

# 机器学习笔记

Notes on Machine Learning

J.R.Tsien

[jade.ray.tsien@gmail.com](mailto:jade.ray.tsien@gmail.com)

# 目 录

第1章 绪论 .....	1
1.1 最优化问题 .....	1
1.1.1 原始问题 .....	1
1.1.2 拉格朗日对偶问题 .....	1
1.2 梯度下降法 (gradient descent) .....	1
1.3 牛顿法和拟牛顿法 .....	1
1.3.1 牛顿法 (Newton Method) .....	2
1.3.2 拟牛顿法 (Quasi Newton Method) .....	2
1.3.2.1 拟牛顿条件 .....	3
1.3.2.2 DFP算法 .....	3
1.3.2.3 BFGS算法 .....	4
1.4 机器学习概论 .....	5
1.4.1 生成模型与判别模型 .....	5
第2章 回归 .....	6
2.1 线性回归 .....	6
2.1.1 直接求解 .....	6
2.1.2 牛顿法 (Newton's Method) .....	7
2.1.3 批处理梯度下降法 (batch gradient descent) .....	8
2.1.4 随机梯度下降法 (stochastic gradient descent) .....	8

2.2 局部加权线性回归 (LWR)	8
2.3 Logistic回归	9
2.3.1 二项逻辑斯谛回归	9
2.3.2 多项逻辑斯谛回归	10
2.4 广义线性模型	11
2.4.1 指数分布族 (the exponential family)	11
2.4.2 构造GLMs	12
2.4.2.1 最小二乘法 (Ordinary Least Square)	12
2.4.2.2 Logistic Regression	13
2.4.2.3 Softmax Regression	13
2.4.3 总结	13
第3章 生成算法	14
3.1 高斯判别分析 (Gaussian Discriminant Analysis)	14
3.1.1 多元正态分布 (Multivariate Normal Distribution)	14
3.1.2 高斯判别分析 (GDA)	14
3.1.3 高斯判别分析与logistic回归的对比	15
3.2 朴素贝叶斯法	16
3.2.1 贝叶斯公式	16
3.2.2 朴素贝叶斯法	17
3.2.2.1 模型	17
3.2.2.2 实现	18
第4章 决策树	19
第5章 最大熵	20
第6章 支持向量机	21
第7章 提升方法	22

第8章 EM方法 .....	23
8.1 Jensen不等式 .....	23
第9章 隐马尔可夫模型 .....	24
9.1 模型 .....	24
9.2 概率计算问题 .....	24
9.2.1 直接计算 .....	24
9.2.2 前向算法 (forward algorithm) .....	24
9.2.3 后向算法 (backward algorithm) .....	24
9.3 学习问题 .....	24
9.4 预测问题 .....	24
第10章 附录 .....	25
10.1 矩阵分析 (matrix analysis) .....	25
10.1.1 迹(Trace)和导数(matrix derivatives) .....	25
10.2 常用不等式 .....	26
10.2.1 柯西不等式 (Cauchy Inequality) .....	26
10.2.2 赫尔德不等式 (Hölder Inequality) .....	27
10.2.3 闵可夫斯基不等式 (Minkowski Inequality) .....	27
参考文献 .....	28



# 第1章 绪论

## § 1.1 最优化问题

### 1.1.1 原始问题

### 1.1.2 拉格朗日对偶问题

## § 1.2 梯度下降法 (gradient descent)

梯度下降法或者最速下降法 (steepest descent) 是求解无约束最优化问题的方法。特点是实现起来比较简单。其原理是如果函数  $f(x)$  在点  $a$  处可微且有定义，那么函数  $f(x)$  在  $a$  点沿着梯度的反方向，即  $-\nabla f(a)$ ，下降最快。

所以，可以从一个初始值  $x_0$  出发，沿梯度反方向迭代的更新解。如下

$$x_{n+1} = x_n - \alpha \nabla f(x_n)$$

直到  $x_n$  的值不再发生变化，或者变化很小，此时， $x_n$  等于或者接近  $f(x)$  的极小值。 $\alpha$  称为学习率 (learning rate)。 $\alpha$  值过大，可能会在最小值附近振荡。 $\alpha$  值过小，可能学习的时间比较长。同时， $\alpha$  值的选取可以是预先设定的固定值，也可以是根据解更新的情况变化的值。

梯度下降法的一个问题在于，能否得到最优解取决于初始值的选取。

## § 1.3 牛顿法和拟牛顿法

### 1.3.1 牛顿法 (Newton Method)

牛顿法，或牛顿-拉夫逊法 (Newton-Raphson Method) 也是求解无约束优化问题的常用方法。牛顿法是二阶收敛的算法（不仅考虑梯度方向，同时考虑梯度的梯度），而梯度下降法是一阶收敛的，因此牛顿法的收敛速度比梯度下降快。换句话说，牛顿法用二次曲面来拟合当前所在位置的局部曲面，然后按照曲率最大的方向下降。而梯度法是用一个平面去拟合局部曲面，然后按照平面的法向量的方向下降。但是一次迭代的代价比较高，因为需要计算矩阵的逆。

考虑无约束的最优化问题

$$\min_{x \in \mathbb{R}^n} f(x)$$

假设 $f(x)$ 有二阶连续偏导数，且设第 $k$ 次迭代的解为 $x_k$ ，将 $f(x)$ 在点 $x_k$ 处进行二阶泰勒展开

$$f(x) = f(x_k) + (x - x_k)f'(x_k) + \frac{1}{2}(x - x_k)^2 f''(x_k)$$

函数 $f(x)$ 在下次迭代点 $x_{k+1}$ 处取得极值的必要条件是 $f'(x_{k+1}) = 0$ ，即

$$f'(x)|_{x_{k+1}} = f'(x_k) + (x_{k+1} - x_k)f''(x_k) = 0$$

解上式得到

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$$

迭代停止的条件可以设定为 $f'(x_k) < \epsilon$ 。

当 $x$ 是向量的时候，其一阶导数要修改成梯度的形式，二阶导数修改成其Hessian矩阵，即

$$f(x) = f(x_k) + (x - x_k)\nabla_x f(x_k) + \frac{1}{2}(x - x_k)^T H(x_k)(x - x_k)$$

当 $H(x_k)$ 是正定矩阵时， $f(x)$ 的极值为极小值。其更新公式是

$$x_{k+1} = x_k - H^{-1}(x_k)\nabla_x f(x_k)$$

迭代终止的条件 $\nabla_x f(x_k) < \epsilon$ 。

当初始点离极值点较远时，牛顿法可能不收敛，因为此时由Hessian矩阵的逆矩阵和梯度规定的牛顿方向不一定是下降方向。

### 1.3.2 拟牛顿法 (Quasi Newton Method)

牛顿法中Hessian矩阵求逆比较复杂，所以实际中会考虑用一个近似的正定对称矩阵来代替Hessian矩阵，这就是拟牛顿法。不同的替代方法形成了不同的拟牛顿法。

### 1.3.2.1 拟牛顿条件

首先将函数 $f(x)$ 在第 $k$ 次迭代点处泰勒展开

$$f(x) = f(x_k) + (x - x_k)\nabla f(x_k) + \frac{1}{2}(x - x_k)^T H(x_k)(x - x_k)$$

求其此时的梯度

$$\nabla f(x) = \nabla f(x_k) + (x - x_k)H(x_k)$$

令 $x = x_{k+1}$ ,

$$\nabla f(x_{k+1}) - \nabla f(x_k) = (x_{k+1} - x_k)H(x_k)$$

左边是梯度的变化, 右边 $(x_{k+1} - x_k)$ 是自变量的变化, 分别记为

$$g_k = \nabla f(x_{k+1}) - \nabla f(x_k)$$

$$\delta_k = x_{k+1} - x_k$$

得到拟牛顿条件

$$g_k = H_k \delta_k \quad (1.1)$$

或

$$H_k^{-1} g_k = \delta_k \quad (1.2)$$

拟牛顿法选择 $G_k \approx H_k^{-1}$ , 或者 $B_k \approx H_k$ 。而且, 尽量使得近似矩阵的更新方式为迭代更新

$$G_{k+1} = G_k + \Delta G_k$$

且 $G_k$ 满足拟牛顿条件

$$G_{k+1} g_k = \delta_k \quad (1.3)$$

一般令 $G_0 = I$ 为单位阵, 所以只需要找到 $\Delta G_k$ 即可。

### 1.3.2.2 DFP算法

DFP算法最早由William C. Davidon于1959年提出, 后由Roger Fletcher和Michael J.D. Powell发展和完善。DFP算法令校正矩阵为

$$\Delta G_k = \frac{\delta_k \delta_k^T}{\delta_k^T g_k} - \frac{G_k g_k g_k^T G_k}{g_k^T G_k g_k}$$

构造过程主要的规则在于保证如果初始矩阵 $G_0$ 正定, 那么每个 $G_k$ 都是正定。

DFP算法

输入: 目标函数 $f(x)$ , 梯度 $\nabla f(x)$ , 精度要求 $\epsilon$

输出:  $f(x)$ 的极小值点 $x^*$



- (1) 选定初始点 $x_0$ ，取 $G_0$ 为正定对称矩阵（单位阵），置 $k = 0$
- (2) 计算 $\nabla f(x_k)$ ，若 $\|\nabla f(x_k)\| < \epsilon$ ，令 $x^* = x_k$ ，停止计算
- (3) 令 $p_k = -G_k \nabla f(x_k)$
- (4) 一维搜索：求 $\lambda_k$ 使得

$$f(x_k + \lambda_k p_k) = \min_{\lambda \geq 0} f(x_k + \lambda p_k)$$

- (5) 令 $x_{k+1} = x_k + \lambda_k p_k$
- (6) 计算 $\nabla f(x_{k+1})$ ，若 $\|\nabla f(x_{k+1})\| < \epsilon$ ，令 $x^* = x_{k+1}$ ，停止计算；否则，按 $G_{k+1} = G_k + \Delta G_k$ 计算 $G_{k+1}$
- (7) 置 $k = k + 1$ ，转（3）

### 1.3.2.3 BFGS算法

BFGS算法（Broyden-Fletcher-Goldfarb-Shanno）是最流行的拟牛顿法。DFP用一个正定矩阵来近似 $H^{-1}$ ，BFGS用一个正定矩阵来近似 $H$ ，其表示如下

$$B_{k+1} = B_k + \Delta B_k \quad (1.4)$$

$$\Delta B_k = \frac{g_k g_k^T}{g_k^T \delta_k} - \frac{B_k \delta_k \delta_k^T B_k}{\delta_k^T B_k \delta_k} \quad (1.5)$$

构造的核心也是满足如果初始矩阵 $B_0$ 是正定的，那么迭代过程中的每一个 $B_k$ 都是正定的。

#### BFGS算法

输入：目标函数 $f(x)$ ，梯度 $\nabla f(x)$ ，精度 $\epsilon$

输出： $f(x)$ 的极小值点 $x^*$

- (1) 选定初始点 $x_0$ ，取 $B_0$ 为正定对称矩阵（单位阵），置 $k = 0$
- (2) 计算 $\nabla f(x_k)$ ，若 $\|\nabla f(x_k)\| < \epsilon$ ，令 $x^* = x_k$ ，停止计算
- (3) 由 $B_k p_k = -\nabla f(x_k)$ 求出 $p_k$
- (4) 一维搜索：求 $\lambda_k$ 使得

$$f(x_k + \lambda_k p_k) = \min_{\lambda \geq 0} f(x_k + \lambda p_k)$$

- (5) 令 $x_{k+1} = x_k + \lambda_k p_k$
- (6) 计算 $\nabla f(x_{k+1})$ ，若 $\|\nabla f(x_{k+1})\| < \epsilon$ ，令 $x^* = x_{k+1}$ ，停止计算；否则，按 $B_{k+1} = B_k + \Delta B_k$ 计算 $B_{k+1}$
- (7) 置 $k = k + 1$ ，转（3）

令 $G_k = B_k^{-1}$ ，对 $B_{k+1} = B_k + \Delta B_k$ 运用两次Sherman-Morrison公式有

$$G_{k+1} = \left(I - \frac{\delta_k g_k^T}{\delta_k^T g_k}\right) G_k \left(I - \frac{\delta_k g_k^T}{\delta_k^T g_k}\right)^T + \frac{\delta_k g_k^T}{\delta_k^T g_k}$$

称为BFGS算法关于 $G_k$ 的迭代公式。

Sherman-Morrison公式：假设 $A$ 是 $n$ 阶可逆矩阵， $u, v$ 是 $n$ 维向量，且 $A + uv^T$ 也是可逆矩阵，则

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1}uv^T A^{-1}}{1 + v^T A^{-1}u}$$

记DFP关于 $G_k$ 的迭代公式得到的 $G_{k+1}$ 为 $G^{DFP}$ ，由BFGS得到的记为 $G^{BFGS}$ ，它们都满足拟牛顿公式，且其线性组合

$$G_{k+1} = \alpha G^{DFP} + (1 - \alpha) G^{BFGS}$$

也满足拟牛顿条件，而且是正定是。其中 $0 \leq \alpha \leq 1$ 。这样得到的一类拟牛顿法称为Broyden类算法。

实际在使用的时候使用的是LBFGS算法（Limited-memory-BFGS）。有很成熟的开源实现，直接调用即可。

## § 1.4 机器学习概论

### 1.4.1 生成模型与判别模型

判别模型（discriminative model）是给定输入 $X$ ，要求预测的输出 $Y$ 的类型。其方法是通过训练数据来学习决策函数 $f(x)$ 或者条件概率分布 $p(y|x)$ 来构造预测的模型。

生成模型（generative model）是通过学习联合分布 $p(x, y)$ 或者 $p(x|y)$ 以及 $p(y)$ 来得到给定 $x$ 的 $y$ 的后验概率 $p(y|x) = \frac{p(x, y)}{p(y)} = \frac{p(x|y)p(y)}{p(y)}$ 。

## 第2章 回归

给定数据集

$$T = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\},$$

其中,  $x^{(i)} = (x_0, x_1, \dots, x_n) \in \mathcal{X} = \mathbb{R}^{n+1}$ ,  $y^{(i)} \in \mathcal{Y}$ ,  $i = 0, 2, \dots, m$ , 且  $x_0 = 1$  (表示截距, intercept term)。回归分析的任务是找出输入  $x$  与输出  $y$  之间的关系。

### § 2.1 线性回归

假设输入与输出之间满足的关系是线性的,  $\theta$  称为参数 (parameters) 或者权重 (weights)

$$y = h_{\theta}(x) = \theta^T x = \sum_{i=0}^n \theta_i x_i, \quad \theta \in \mathbb{R}^{n+1}$$

对于这个模型, 需要有一个损失函数 (cost function) 来表示其对训练数据的拟合程度

$$J(\theta) = \frac{1}{2} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2.$$

这个被称为最小二乘方效用函数 (least-square cost function)。

则这个问题的求解可以表述为如下的无约束最优化问题

$$\min_{\theta} J(\theta)$$

#### 2.1.1 直接求解

直接求  $J(\theta)$  对  $\theta$  的极值。首先定义设计矩阵 (design matrix)  $X$

$$X = \begin{bmatrix} (x^{(1)})^T & (x^{(2)})^T & \dots & (x^{(m)})^T \end{bmatrix}^T$$

$$Y = \begin{bmatrix} y^{(1)} & y^{(2)} & \dots & y^{(m)} \end{bmatrix}^T$$

则有

$$X\theta - Y = \begin{bmatrix} h_\theta(x^{(1)}) - y^{(1)} & \dots & h_\theta(x^{(m)}) - y^{(m)} \end{bmatrix}^T$$

考虑到  $z^T z = \sum_i z_i^2$ , 有

$$J(\theta) = \frac{1}{2}(X\theta - Y)^T(X\theta - Y) = \frac{1}{2} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})^2$$

使用公式  $\nabla_A \text{tr} ABA^T C = CAB + C^T AB^T$ , 且由于  $J(\theta)$  只是个实数, 所以

$$\begin{aligned} \nabla_\theta J(\theta) &= \nabla_\theta \text{tr} J(\theta) \\ &= X^T X\theta - X^T Y \end{aligned}$$

令  $\nabla_\theta J(\theta) = 0$  得到

$$\theta = (X^T X)^{-1} X^T Y$$

不过, 这个公式用来直接计算  $\theta$  不现实, 因为矩阵求逆比较麻烦, 同时可能会是数值不稳定的矩阵。

### 2.1.2 牛顿法 (Newton's Method)

已知最优化问题

$$\min_{\theta} J(\theta) = \frac{1}{2} \sum_{i=1}^m (h_\theta(x^{(i)}) - y^{(i)})^2$$

函数  $J(\theta)$  的 Hessian 矩阵是 (注意  $x^{(i)}$  和  $\theta$  都是列向量)

$$H(\theta) = \begin{bmatrix} \frac{\partial^2 J(\theta)}{\partial \theta_1 \partial \theta_1} & \frac{\partial^2 J(\theta)}{\partial \theta_1 \partial \theta_2} & \dots & \frac{\partial^2 J(\theta)}{\partial \theta_1 \partial \theta_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 J(\theta)}{\partial \theta_n \partial \theta_1} & \frac{\partial^2 J(\theta)}{\partial \theta_n \partial \theta_2} & \dots & \frac{\partial^2 J(\theta)}{\partial \theta_n \partial \theta_n} \end{bmatrix} = \sum_{i=1}^m x^{(i)} (x^{(i)})^T = X^T X$$

又,  $J(\theta) = \frac{1}{2}(X\theta - Y)^T(X\theta - Y)$ , 所以其梯度

$$\nabla_\theta J(\theta) = X^T X\theta - X^T Y$$

代入牛顿法的迭代公式

$$\theta_{n+1} = \theta_n - (X^T X)^{-1} (X^T X\theta_n - X^T Y) = (X^T X)^{-1} X^T Y$$

### 2.1.3 批处理梯度下降法 (batch gradient descent)

$J(\theta)$ 对 $\theta$ 的梯度

$$\frac{\partial J(\theta)}{\partial \theta} = \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})x^{(i)}$$

所以，更新公式为

$$\begin{aligned}\theta_{n+1} &= \theta_n - \alpha \frac{\partial J(\theta)}{\partial \theta} \\ &= \theta_n - \alpha \sum_{i=1}^m (h_{\theta_n}(x^{(i)}) - y^{(i)})x^{(i)}\end{aligned}$$

这个公式更新时每次都需要全部的训练数据集，所以称之为批处理梯度下降法。当数据集比较大时，进行一次更新就比较耗费时间。

### 2.1.4 随机梯度下降法 (stochastic gradient descent)

随机梯度下降法，也称增量梯度下降 (incremental gradient descent)，一次使用一条训练数据来更新参数。算法如下

$$\theta := \theta - \alpha (h_{\theta}(x^{(i)}) - y^{(i)})x^{(i)}, i = 1, 2, \dots, m$$

随机梯度下降法更新的速度比梯度下降法快，但可能收敛不到最优值，不过通过调节学习率可以使得算法得到较好的解。而且随机梯度下降法可以用作在线学习的算法。

## § 2.2 局部加权线性回归 (LWR)

线性回归的方法是参数学习算法 (parametric learning algorithm)，其参数的个数，即特征的个数是固定的，一旦算法学习完成，训练数据集就不再对参数产生影响。但是，当选取的参数过多时，可能存在过拟合问题，而当选取的参数过少时，存在欠拟合问题。局部加权线性回归 (local weighted linear regression, LWR) 是一种非参学习算法 (non-parametric learning algorithm)，其参数是随着预测点的不同而发生变化的，每有一个新的预测点，就需要整个训练数据集重新参与学习。所谓局部，是因为目标函数的逼近仅仅根据查询点附近的数据。所谓加权，是因为每个训练样例的贡献都是由它与查询点间的距离加权的。而回归是指数值逼近的方法。

线性回归的优化目标是

$$\min_{\theta} \sum_{i=1}^m (\theta^T x^{(i)} - y^{(i)})^2$$

而LWR的优化目标是在上述公式上增加一个距离乘法项

$$\min_{\theta} \sum_{i=1}^m w^{(i)} (\theta^T x^{(i)} - y^{(i)})^2$$

一个相对标准的权重选择是

$$w^{(i)} = \exp\left(-\frac{(x^{(i)} - x)^2}{2\tau^2}\right)$$

其中,  $x$ 是要预测的输入数据,  $\tau$ 称为带宽 (bandwidth)。权重项使得离输入数据 $x$ 越近的点影响越大。

非参数学习算法在局部预测能力上有时要比参数学习算法好, 但是缺点是每次做预测都要重新学习, 耗费时间空间。

## § 2.3 Logistic回归

### 2.3.1 二项逻辑斯谛回归

逻辑斯谛回归 (logistic regression) 是分类模型, 跟线性回归模型有相同的输入, 但是其输出是离散值。但是, 之所以叫“回归”是因为它依然采用线性回归的算法来预测 $x$ 与 $y$ 之间的关系。也因此, 需要限制模型中 $y$ 值的输出, 令

$$h_{\theta}(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}}$$

其中

$$g(z) = \frac{1}{e^{-z} + 1}$$

称为逻辑斯谛函数 (logistic function), 或者S形曲线 (sigmoid curve)。

为了推导方便, 首先给出 $g(z)$ 的导数

$$g'(z) = g(z)(1 - g(z))$$

设

$$P(y = 1|x; \theta) = h_{\theta}(x)$$

$$P(y = 0|x; \theta) = 1 - h_{\theta}(x)$$

或者简写成

$$p(y|x; \theta) = (h_\theta(x))^y (1 - h_\theta(x))^{1-y}$$

其对数似然函数

$$\begin{aligned} l(\theta) &= \log \prod_{i=1}^m p(y^{(i)}|x^{(i)}; \theta) = \sum_{i=1}^m (y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log(1 - h_\theta(x^{(i)}))) \\ &= \sum_{i=1}^m (y^{(i)} h_\theta(x^{(i)}) + \log(1 - h_\theta(x^{(i)}))) \end{aligned}$$

其中， $m$ 是训练样本的个数。对 $\theta$ 求导，得

$$\frac{\partial l(\theta)}{\partial \theta} = \sum_{i=1}^m (y^{(i)} - h_\theta(x^{(i)})) x^{(i)}$$

则对于单条训练样本的更新公式是

$$\theta := \theta + \alpha (y^{(i)} - h_\theta(x^{(i)})) x^{(i)}$$

注意，之所以是+号，是因为所求的最优化问题是似然函数的最大化，也即按照梯度的方向增加。

这个更新公式跟LMS（least mean squares）更新准则一样。也即，更新的幅度由学习率以及特征向量和误差的乘积决定。这是由梯度下降法所决定的。

### 2.3.2 多项逻辑斯谛回归

逻辑斯谛回归其实是建立了逻辑斯谛概率模型之后依据似然函数最大化准备建立的回归模型，其学习算法一般是梯度下降或者牛顿法。多项逻辑斯谛回归的概率模型可以表述为：如果离散随机变量 $Y$ 的取值集合是 $\{1, 2, \dots, K\}$ ，那么多项逻辑斯谛回归模型是

$$P(Y = k|x; \theta) = \frac{e^{\theta^T x}}{1 + \sum_{i=1}^{K-1} e^{\theta^T x}}, k = 1, 2, \dots, K-1 \quad (2.1)$$

$$P(Y = K|x; \theta) = \frac{1}{1 + \sum_{i=1}^{K-1} e^{\theta^T x}}, k = K \quad (2.2)$$

其中， $x, \theta \in \mathbb{R}^{n+1}$ 。

## § 2.4 广义线性模型

广义线性模型（Generalized Linear Models, GLMs）是线性回归（linear regression）和逻辑斯谛回归（logistic regression）的统一表述。

### 2.4.1 指数分布族（the exponential family）

首先定义指数分布族

$$p(y; \eta) = b(y)e^{\eta^T T(y) - a(\eta)} \quad (2.3)$$

其中， $\eta$ 称为分布的自然参数（natural parameter）或者标准参数（canonical parameter）， $\eta^T$ 是取 $\eta$ 的转置的意思，当 $\eta$ 是标量（scalar）的时候， $\eta^T = \eta$ ，当 $\eta$ 是向量（vector）的时候，上式就是向量的乘积， $T(y)$ 是充分统计量（sufficient statistic），通常令 $T(y) = y$ ， $a(\eta)$ 称为对数分割函数（log partition function）。注意到 $e^{-a(\eta)}$ 实际上起到的是归一化的作用，即保证分布 $p(y; \eta)$ 对 $y$ 的和或积分是1。固定 $T, a, b$ 就得到了以 $\eta$ 为参数的分布族。下面将伯努利分布和高斯分布表示成指数族的形式。

伯努利分布（Bernoulli distribution）的形式为

$$p(y; \phi) = \phi^y (1 - \phi)^{1-y} \quad (2.4)$$

表示的是一次Bernoulli实验（二值）的结果。 $y$ 表示实验成功与否：成功，则 $y = 1$ ；否则， $y = 0$ 。

将上式表示成

$$\begin{aligned} p(y; \phi) &= \phi^y (1 - \phi)^{1-y} \\ &= \exp(y \log \phi + (1 - y) \log(1 - \phi)) \\ &= \exp((\log(\frac{\phi}{1 - \phi}))y + \log(1 - \phi)) \end{aligned}$$

此时，令 $\eta = \log \frac{\phi}{1 - \phi}$ 就得到了Bernoulli分布的指数族形式。其中

$$T(y) = y$$

$$b(\eta) = 1$$

$$a(\eta) = -\log(1 - \phi) = \log(1 + e^\eta)$$

注意到 $\phi = \frac{1}{1 + e^{-\eta}}$ 刚好是sigmoid函数。



高斯分布（Gaussian distribution）的一般形式

$$p(y; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y - \mu)^2}{2\sigma^2}\right) \quad (2.5)$$

考虑到 $\sigma$ 对假设函数没有影响，所以可以令 $\sigma = 1$

$$p(y; \mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(y - \mu)^2\right) \quad (2.6)$$

表示为GMLs形式，有

$$\eta = \mu$$

$$T(y) = y$$

$$a(\eta) = \frac{\mu^2}{2} = \frac{\eta^2}{2}$$

$$b(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}}$$

如果要考虑 $\sigma$ 的话，需要用到GLMs的更为一般的模型

$$p(y; \eta, \tau) = b(a, \tau) \exp\left(\frac{\eta^T T(y) - a(\eta)}{c(\tau)}\right) \quad (2.7)$$

其中， $\eta = (\mu, \sigma) \in \mathbb{R}^2$ ， $\tau$ 称为离差参数（dispersion parameter），对于高斯分布来说， $c(\tau) = \sigma^2$ 。

## 2.4.2 构造GLMs

构造GLMs有三个假设：

1.  $y|x; \theta \sim \text{ExponentialFamily}(\eta)$ ，即给定特征向量 $x$ 和参数 $\theta$ ，输出 $y$ 符合以 $\eta$ 为参数的指数分布族。
2. 给定 $x$ ，目的是要求 $T(y)$ 的期望值，即 $h(x) = E[T(y)|x]$ 。 $T(y)$ 是 $y$ 的充分统计量，一般是 $T(y) = y$ 。
3. 自然参数 $\eta$ 和输入 $x$ 之间满足线性关系： $\eta = \theta^T x$ 。

### 2.4.2.1 最小二乘法（Ordinary Least Square）

考虑到目标变量（target variable）或者响应变量（response variable） $y$ 是连续的，假定其服从高斯分布，即 $y|x; \theta \sim \mathcal{N}(\mu, \sigma^2)$ 。注意高斯分布的期望是 $\mu$ ，标准差是 $\sigma$ 。此时， $T(y) = y$ ，所以

$$h_\theta(x) = E[T(y)|x; \theta] = E[y|x; \theta]$$

$$= \mu$$

$$= \eta$$

$$= \theta^T x$$

第一行基于假设2，第二行基于高斯分布的特性，第三行基于假设1，即高斯分布作为指数分布族的一个特例，第四行基于假设3。

### 2.4.2.2 Logistic Regression

LogisticRegression的输出只有 $y \in \{0, 1\}$ 两个值，所以考虑输出满足伯努利分布，即 $y|x; \theta \sim \text{Bernoulli}(\phi)$ ，而伯努利分布的期望 $E[y|x; \theta] = \phi$ ，所以有

$$h_{\theta}(x) = E[T(y)|x; \theta] = E[y|x; \theta]$$

$$= \phi$$

$$= \frac{1}{1 + e^{-\eta}}$$

$$= \frac{1}{1 + e^{-\theta^T x}}$$

这样就得到了logistic函数 $\frac{1}{1+e^{-z}}$ 。

### 2.4.2.3 Softmax Regression

考虑一个多类分类问题，假设 $y \in \{1, 2, \dots, k\}$ ，此时可以用多项式分布对 $y$ 进行建模。最终可以得到多元逻辑斯谛回归模型。

## 2.4.3 总结

从上面可以看出，广义线性模型（GLMs）的作用其实是提供一个统一的方法来实现训练模型的选择。其流程如下所示

$y|x; \theta \sim GLMs \rightarrow$  模型（logistic回归、softmax回归及最小二乘法等）

$\rightarrow$  由likelihood函数建立最优化模型

$\rightarrow$  参数求解（梯度法、牛顿法等）

## 第3章 生成算法

### § 3.1 高斯判别分析（Gaussian Discriminant Analysis）

高斯判别法是生成模型，虽然名字中有“判别”两个字。

#### 3.1.1 多元正态分布（Multivariate Normal Distribution）

假设  $x \sim \mathcal{N}(\mu, \Sigma)$ ，即  $x$  服从以  $\mu \in \mathbb{R}^n$  为均值向量（mean vector）， $\Sigma \in \mathbb{R}^{n \times n}$  为协方差矩阵（covariance matrix）的多元高斯分布，则  $x$  的概率密度函数（PDF）为

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1} (x - \mu)\right)$$

#### 3.1.2 高斯判别分析（GDA）

生成模型对每一个分类分别进行建模，即首先做关于分类和不同分类的输入特征的分布的假设

$$y \sim \text{Bernoulli}(\phi)$$

$$x|y = 0 \sim \mathcal{N}(\mu_0, \Sigma)$$

$$x|y = 1 \sim \mathcal{N}(\mu_1, \Sigma)$$

由此，参数为  $\phi, \Sigma, \mu_0, \mu_1$ 。注意两个高斯分布虽然有相同的均值向量，但是使用不同的协方差矩阵。

其对数似然函数为

$$\uparrow(\phi, \mu_0, \mu_1, \Sigma) = \log \prod_{i=1}^m p(x^{(i)}, y^{(i)}; \phi, \mu_0, \mu_1, \Sigma)$$

$$= \log \prod_{i=1}^m p(x^{(i)}|y^{(i)}; \mu_0, \mu_1, \Sigma) p(y^{(i)}; \phi)$$

解得参数的最大似然估计

$$\begin{aligned}\phi &= \frac{1}{m} \sum_{i=1}^m 1\{y^{(i)} = 1\} \\ \mu_0 &= \frac{\sum_{i=1}^m 1\{y^{(i)} = 0\} x^{(i)}}{\sum_{i=1}^m 1\{y^{(i)} = 0\}} \\ \mu_1 &= \frac{\sum_{i=1}^m 1\{y^{(i)} = 1\} x^{(i)}}{\sum_{i=1}^m 1\{y^{(i)} = 1\}} \\ \Sigma &= \frac{1}{m} \sum_{i=1}^m (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T.\end{aligned}$$

### 3.1.3 高斯判别分析与logistic回归的对比

在高斯判别分析中, 如果记 $p(y = 1|x; \phi, \mu_0, \mu_1, \Sigma)$ 为 $x$ 的函数, 则有如下形式

$$p(y = 1|x; \phi, \mu_0, \mu_1, \Sigma) = \frac{1}{1 + \exp(-\theta^T x)}$$

$\theta$ 是 $\phi, \Sigma, \mu_0, \mu_1$ 的函数。

事实上,

$$\left. \begin{aligned} p(x|y = 0) &\sim \mathcal{N}(\mu_0, \Sigma) \\ p(x|y = 1) &\sim \mathcal{N}(\mu_1, \Sigma) \end{aligned} \right\} \Rightarrow p(y = 1|x) \sim \text{LogitsticRegression}$$

反之, 不成立; 此外, 还有

$$\left. \begin{aligned} p(x|y = 0) &\sim \text{Poisson}(\lambda_0) \\ p(x|y = 1) &\sim \text{Poisson}(\lambda_1) \end{aligned} \right\} \Rightarrow p(y = 1|x) \sim \text{LogitsticRegression}$$

反之, 不成立; 更一般的, 有

$$\left. \begin{aligned} p(x|y = 0) &\sim \text{ExponentialFamily}(\eta_0) \\ p(x|y = 1) &\sim \text{ExponentialFamily}(\eta_1) \end{aligned} \right\} \Rightarrow p(y = 1|x) \sim \text{LogitsticRegression}$$

反之, 不成立。

证明一下最后一个公式。显然有

$$p(y) = \phi^y (1 - \phi)^{1-y}$$

$$p(x|y=0) = b(x)e^{\eta_0^T T(x) - a(\eta_0)}$$

$$p(x|y=1) = b(x)e^{\eta_1^T T(x) - a(\eta_1)}$$

则

$$\begin{aligned} p(y=1|x) &= \frac{p(x|y=1)p(y=1)}{p(x|y=1)p(y=1) + p(x|y=0)p(y=0)} \\ &= \frac{1}{1 + \frac{p(x|y=0)p(y=0)}{p(x|y=1)p(y=1)}} \\ &= \frac{1}{1 + \frac{1-\phi}{\phi} \exp(\eta_0^T T(x) - a(\eta_0) - \eta_1^T T(x) + a(\eta_1))} \\ &= \frac{1}{1 + \exp[(\eta_0^T T(x) - a(\eta_0) - \eta_1^T T(x) + a(\eta_1)) \ln \frac{1-\phi}{\phi}]} \end{aligned}$$

高斯判别分析事实上包含有比logistic回归更强的假设。高斯判别分析对输入数据分布做了假设，而logistic回归对输入数据符合什么分布，并没有做一般性的假设。所以logistic回归的泛化性能好一些，比高斯判别分析的鲁棒性好。但是高斯判别分析需要的训练数据比logistic回归需要的训练数据要少，也即使只有少量的训练数据，即使其不完全符合高斯分布，高斯判别分析也能得出很好的分类模型。

## § 3.2 朴素贝叶斯法

### 3.2.1 贝叶斯公式

**划分** 设 $S$ 为实验 $E$ 的样本空间， $B_1, B_2, \dots, B_n$ 为 $E$ 的一组事件，若

(1)  $B_i B_j = \emptyset, i \neq j, i, j = 1, 2, \dots, n$

(2)  $B_1 \cup B_2 \cup \dots \cup B_n = S$

则称 $B_1, B_2, \dots, B_n$ 为样本空间 $S$ 的一个**划分**。

**全概率公式** 设实验 $E$ 的样本空间为 $S$ ， $A$ 为 $E$ 的事件， $B_1, B_2, \dots, B_n$ 为 $S$ 的一组划分，且 $P(B_i) > 0, i = 1, 2, \dots, n$ ，则

$$P(A) = \sum_{i=1}^n P(A|B_i)$$

称为**全概率公式**。

**贝叶斯公式** 设实验 $E$ 的样本空间为 $S$ ， $A$ 为 $E$ 的事件， $B_1, B_2, \dots, B_n$ 为 $S$ 的一组划分，

且 $P(A) > 0, P(B_i) > 0, i = 1, 2, \dots, n$ , 则

$$P(B_i|A) = \frac{P(A|B_i)P(B_i)}{P(A)} = \frac{P(A|B_i)P(B_i)}{\sum_{i=1}^n P(A|B_i)}$$

称为贝叶斯公式。

其中,  $P(A), P(B_i)$  分别称为先验概率 (prior probability)。 $P(A|B_i), P(B_i|A)$  称为后验概率 (posterior probability)。

### 3.2.2 朴素贝叶斯法

#### 3.2.2.1 模型

朴素贝叶斯法 (naive Bayes) 之所以被称为朴素的, 是因为其包含有一些朴素的假设: 类条件独立假设。

假设  $D$  是训练元组及其分类标号的集合。特征向量表示为  $X = \{x_1, x_2, \dots, x_n\}$ , 其中  $x_1, x_2, \dots, x_n$  为  $n$  个属性的取值。假设有  $m$  个类,  $C_1, C_2, \dots, C_m$ 。给定元组  $X$ , 朴素贝叶斯分类法将  $X$  分给后验概率最大的类。即,  $X$  属于  $C_i$  类, 当且仅当

$$P(C_i|X) > P(C_j|X), j = 1, 2, \dots, m, i \neq j$$

则这个分类过程等价于最大化后验概率  $P(C_i|X)$ 。由贝叶斯公式得

$$\max_i P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}$$

其中,  $P(X)$  对所有分类都是一样的,  $P(C_i)$  如果做等概率假设, 即  $P(C_1) = P(C_2) = \dots = P(C_m)$ , 则上式等价于

$$\max_i P(X|C_i)$$

否则, 用训练样本里面每个分类出现的频次作为其先验概率估计值, 即  $P(C_i) = \frac{freq(C_i)}{|D|}$ , 则最优化问题等价于

$$\max_i P(X|C_i)P(C_i)$$

下面考虑  $P(X|C_i)$  的计算方法。为了简化计算量, 可以做类条件独立的朴素假设, 即属性之间不存在依赖关系, 则

$$P(X|C_i) = \prod_{k=1}^n P(x_k|C_i) = P(x_1|C_i)P(x_2|C_i) \cdots P(x_n|C_i)$$

这样, 求  $P(X|C_i)$  的问题就转化为求属性  $x_k$  在分类  $C_i$  中出现频次的问题了。此时, 需要区分属性取值是离散的, 还是连续的。

(1) 如果属性  $k$  的取值是离散的, 则有  $P(x_k|C_i) = \frac{freq(<x_k, C_i>)}{freq(C_i)}$ 。

(2)如果属性 $k$ 的取值是连续的, 假定连续值属性的取值 $x_k \sim \mathcal{N}(\mu, \sigma)$ , 则定义

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

同时, 有

$$P(x_k | C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i})$$

$\mu_{C_i}$  和  $\sigma_{C_i}$  分别表示  $C_i$  类中属性  $k$  的均值和标准差。

### 3.2.2.2 实现

计算概率及其乘积的过程涉及到数据的一些处理。首先是数据平滑问题, 有些属性值可能在某类的统计数据中并不出现, 这样会导致某一属性的概率值为0, 影响结果。实际使用时可以通过对分子分母同时加上一个常数来避免概率为0。

另一问题是数据下溢出的问题, 因为涉及到很多小数的乘积。一个解决方法是对概率乘积取自然对数。这样  $p_1 p_2$  就可以表示为  $\ln p_1 + \ln p_2$ , 取自然对数保证结果的单调性是相同的。

## 第4章 决策树



## 第5章 最大熵

## 第6章 支持向量机

## 第7章 提升方法

## 第8章 EM方法

### § 8.1 Jensen不等式

# 第9章 隐马尔可夫模型

## § 9.1 模型

## § 9.2 概率计算问题

### 9.2.1 直接计算

### 9.2.2 前向算法 (forward algorithm)

### 9.2.3 后向算法 (backward algorithm)

## § 9.3 学习问题

## § 9.4 预测问题

# 第10章 附录

## § 10.1 矩阵分析 (matrix analysis)

### 10.1.1 迹(Trace)和导数(matrix derivatives)

令  $f : \mathbb{R}^{m \times n} \mapsto \mathbb{R}$  表示将  $m \times n$  ( $m$ -by- $n$ ) 矩阵映射为实数的函数。定义  $f$  对矩阵  $\mathbf{A}$  的导数

$$\nabla_{\mathbf{A}} f(\mathbf{A}) = \begin{pmatrix} \frac{\partial f(\mathbf{A})}{\partial a_{11}} & \cdots & \frac{\partial f(\mathbf{A})}{\partial a_{1n}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f(\mathbf{A})}{\partial a_{n1}} & \cdots & \frac{\partial f(\mathbf{A})}{\partial a_{nn}} \end{pmatrix}$$

矩阵的迹 (trace) 表示的是矩阵的对角元素的和,

$$\text{tr} \mathbf{A} = \sum_{i=1}^n a_{ii}$$

假设  $A, B, C, D$  均是方阵

$$\text{tr} ABCD = \text{tr} DABC = \text{tr} CDAB = \text{tr} BCDA \quad (10.1)$$

循环将最右边矩阵放到最左边。假设  $a$  是实数

$$\text{tr} A = \text{tr} A^T \quad (10.2)$$

$$\text{tr}(A + B) = \text{tr} A + \text{tr} B \quad (10.3)$$

$$\text{tr} aA = a \text{tr} A \quad (10.4)$$

下面的一些公式出自 Andrew Ng 的机器学习讲义, 这里证明一下。

$$\nabla_A \text{tr} AB = B^T \quad (10.5)$$

$$\nabla_{A^T} f(A) = (\nabla_A f(A))^T \quad (10.6)$$

$$\nabla_A \text{tr} ABA^T C = CAB + C^T AB^T \quad (10.7)$$

$$\nabla_A |A| = |A|(A^{-1})^T \quad (10.8)$$

**证明** (1)  $(\nabla_A \text{tr} AB)_{ij} = \frac{\partial \text{tr} AB}{\partial a_{ij}} = \frac{\partial \sum_m \sum_k a_{mk} b_{km}}{\partial a_{ij}}$ , 只有当  $m = i, k = j$  时才有  $a_{ij}$  的系数, 所以  $(\nabla_A \text{tr} AB)_{ij} = b_{ji}$ , 即证。

(2)  $(\nabla_{A^T} f(A))_{ij} = \frac{\partial f(A)}{\partial a_{ji}}$ , 即证。

(3)  $\text{tr} ABA^T C = \sum_m \sum_k \sum_t \sum_s a_{mk} b_{kt} a_{st} c_{sm}$ , 所以

$$\begin{aligned} (\nabla_A \text{tr} ABA^T C)_{ij} &= \frac{\partial \sum_m \sum_k \sum_t \sum_s a_{mk} b_{kt} a_{st} c_{sm}}{\partial a_{ij}} \\ &= \sum_m \sum_k \sum_t \sum_s \frac{\partial a_{mk}}{\partial a_{ij}} b_{kt} a_{st} c_{sm} + \sum_m \sum_k \sum_t \sum_s a_{mk} b_{kt} \frac{\partial a_{st}}{\partial a_{ij}} c_{sm} \end{aligned}$$

左边, 令  $m = i, k = j$ , 右边, 令  $s = i, t = j$ ,

$$\begin{aligned} (\nabla_A \text{tr} ABA^T C)_{ij} &= \sum_t \sum_s b_{jt} a_{st} c_{si} + \sum_m \sum_k a_{mk} b_{kj} c_{im} \\ &= \sum_t \sum_s b_{jt} a_{st} c_{si} + \sum_m \sum_k c_{im} a_{mk} b_{kj} \\ &= (BA^T C)_{ji} + (CAB)_{ij} \\ &= (C^T AB^T + CAB)_{ij} \end{aligned}$$

## § 10.2 常用不等式

### 10.2.1 柯西不等式 (Cauchy Inequality)

柯西不等式, 又称柯西-施瓦茨不等式 (Cauchy-Schwarz inequality)。对于一个内积空间所有向量  $\mathbf{x}$  和  $\mathbf{y}$ ,

$$|\langle \mathbf{x}, \mathbf{y} \rangle|^2 \leq \langle \mathbf{x}, \mathbf{x} \rangle \cdot \langle \mathbf{y}, \mathbf{y} \rangle$$

其中  $\langle \cdot, \cdot \rangle$  表示内积 (点积), 当且仅当  $\mathbf{x}$  与  $\mathbf{y}$  线性相关时等式成立。

对于欧几里得空间 $\mathbb{R}^2$ ,

$$\left(\sum_{i=1}^n x_i y_i\right)^2 \leq \left(\sum_{i=1}^n x_i^2\right) \left(\sum_{i=1}^n y_i^2\right)$$

当且仅当 $\frac{x_1}{y_1} = \frac{x_2}{y_2} = \cdots = \frac{x_n}{y_n}$ 时等式成立。

### 10.2.2 赫尔德不等式 (Hölder Inequality)

赫尔德不等式揭示了 $L^p$ 空间的相互关系。设 $S$ 为测度空间,  $1 \leq p, q \leq \infty$ , 且 $\frac{1}{p} + \frac{1}{q} = 1$ , 若 $f \in L^p(S)$ ,  $g \in L^q(S)$ , 则 $fg \in L^1(S)$ , 且

$$\|fg\|_1 \leq \|f\|_p \|g\|_q$$

写成序列或向量的形式

$$\sum_{i=1}^n |a_i b_i| \leq \left(\sum_{i=1}^n |a_i|^p\right)^{\frac{1}{p}} \left(\sum_{i=1}^n |b_i|^q\right)^{\frac{1}{q}}$$

### 10.2.3 闵可夫斯基不等式 (Minkowski Inequality)

闵可夫斯基不等式表明 $L^p$ 空间是一个赋范向量空间。设 $S$ 是一个度量空间,  $f, g \in L^p(S)$ ,  $1 \leq p \leq \infty$ , 那么 $f+g \in L^p(S)$ , 有

$$\|f+g\|_p \leq \|f\|_p + \|g\|_p$$

写成序列或向量的形式

$$\left(\sum_{k=1}^n |x_k + y_k|^p\right)^{\frac{1}{p}} \leq \left(\sum_{k=1}^n |x_k|^p\right)^{\frac{1}{p}} + \left(\sum_{k=1}^n |y_k|^p\right)^{\frac{1}{p}}$$



## 参考文献

- [1] 李航著. 《统计学习方法》. 北京:清华大学出版社, 2012, 3
- [2] Jiawei Han, Micheline Kamber, Jian Pei 著.范明, 孟小峰译. 《数据挖掘: 概念与技术》. 机械工业出版社, 2012, 8
- [3] Tom M. Mitchell 著.曾华军等译. 《机器学习》.机械工业出版社, 2003, 1
- [4] Andrew Ng, 《Machine Learning》公开课讲义