# 机器学习笔记–多元分类

空修菜

## 1 多元分类——Softmax Regression

笔记分为两部分, 第一部分是笔记的总结; 第二部分是模型相关的推导过程.

### 1.1 Summary

1. softmax 模型处理的多分类问题, 即对于每个实例 $x$ 的标签 $y$ 有 $y \in \{1, 2, \cdots, k\}$. 它是罗辑回归的一般情形, 当 $k = 2$ 时，softmax 模型就是罗辑回归模型;

2. 假设函数 $h(\theta, x)$ 是一个关于 $x$ 的概率向量 (分量元素大于等于 0 且相加为 1), 它给出 $x$ 的每个类别的概率;

$$h(\theta, x) = \frac{1}{\sum_{j=1}^{k} e^{\theta_j^T x}} \begin{pmatrix} e^{\theta_1^T x} \\ e^{\theta_2^T x} \\ \vdots \\ e^{\theta_k^T x} \end{pmatrix}$$

3. 在上面的 $h(\theta, x)$ 中, $\theta_i$ $(1 \le i \le k)$ 是第 $i$ 类的权重向量. $x$ 是第 $i$ 类的概率就是

$$p(y = k \mid x) = \frac{e^{\theta_i^T x}}{\sum_{j=1}^{k} e^{\theta_j^T x}},$$

因此, $x$ 的类别 $\hat{y}$ 就是概率向量 $h(\theta, x)$ 中最大值的索引, 即

$$\hat{y} = \underset{i=1,2,\cdots,k}{\arg\max}\, h(\theta, x).$$

4. 最终的任务是确定每一个类别 $s$ 的权重向量 $\theta_s$, 这由梯度下降法给出, 每一个权重向量 $\theta_s$ $(1 \le s \le k)$ 的迭代公式为:

$$\theta_s := \theta_s + \frac{\beta}{m} \sum_{i=1}^{m} x^{(i)} \left( 1\{y^{(i)} = s\} - \frac{e^{\theta_s^T x^{(i)}}}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}} \right).$$

5. 迭代公式的矩阵表示. 假设共分为 $k$ 个类, $\theta = (\theta_1, \cdots, \theta_k)$,

$$\theta = \begin{pmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1n} \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{k1} & \theta_{k2} & \cdots & \theta_{kn} \end{pmatrix},$$

同时, 为简便记, 令

$$\alpha_s^{(i)} = 1\{y^{(i)} = s\} - \frac{e^{\theta_s^T x^{(i)}}}{\sum_{j=1}^k e^{\theta_j^T x^{(i)}}},$$

则有

$$\alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_k \end{pmatrix} = \begin{pmatrix} \alpha_1^{(1)} & \alpha_1^{(2)} & \cdots & \alpha_1^{(m)} \\ \alpha_2^{(1)} & \alpha_2^{(2)} & \cdots & \alpha_2^{(m)} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_k^{(1)} & \alpha_k^{(2)} & \cdots & \alpha_k^{(m)} \end{pmatrix}$$

所以, 关于参数 $\theta$ 的迭代矩阵就可以写为:

$$\theta := \begin{pmatrix} \theta_{11} & \theta_{12} & \cdots & \theta_{1n} \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{k1} & \theta_{k2} & \cdots & \theta_{kn} \end{pmatrix} + \frac{\beta}{m} \begin{pmatrix} \alpha_1^{(1)} & \alpha_1^{(2)} & \cdots & \alpha_1^{(m)} \\ \alpha_2^{(1)} & \alpha_2^{(2)} & \cdots & \alpha_2^{(m)} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_k^{(1)} & \alpha_k^{(2)} & \cdots & \alpha_k^{(m)} \end{pmatrix} \begin{pmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_n^{(1)} \\ x_1^{(2)} & x_2^{(2)} & \cdots & x_n^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ x_1^{(m)} & x_2^{(m)} & \cdots & x_n^{(m)} \end{pmatrix}$$

6. **问题：冗余参数**. $\theta$ 是一个 $n \times k$ 的矩阵, $v$ 是一个 $n$ 维列向量, 若用 $\theta - v$, 则有如下结果:

$$\begin{aligned} p(y = s \mid x; \theta - v) &= \frac{\exp\left((\theta_s - v)^T x\right)}{\sum_{i=1}^k \exp\left((\theta_i - v)^T x\right)} \\ &= \frac{\exp(\theta_s^T x)}{\sum_{i=1}^k \exp(\theta_i - v)^T x)} \\ &= p(y = s \mid x; \theta) \end{aligned}$$

这就意味着 $\theta$ 并不唯一。解决方法是在函数 $\ell(\theta)$ 中加入一个约束项:

$$\frac{\lambda}{2} \sum_{i=1}^k \sum_{s=1}^n \theta_{is}^2,$$

这个约束项实际上就是将矩阵 $\theta$ 的每一个列向量 $\theta_s$ 的元素的平方和，然后再将每一个列向量的元素平方和相加。

7. 加入约束项的迭代公式为:

$$\theta_s := (1-\frac{\alpha\lambda}{n})\theta_s + \frac{\alpha}{n}\sum_{i=1}^{n} x^{(i)} \left(1\{y^{(i)}=s\} - \frac{e^{\theta_s^T x^{(i)}}}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}}\right).$$

### 1.2 推导过程

Softmax 回归是 $y$ 取多个值时的情形。设 $y$ 可以取 $k$ 个值, $y=1,2,...,k$。每一个 $k$ 都对应了一个 $\theta$。假设每个类的概率分别为 $\phi_i$, 所以有 $\phi_1,\phi_2,...,\phi_k$。又因为 $\sum_{i=1}^{k}=1$, 所以当我们确定了前面 $k$ 个数之后，就能确定最后一个数。所以我们只考虑 $y=1,y=2,...,y=k-1$ 中情况。$p(y=i)=\phi_i$. 这里为方便记录，将 $y$ 写成列向量的形式, 每一个列向量共有 $k-1$ 行。

$$T(1)=\begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix},T(2)=\begin{pmatrix}0\\1\\\vdots\\0\end{pmatrix},....,T(k-1)=\begin{pmatrix}0\\0\\\vdots\\1\end{pmatrix},T(k)=\begin{pmatrix}0\\0\\\vdots\\0\end{pmatrix}$$

令 $(T(y))_i$ 表示向量 $T(y)$ 的第 $i$ 个元素，所以 $(T(y))_i$ 不是 0 就是 1, 所以 $(T(y))_i=1\{y=i\}$, 若 $y=i$，取 1；若 $y \neq$i, 则取 0 。

所以，对于 $y=1,2,...,k-1$, 上面的向量就可以统一地写为:

$$T(y)=\begin{pmatrix}(T(y))_1\\(T(y))_2\\\vdots\\(T(y))_{k-1}\end{pmatrix}$$

在参数 $\theta$ 和 $x$ 都给定时, 有 $p(T(y)_i)=p(y=i|x;\theta)=\phi_i$. 由数学期望的定义，当 $i$ 是固定的时候，对于 $(T(y))_i$ 的数学期望有（要明白此时对于固定的 $i$, $(T(y))_i$ 只取 0 或 1 两个值）。

$$\mathbb{E}[(T(y))_i]=1\cdot p(T(y)_i)+0\cdot p(T(y)_i)=p(y=i)=\phi_i.$$

此时，利用类似于两点分布的表示方式，即 $p^y(1-p)^{1-y}$, 我们就可以将

$p(y;\phi)$ 表示为

$$
\begin{aligned}
p(y;\phi) &= \phi_1^{1\{y=1\}}\phi_2^{1\{y=2\}}\cdots\phi_k^{1\{y=k\}} \\
&= \phi_1^{1\{y=1\}}\phi_2^{1\{y=2\}}\cdots\phi_k^{1-\sum_{i=1}^{k-1}1\{y=i\}} \\
&= \phi_1^{(T(y))_1}\phi_2^{(T(y))_2}\cdots\phi_k^{1-\sum_{i=1}^{k-1}(T(y))_i} \\
&= \exp\Big((T(y))_1\log\phi_1 + (T(y))_2\log\phi_2 + \cdots (1-\sum_{i=1}^{k-1}(T(y))_i)\log\phi_k\Big) \\
&= \exp\Big((T(y))_1\log\frac{\phi_1}{\phi_k} + (T(y))_2\log\frac{\phi_2}{\phi_k} + \cdots + (T(y))_{k-1}\log\frac{\phi_{k-1}}{\phi_k} + \log\phi_k\Big) \\
&= b(y)\exp(\eta^T T(y) - a(\eta))
\end{aligned}
$$

令 $\eta^T = (\log\frac{\phi_1}{\phi_k}, log\frac{\phi_2}{\phi_k}, ..., log\frac{\phi_{k-1}}{\phi_k})$, 进一步地

$$
\begin{aligned}
\eta &= \begin{pmatrix} \log\frac{\phi_1}{\phi_k} \\ log\frac{\phi_2}{\phi_k} \\ \vdots \\ log\frac{\phi_{k-1}}{\phi_k} \end{pmatrix}, \\
b(y) &= 1 \\
a(y) &= -\log\phi_k.
\end{aligned}
$$

由上面的表示我们知道，对于 $i = 1, ..., k-1$, $\eta^T = (\eta_1, \eta_2, ..., \eta_{k-1})$, 有

$$\eta_i = \log\frac{\phi_i}{\phi_k},$$

所以

$$e^{\eta_i} = \frac{\phi_i}{\phi_k},$$

所以

$$\phi_k e^{\eta_i} = \phi_i,$$

同时，令 $\eta_k = \log\frac{\phi_k}{\phi_k} = 0$. 因此

$$1 = \sum_{i=1}^{k}\phi_i = \sum_{i=1}^{k}\phi_k e^{\eta_i} = \phi_k\sum_{i=1}^{k}e^{\eta_i}$$

因此，

$$\phi_k = \frac{1}{\sum_{i=1}^{k}e^{\eta_i}},$$

所以

$$\phi_i = \frac{e^{\eta_i}}{\sum_{i=1}^k e^{\eta_i}}.$$

这就得到了每一个类，即 $y=i$ 时的概率大具体表达式。将 $\eta$ 映射成 $\phi$ 的函数称为 **Softmax function.**

由之前的关于 GLM 的假设 2,即 $\eta = \theta^T x$。设存在 $k$ 个向量 $\theta_1, \theta_2, ..., \theta_{k-1}$, $\theta_i$ 的维数与 $x^{(i)}$ 的维数相同。$\theta_k$ 是一个 0 向量。$\eta_i = \theta_i^T x$, 所以

$$\begin{aligned} p(y=i|x;\theta) &= \phi_i \\ &= \frac{e^{\eta_i}}{\sum_{j=1}^k e^{\eta_j}} \\ &= \frac{e^{\theta_i^T x}}{\sum_{j=1}^k e^{\theta_j^T x}} \end{aligned}$$

再由 GLM 的假设 1，就可以得到 hypothesis function $h_\theta(x)$,

$$\begin{aligned} h_\theta(x) &= \mathbb{E}[T(y)|x;\theta] \\ &= \mathbb{E}\begin{bmatrix} (T(y))_1 \\ (T(y))_2 \\ \vdots \\ (T(y))_{k-1} \end{bmatrix} = \begin{pmatrix} \mathbb{E}[(T(y))_1] \\ \mathbb{E}[(T(y))_2] \\ \vdots \\ \mathbb{E}[(T(y))_{k-1}] \end{pmatrix} \\ &= \begin{pmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{k-1} \end{pmatrix} \\ &= \begin{pmatrix} \frac{e^{\theta_1^T x}}{\sum_{j=1}^k e^{\theta_j^T x}} \\ \frac{e^{\theta_2^T x}}{\sum_{j=1}^k e^{\theta_j^T x}} \\ \vdots \\ \frac{e^{\theta_{k-1}^T x}}{\sum_{j=1}^k e^{\theta_j^T x}} \end{pmatrix} \end{aligned}$$

由上面的讨论可以知道，我们最后得到的是一个列向量。在给定了 $\theta$ 和 $x$ 时，列向量里的每一个元素是 $y$ 可能出现在每一个类的概率。同时，由于列

向量 $h_\theta(x)$ 只有 $k-1$ 个元素，由概率的定义，用 1 减去 $h_\theta(x)$ 中的 $k-1$ 个元素，将得到第 $k$ 类的可能概率。

最后，写出以 $\theta$ 为自变量的概率的表达式，以此为基础就好进行极大似然估计。由于 log 是单调的，所以可以利用这个函数。由于 $\phi_i \in [0,1]$，所以就要取 $-\log f(\theta)$ 的最大值或 $\log f(\theta)$ 的最小值。

$$
\begin{aligned}
\ell(\theta) &= -\log \prod_{i=1}^{n} p(y^{(i)}|x^{(i)};\theta) \\
&= -\sum_{i=1}^{n} \log p(y^{(i)}|x^{(i)};\theta) \\
&= -\sum_{i=1}^{n} \log \prod_{l=1}^{k} \left( \frac{e^{\theta_l^T x^{(i)}}}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}} \right)^{1\{y^{(i)}=l\}} \\
&= -\sum_{i=1}^{n} \sum_{l=1}^{k} 1\{y^{(i)}=l\} \log \left( \frac{e^{\theta_l^T x^{(i)}}}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}} \right)
\end{aligned}
$$

对于 $i=1,2,...,k$, $\theta_i \in \mathbb{R}^{d+1}$, $\theta_i = (\theta_{i1}, \theta_{i2}, ..., \theta_{i(d+1)})$, 即 $\theta_{i,s}$ 表示向量 $\theta_1$ 的第 $s$ 个元素（$1 \le s \le d+1$）。

令

$$
g_l = \frac{e^{\eta_l}}{\sum_{j=1}^{k} e^{\eta_j}}
$$

如果 $l = s$, 则

$$
\frac{\partial g_l}{\partial \eta_s} = \frac{e^{\eta_l}\sum_{j=1}^{k} e^{\eta_j} - e^{\eta_l}e^{\eta_s}}{(\sum_{j=1}^{k} e^{\eta_j})^2} = \frac{e^{\eta_l}}{\sum_{j=1}^{k} e^{\eta_j}} \frac{\sum_{j=1}^{k} e^{\eta_j} - e^{\eta_l}}{\sum_{j=1}^{k} e^{\eta_j}} = g_l(1-g_l),
$$

若 $l \neq s$, 则

$$
\frac{\partial g_l}{\partial \eta_s} = \frac{0 - e^{\eta_s}e^{\eta_l}}{(\sum_{j=1}^{k} e^{\eta_j})^2} = -g_l g_s,
$$

令

$$
\ell'(\theta) = -\sum_{l=1}^{k} 1\{y^{(i)}=l\} \log \left( \frac{e^{\theta_l^T x^{(i)}}}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}} \right) = -\sum_{l=1}^{k} 1\{y^{(i)}=l\} \log g_l
$$

所以

$$
\begin{aligned}
\frac{\partial \ell'(\theta)}{\partial \eta_s} &= -\sum_{l=1}^{k} 1\{y^{(i)} = l\}\frac{1}{g_l}\frac{\partial g_l}{\partial \eta_s} \\
&= -1\{y^{(i)} = s\}\frac{1}{g_s}g_s(1-g_s) - \sum_{l\neq s}^{k} 1\{y^{(i)} = l\}\frac{1}{g_l}(-g_l g_s) \\
&= -1\{y^{(i)} = s\} + 1\{y^{(i)} = s\}g_s + \sum_{l\neq s}^{k} 1\{y^{(i)} = l\}g_s \\
&= -1\{y^{(i)} = s\} + g_s\sum_{l=1}^{k} 1\{y^{(i)} = l\} \\
&= g_s - 1\{y^{(i)} = s\}
\end{aligned}
$$

我们对 $\ell'$ 关于 $\theta_{st}$ 求导数，其中 $1 \leq s \leq k, 1 \leq t \leq d+1$。由链式法则我们知道

$$
\frac{\partial \ell'(\theta)}{\partial \theta_{st}} = \frac{\partial \ell'(\theta)}{\partial \eta s}\frac{\partial \eta_s}{\partial \theta_{st}}
$$

而

$$
\frac{\partial \eta_s}{\partial \theta_{st}} = \frac{\partial \theta_s^T x^{(i)}}{\partial \theta_{st}} = x_t^{(i)},
$$

所以我们就得到了

$$
\frac{\partial \ell'(\theta)}{\partial \theta_{st}} = (g_s - 1\{y^{(i)} = s\})x_t^{(i)}.
$$

由上面的讨论以及 $\ell(\theta)$ 的定义我们知道

$$
\frac{\partial \ell(\theta)}{\partial \theta_{st}} = \sum_{i=1}^{n}(g_s - 1\{y^{(i)} = s\})x_t^{(i)},
$$

当 $t$ 分别取 $1, 2, ..., d+1$ 时，我们就得到了向量 $\theta_s$ 的导数。此处的对 $\theta_s$ 求

导在符号上并不是严格意义上的求导。

$$
\begin{aligned}
\frac{\partial \ell(\theta)}{\partial \theta_s} = \begin{pmatrix} \frac{\partial \ell(\theta)}{\partial \theta_{s1}} \\ \frac{\partial \ell(\theta)}{\partial \theta_{s2}} \\ \vdots \\ \frac{\partial \ell(\theta)}{\partial \theta_{s(d+1)}} \end{pmatrix} &= \begin{pmatrix} \sum_{i=1}^{n}(g_s - 1\{y^{(i)} = s\})x_1^{(i)} \\ \sum_{i=1}^{n}(g_s - 1\{y^{(i)} = s\})x_2^{(i)} \\ \vdots \\ \sum_{i=1}^{n}(g_s - 1\{y^{(i)} = s\})x_{d+1}^{(i)} \end{pmatrix} \\
&= \sum_{i=1}^{n} \left[ (g_s - 1\{y^{(i)} = s\}) \begin{pmatrix} x_1^{(i)} \\ x_2^{(i)} \\ x_3^{(i)} \\ \vdots \\ x_{(d+1)}^{(i)} \end{pmatrix} \right] \\
&= \sum_{i=1}^{n}(g_s - 1\{y^{(i)} = s\})x^{(i)}
\end{aligned}
$$

所以，由 gradient descent 我们立即得到关于参数 $\theta$ 的拟合回归表达式

$$
\begin{aligned}
\theta_s := \theta_s - \frac{\alpha}{n}\sum_{i=1}^{n}(g_s - 1\{y^{(i)} = s\})x^{(i)} \\
= \theta_s + \frac{\alpha}{n}\sum_{i=1}^{n} x^{(i)} \left( 1\{y^{(i)} = s\} - \frac{e^{\theta_s^T x^{(i)}}}{\sum_{j=1}^{k} e^{\theta_j^T x^{(i)}}} \right).
\end{aligned}
$$

1. 学习率 $\alpha$ 应该尽量大些，经验值 $0.1 \le \alpha \le 0.5$。

2. 上面的迭代公式中 $x^{(i)}$ 是一个向量。这个迭代方式得到的是一个向量 $\theta$，所以进行性批量梯度下降得到就是一个矩阵。