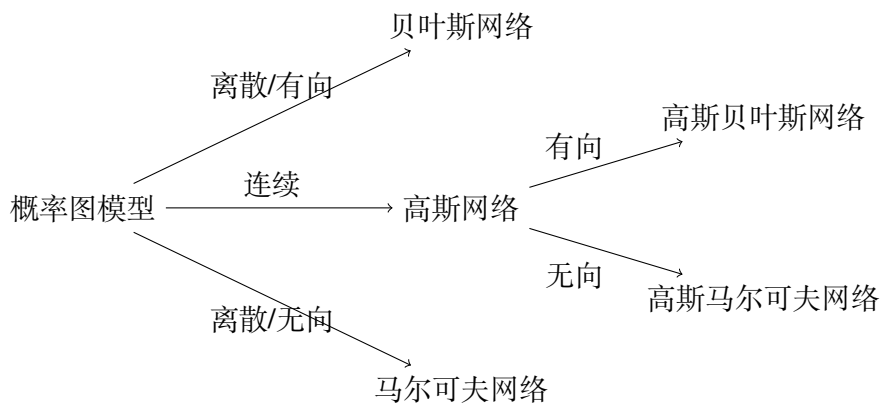
$$\hat{\sigma}_k^2 = \frac{\sum_{j=1}^N \hat{\gamma}_{jk} (y_j - \mu_k)^2}{\sum_{j=1}^N \hat{\gamma}_{jk}}, \quad k = 1, 2, \dots, K$$

$$\hat{\alpha}_k = \frac{n_k}{N} = \frac{\sum_{j=1}^N \hat{\gamma}_{jk}}{N}, \quad k = 1, 2, \dots, K$$

(4) 重复以上计算, 直到对数似然函数值不再有明显的变化为止。



第 11 章 概率图模型



11.1 贝叶斯网络

11.2 条件独立

11.3 马尔科夫随机场

11.4 图模型中的推断

11.5 隐马尔可夫模型

隐马尔可夫模型 (hidden Markov model, HMM) 是可用于标注问题的统计学习模型, 描述由隐藏的马尔可夫链随机生成观测序列的过程, 属于生成模型。本节首先介绍隐马尔可夫模型的基本概念, 然后分别叙述隐马尔可夫模型的概率计算算法、学习算法以及预测算法。隐马尔可夫模型在语音识别、自然语言处理、生物信息、模式识别等领域有着广泛的应用。

隐马尔可夫模型的基本概念

隐马尔可夫模型是关于时序的概率模型, 描述由一个隐藏的马尔可夫链随机生成不可观测的状态随机序列, 再由各个状态生成一个观测而产生观测随机序列的过程。隐藏的马尔可夫链随机生成的状态的序列, 称为状态序列 (state sequence); 每个状态生成一个观测, 而由此产生的观测的随机序列, 称为观测序列 (observation sequence)。序列的每一个位置又可以看作是一个时刻。隐马尔可夫模型由初始概率分布、状态转移概率分布以及观测概率分布确定。隐马尔可夫模型定义如下:

设 Q 是所有可能的状态的集合, V 是所有可能的观测的集合。

$$Q = \{q_1, q_2, \dots, q_N\}, \quad V = \{v_1, v_2, \dots, v_M\} \quad (11.1)$$

设 N 是所有可能的状态数, M 是可能的观测数。

I 是长度为 T 的状态序列, O 是对应的观测序列。

$$I = \{i_1, i_2, \dots, i_T\}, \quad O = \{o_1, o_2, \dots, o_T\} \quad (11.2)$$

A 是状态转移概率矩阵:

$$A = \left[a_{ij} \right]_{N \times N} \quad (11.3)$$

其中,

$$a_{ij} = P(i_{t+1} = q_j | i_t = q_i), i = 1, 2, \dots, N; j = 1, 2, \dots, N \quad (11.4)$$

是在时刻 t 处于状态 q_i 的条件下在时刻 $t+1$ 转移到状态 q_j 的概率。

B 是观测概率矩阵:

$$B = \left[b_j(k) \right]_{N \times M} \quad (11.5)$$

其中,

$$b_j(k) = P(o_t = v_k | i_t = q_j), k = 1, 2, \dots, M; j = 1, 2, \dots, N \quad (11.6)$$

是在时刻 t 处于状态 q_j 的条件下生成观测 v_k 的概率。 π 是初始状态概率向量:

$$\pi = (\pi_i) \quad (11.7)$$

其中,

$$\pi_i = P(i_1 = q_i), i = 1, 2, \dots, N \quad (11.8)$$

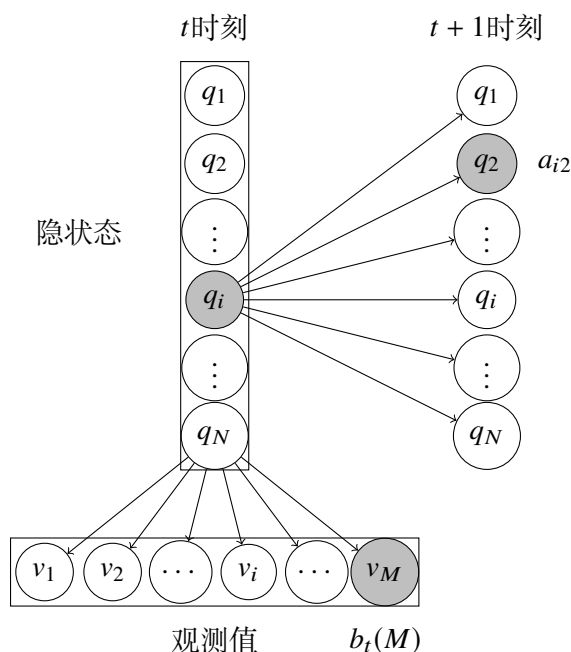
是时刻 $t=1$ 处于状态 q_i 的概率。

隐马尔可夫模型由初始状态概率向量 π 、状态转移概率矩阵 A 和观测概率矩阵 B 决定。 π 和 A 决定状态序列, B 决定观测序列。因此, 隐马尔可夫模型 λ 可以用三元符号表示, 即

$$\lambda = (A, B, \pi) \quad (11.9)$$

A, B, π 称为隐马尔可夫模型的三要素。从定义可知, 隐马尔可夫模型作了两个基本假设:

1. 齐次马尔可夫性假设, 即假设隐藏的马尔可夫链在任意时刻 t 的状态只依赖于其前一时刻的状态, 与其他时刻的状态及观测无关, 也与时刻 t 无关。
2. 观测独立性假设, 即假设任意时刻的观测只依赖于该时刻的马尔可夫链的状态, 与其他观测及状态无关。



隐马尔可夫模型的3个基本问题

1. 概率计算问题。给定模型 $\lambda = (A, B, \pi)$ 和观测序列 $O = (o_1, o_2, \dots, o_T)$, 计算在模型 λ 下观测序列 O 出现的概率 $P(O|\lambda)$
2. 学习问题。已知观测序列 $O = (o_1, o_2, \dots, o_T)$, 估计模型 $\lambda = (A, B, \pi)$ 参数, 使得在该模型下观测序列概率 $P(O|\lambda)$ 最大。即用极大似然估计的方法估计参数。
3. 预测问题。也称为解码 (decoding) 问题。已知模型 $\lambda = (A, B, \pi)$ 和观测序列 $O = (o_1, o_2, \dots, o_T)$, 求对给定观测序列条件概率 $P(I|\lambda)$ 最大的状态序列 $I = (i_1, i_2, \dots, i_T)$ 。即给定观测序列, 求最有可能的对应的状态序列。

问题一: 概率计算算法

直接计算法

给定模型 $\lambda = (A, B, \pi)$ 和观测序列 $O = (o_1, o_2, \dots, o_T)$, 计算在模型 λ 下观测序列 O 出现的概率 $P(O|\lambda)$ 。最直接的方法是按概率公式直接计算。

$$\begin{aligned}
 P(O|\lambda) &= \sum_I P(O|I, \lambda) P(I|\lambda) \\
 &= \sum_{i_1, i_2, \dots, i_T} \pi_{i_1} b_{i_1}(o_1) a_{i_1 i_2} \dots a_{i_{T-1} i_T} b_{i_T}(o_T)
 \end{aligned} \tag{11.10}$$

利用公式计算量大, 这种算法不可行。

前向算法

首先定义前现概率

定义 11.1. 前向概率

给定隐马尔可夫模型 λ , 定义到时刻 t 部分观测序列为 $O = (o_1, o_2, \dots, o_t)$ 且状态为 q_i 的概率为前向概率, 记作

$$\alpha_t(i) = P(o_1, o_2, \dots, o_t, i_t = q_i | \lambda) \quad (11.11)$$

可以递推地求得前向概率 $\alpha_t(i)$ 及观测序列概率 $P(O|\lambda)$

输入: 隐马尔可夫模型 λ , 观测序列 O ;

输出: 观测序列概率 $P(O|\lambda)$

(1) 初值

$$\alpha_1(i) = \pi_i b_i(o_1), \quad i = 1, 2, \dots, N \quad (11.12)$$

(2) 递推, 对 $t = 1, 2, \dots, T-1$

$$\alpha_{t+1}(i) = \left[\sum_{j=1}^N \alpha_t(j) a_{ji} \right] b_i(o_{t+1}) \quad (11.13)$$

(3) 终止

$$P(O|\lambda) = \sum_{i=1}^N \alpha_T(i) \quad (11.14)$$

前向算法实际是基于“状态序列的路径结构”递推计算 $P(O|\lambda)$ 的算法。前身算法的高效的关键是其局部计算前向概率, 然后利用路径结构将前向概率“递推”到全局, 得到 $P(O|\lambda)$ 。

后向概率**定义 11.2. 后向算法**

给定隐马尔可夫模型 λ , 定义在时刻 t 状态为 q_i 的条件下, 从 $t+1$ 到 T 的部分观测序列为 $o_{t+1}, o_{t+2}, \dots, o_T$ 的概率为后向概率, 记作

$$\beta_t(i) = P(o_{t+1}, o_{t+2}, \dots, o_T | i_t = q_i, \lambda) \quad (11.15)$$

可以用递推的方法求得后向概率 $\beta_t(i)$ 及观测序列概率 $P(O|\lambda)$

输入: 隐马尔可夫模型 λ , 观测序列 O ;

输出: 观测序列概率 $P(O|\lambda)$

(1) 初值

$$\beta_T(i) = 1, \quad i = 1, 2, \dots, N \quad (11.16)$$

(2) 递推, 对 $t = T-1, T-2, \dots, 1$

$$\beta_t(i) = \sum_{j=1}^N a_{ij} b_j(o_{t+1}) \beta_{t+1}(j), \quad i = 1, 2, \dots, N \quad (11.17)$$

(3) 终止

$$P(O|\lambda) = \sum_{i=1}^N \pi_i b_i(o_1) \beta_1(i) \quad (11.18)$$

利用前向概率和后向概率的定义可以将观测序列概率 $P(O|\lambda)$ 统一写成

$$P(O|\lambda) = \sum_{i=1}^N \sum_{j=1}^N \alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j), \quad t = 1, 2, \dots, T-1 \quad (11.19)$$

一些概率与期望值的计算

1. 给定模型 λ 和观测 O , 在时刻 t 处于状态 q_j 的概率。记 $\gamma_t(i)$
2. 给定模型 λ 和观测 O , 在时刻 t 处于状态 q_j 且在时刻 $t+1$ 处于状态 q_j 的概率。记 $\xi_t(i, j)$
3. 一些有用的期望值

问题二:学习算法

隐马尔可夫模型的学习, 根据训练数据是包括观测序列和对应的状态序列还是只有观测序列, 可以分别由监督学习与非监督学习实现。

监督学习算法

假设训练数据包含 S 个长度相同的观测序列和对应的状态序列 $\{(O_1, I_1), \dots, (O_S, I_S)\}$, 那么可以利用**极大似然法**来估计隐马尔可夫模型的参数。具体方法如下。

1. 转移概率 a_{ij} 的估计

设样本中时刻 t 处于状态 i 时刻 $t+1$ 转移到状态 j 的频数为 A_{ij} , 那么状态转移概率 a_{ij} 的估计是

$$\hat{a}_{ij} = \frac{A_{ij}}{\sum_{j=1}^N A_{ij}}, \quad i = 1, 2, \dots, N; j = 1, 2, \dots, N \quad (11.20)$$

2. 观测概率 $b_j(k)$ 的估计

设样本中状态为 j 并观测为 k 的频数是 B_{jk} , 那么状态为 j 观测为 k 的概率 $b_j(k)$ 的估计是

$$\hat{b}_j(k) = \frac{B_{jk}}{\sum_{k=1}^M B_{jk}}, \quad j = 1, 2, \dots, k = 1, 2, \dots, M \quad (11.21)$$

3. 初始状态概率 π_i 的估计 $\hat{\pi}_i$ 为 S 个样本中初始状态为 q_i 的频率

Baum-Welch 算法

假设训练数据包含 S 个长度相同的观测序列 $\{(O_1, I_1), \dots, (O_S, I_S)\}$ 而没有对应的状态序列, 目标是学习隐马尔可夫模型 $\lambda = (A, B, \pi)$ 的参数。我们将观测序列数据看作观测数据 O , 状态序列数据看作不可观测的隐数据 I , 那么隐马尔可夫模型事实上是一个含有隐变量的概率模型

$$P(O|\lambda) = \sum_I P(O|I, \lambda)P(I|\lambda) \quad (11.22)$$

它的参数学习可以由 EM 算法实现。

1. 确定完全数据的对数似然函数

所有观测数据写成 $O = (o_1, o_2, \dots, o_T)$, 所有隐数据写成 $I = (i_1, i_2, \dots, i_T)$ 完全数据是 $(O, I) = (o_1, o_2, \dots, o_T, i_1, i_2, \dots, i_T)$ 。完全数据的对数似然函数是 $\log P(O, I|\lambda)$

2. EM 算法的 E 步: 求 Q 函数 $Q(\lambda, \bar{\lambda})$

$$E_I[\log P(O, I|\lambda)|O, \bar{\lambda}] = \sum_I \log P(O, I|\lambda)P(I|O, \bar{\lambda}) \quad (11.23)$$

$$Q(\lambda, \bar{\lambda}) = \sum_I \log P(O, I|\lambda)P(O, I|\bar{\lambda}) \quad (11.24)$$

$P(I|O, \bar{\lambda}) = P(I, O|\bar{\lambda})/P(O|\bar{\lambda})$ 省略了对 λ 而言的常数因子。

$$P(O, I|\lambda) = \pi_{i_1} b_{i_1}(o_1) a_{i_1 i_2} \dots a_{i_{T-1} i_T} b_{i_T}(o_T) \quad (11.25)$$

于是函数 $Q(\lambda, \bar{\lambda})$ 可以写成:

$$\begin{aligned} Q(\lambda, \bar{\lambda}) &= \sum_I \log \pi_{i_1} P(O, I|\bar{\lambda}) \\ &\quad + \sum_I \left(\sum_{t=1}^{T-1} \log a_{i_t i_{t+1}} \right) P(O, I|\bar{\lambda}) \\ &\quad + \sum_I \left(\sum_{t=1}^T \log b_{i_t}(o_t) \right) P(O, I|\bar{\lambda}) \end{aligned} \quad (11.26)$$

3. EM 算法的 M 步: 极大化 $Q(\lambda, \bar{\lambda})$ 求模型参数 A, B, π

由于要极大化的参数在式 11.26 中单独地出现在 3 个项中, 所以只需对各项分别极大化。注意到

$$\sum_{i=1}^N \pi_i = 1, \sum_{j=1}^N a_{ij} = 1, \sum_{k=1}^M b_j(k) = 1,$$

利用拉格朗日乘子法,求得

$$\pi_i = \frac{P(O, i_1 = i | \bar{\lambda})}{P(O | \bar{\lambda})} = \gamma_1(i) \quad (11.27)$$

$$a_{ij} = \frac{\sum_{t=1}^{T-1} P(O, i_t = i, i_{t+1} = j | \bar{\lambda})}{\sum_{t=1}^{T-1} P(O, i_t = i | \bar{\lambda})} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \quad (11.28)$$

$$b_j(k) = \frac{\sum_{t=1}^T P(O, i_t = j | \bar{\lambda}) I(o_t = v_k)}{\sum_{t=1}^T P(O, i_t = j | \bar{\lambda})} = \frac{\sum_{t=1, o_t=v_k}^T \gamma_t(j)}{\sum_{t=1}^T \gamma_t(j)} \quad (11.29)$$

问题三:预测算法

近似算法

近似算法的想法是,在每个时刻 t 选择在该时刻最有可能出现的状态 i_t^* ,从而得到一个状态序列 $I^* = (i_1^*, i_2^*, \dots, i_T^*)$,将它作为预测的结果。近似算法的优点是计算简单,其缺点是不能保证预测的状态序列整体是最有可能的状态序列,因为预测的状态序列可能有实际不发生的部分。

维特比算法

维特比算法实际是用动态规划解隐马尔可夫模型预测问题,即用动态规划 (dynamic programming) 求概率最大路径 (最优路径)。这时一条路径对应着一个状态序列。

首先导入两个变量 δ 和 ψ 。定义在时刻 t 状态为 i 的所有单个路径 i_1, i_2, \dots, i_t 中概率最大值为

$$\begin{aligned} \delta_{t+1} &= \max_{i_1, i_2, \dots, i_t} P(i_t = i, i_{t-1}, \dots, i_1, o_{t+1}, \dots, o_1 | \lambda) \\ &= \max_{1 \leq j \leq N} [\delta_t(j) a_{ji}] b_i(o_{t+1}), \quad i = 1, 2, \dots, N \end{aligned} \quad (11.30)$$

定义在时刻 t 状态为 i 的所有单个路径 i_1, i_2, \dots, i_t 中概率最大的路径的第 $t-1$ 个结点为

$$\psi_t(i) = \arg \max_{1 \leq j \leq N} [\delta_{t-1}(j) a_{ji}], \quad i = 1, 2, \dots, N \quad (11.31)$$

(维特比算法)

输入:模型 $\lambda = (A, B, \pi)$, 观测序列 $O = (o_1, o_2, \dots, o_T)$;

输出:最优路径 $I^* = (i_1^*, i_2^*, \dots, i_T^*)$

1. 初始化

$$\begin{aligned} \delta_1 &= \pi_i b_i(o_1), \\ \psi_1(i) &= 0, \quad i = 1, 2, \dots, N \end{aligned}$$

2. 递推。对 $t = 2, 3, \dots, T$

$$\delta_{t+1} = \max_{1 \leq j \leq N} [\delta_t(j) a_{ji}] b_i(o_{t+1})$$

$$\psi_t(i) = \arg \max_{1 \leq j \leq N} [\delta_{t-1}(j) a_{ji}], \quad i = 1, 2, \dots, N$$

3. 终止

$$P^* = \max_{1 \leq i \leq N} \delta_T(i)$$

$$i_T^* = \arg \max_{1 \leq i \leq N} \delta_T(i)$$

4. 最优路径回溯。对 $t = T - 1, T - 2, \dots, 1$

$$i_t^* = \psi_{t+1}(i_{t+1}^*)$$

求得最佳路径 $I^* = (i_1^*, i_2^*, \dots, i_T^*)$

11.6 条件随机场

第 12 章 采样方法



12.1 基本采样算法

12.2 马尔科夫链蒙特卡罗

12.3 吉布斯采样

12.4 切片采样

12.5 混合蒙特卡罗算法

12.6 估计划分函数

第 13 章 近似推断

13.1 变分推断

13.2 变分线性回归

13.3 指数族分布

13.4 局部变分方法

13.5 变分 logistic 回归

13.6 期望传播

第 14 章 连续潜在变量

14.1 主成分分析

主成分分析,或者称为 PCA,是一种被广泛使用的技术,应用的领域包括维度降低、有损数据压缩、特征抽取「数据可视化。它也被称为 Karhunen-Loeve 变换。

有两种经常使用的 PCA 的定义,它们会给出同样的算法。PCA 可以被定义为数据在低维线性空间上的正交投影,这个线性空间被称为主空间 (principal subspace),使得投影数据的方差被最大化。等价地,它也可以被定义为使得平均投影代价最小的线性投影。平均投影代价是指数据点和它们的投影之间的平均平方距离。

考察一组观测数据集 $\{x_n\}$, 其中 $n = 1, \dots, N$, 因此 x_n 是一个 D 维欧几里德空间中的变量。我们的目标是将数据投影到维度 $M < D$ 的空间中,同时最大化投影数据的方差。

样本集合

$$X = (x_1 \ x_2 \ \dots \ x_N)_{N \times p}^T = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \dots & \dots & \dots & \dots \\ x_{N1} & x_{N2} & \dots & x_{Np} \end{pmatrix}_{N \times p} \quad (14.1)$$

样本均值

$$\bar{x} = \frac{1}{N} \sum_{n=1}^N x_n = \frac{1}{N} (x_1 \ x_2 \ \dots \ x_N) \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}_{N \times 1} = \frac{1}{N} X^T I_N \quad (14.2)$$

样本方差

$$\begin{aligned} S &= \frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})(x_n - \bar{x})^T \\ &= \frac{1}{N} (x_1 - \bar{x} \ x_2 - \bar{x} \ \dots \ x_N - \bar{x}) \begin{pmatrix} (x_1 - \bar{x})^T \\ (x_2 - \bar{x})^T \\ \vdots \\ (x_N - \bar{x})^T \end{pmatrix} \\ &= \frac{1}{N} X^T \overbrace{(I_N - \frac{1}{N} I_N I_N^T)}^{H_N - \text{中心矩阵}} \cdot (I_N - \frac{1}{N} I_N I_N^T)^T \cdot X \\ &= \frac{1}{N} X^T H X \end{aligned} \quad (14.3)$$

中心矩阵 H 有以下良好的性质

$$H = I_N - \frac{1}{N} I_N I_N^T \quad (14.4)$$

$$H^T = H \quad (14.5)$$

$$H^n = H \quad (14.6)$$

最大方差形式

首先,考虑在一维空间 ($M = 1$) 上的投影。我们可心使用 D 维向量 u_1 定义这个空间的方向。不失一般性,我们假定选择一个单位向量,从而 $u_1^T u_1 = 1$ (注意,我们只对 u_1 的方向感兴趣,而对 u_1 本身的大小不感兴趣)。这样,每个数据点 x_n 被投影到一个标题值 $x_1^T x_n$ 上。投影数据的均值是 $u_1^T \bar{x}$ 。投影数据的方差为

$$\begin{aligned} J &= \frac{1}{N} \sum_{n=1}^N \{u_1^T x_n - u_1^T \bar{x}\}^2 \\ &= u_1^T \underbrace{\sum_{n=1}^N \frac{1}{N} (x_n - \bar{x}) \cdot (x_n - \bar{x})^T}_{S} u_1 \\ &= u_1^T S u_1 \end{aligned} \quad (14.7)$$

S 是数据的协方差矩阵,优化问题变成

$$\begin{aligned} \hat{u}_1 &= \arg \max u_1^T S u_1 \\ s.t. \quad &u_1^T u_1 = 1 \end{aligned} \quad (14.8)$$

利用拉格朗日乘子法,求偏导并令其为零,我们看到驻点满足

$$S u_1 = \lambda_1 u_1 \quad (14.9)$$

表明 u_1 一定是 S 的一个特征向量,这个特征向量被称为第一主成分。

我们可以用一种增量的方式定义额外的主成分,方法为:在所有与那些已经考虑过的方向正交的所有可能的方向中,将新的方向选择为最大化投影方差的方向。

总结一下,主成分分析涉及到计算数据集的均值 \bar{x} 和协方差矩阵 S ,然后寻找 S 的对应于 M 个最大特征值的 M 个特征向量。

最小误差形式

引入 D 维基向量的一个完整的单位正交集 $\{u_i\}$,其中 $i = 1, \dots, D$,且满足

$$u_i^T u_j = \delta_{ij} \quad (14.10)$$

由于基是完整的,因此每个数据点可以精确地表示为基向量的一个纯性组合,即

$$x_n = \sum_{i=1}^D a_{ni} u_i \quad (14.11)$$

其中,系数 a_{ni} 对于不同的数据点来说是不同的。这对应于将坐标系旋转到了一个由 $\{u_i\}$ 定义的新坐标系,原始的 D 个分量 $\{x_{n1}, \dots, x_{nD}\}$ 被替换为一个等价的集合 $\{a_{n1}, \dots, a_{nD}\}$ 。我们有 $a_{nj} = x_n^T u_j$, 因此不失一般性

$$x_n = \sum_{i=1}^D (x_n^T u_i) u_i \quad (14.12)$$

然而,我们的目标是使用限定数量 $M < D$ 个变量的一种表示方法来近似数据点,这对应于在低维子空间上的一个投影。不失一般性, M 维线性子空间可以用前 M 个基向量表示,因此我们可以用下式来近似每个数据点 x_n

$$\tilde{x}_n = \sum_{i=1}^M z_{ni} u_i + \sum_{i=M+1}^D b_i u_i \quad (14.13)$$

其中 $\{z_{ni}\}$ 依赖于特定的数据点,而 $\{b_i\}$ 是常数,对于所有数据点都相同。我们可以任意选择 $\{u_i\}, \{z_{ni}\}, \{b_i\}$, 从而最小化由维度降低所引入的失真。作为失真的度量,我们使用原始数据点与它的近似点 \tilde{x}_n 之间的平方距离,在数据集上取平均。因此我们的目标是最小化

$$J = \frac{1}{N} \sum_{n=1}^N \|x_n - \tilde{x}_n\|^2 = \frac{1}{N} \sum_{n=1}^N \sum_{i=M+1}^D (x_n^T u_i - \bar{x}^T u_i)^2 = \sum_{i=M+1}^D u_i^T S u_i \quad (14.14)$$

剩下的任务是关于 $\{u_i\}$ 对 J 进行最小化。同样利用拉格朗日乘子法能够得到

$$S u_i = \lambda_i u_i \quad (14.15)$$

PCA 的应用

14.2 概率 PCA

前一节讨论的 PCA 的形式所基于的是将数据线性投影到比原始数据空间维度更低的子空间内。PCA 也可以被视为概率潜在变量模型的最大似然解。PCA 的这种形式,被称为概率 PCA(probabilistic PCA),与传统的 PCA 相比,会带来如下几个优势。

- 概率 PCA 表示高斯分布的一个限制形式,其中自由参数的数量可以受到限制,同时仍然使得模型能够描述数据集的主要的相关关系。
- 我们可以为 PCA 推导一个 EM 算法,这个算法在只有几个主要的特征向量需要求出的情况下,计算效率比较高,并且避免了计算数据协方差的中间步骤。

- 概率模型与 EM 的结合使得我们能够处理数据集里缺失值的问题。
- 概率 PCA 混合模型可以用一种有理有据的方式进行形式化, 并且可以使用 EM 算法进行训练。
- 概率 PCA 构成了 PCA 的贝叶斯方法的基础, 其中主子空间的维度可以自动从数据中找到。
- 似然函数的存在使得直接与其他概率密度模型进行对比成为可能。相反, 传统的 PCA 会给接近主子空间的数据点分配一个较低的重建代价, 即使这些数据点的位置距离训练数据任意远。
- 概率 PCA 可以被用来对类条件概率密度建模, 因此可以应用于分类问题。
- 概率 PCA 模型可以用一种生成式的方式运行, 从而可以按照某个概率分布生成样本。

最大似然 PCA

用于 PCA 的 EM 算法

14.3 核 PCA

14.4 非线性隐变量模型