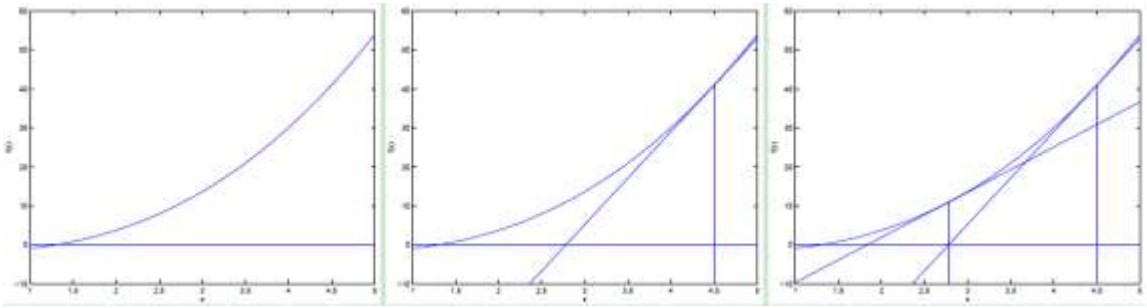
斯坦福 ML 公开课笔记 4

这篇笔记针对公开课的第 4 个视频，主要内容包括牛顿方法、指数分布族、广义线性模型。

牛顿方法

牛顿方法（NewTon's Method）与梯度下降（gradient descent）方法的功能一样，都是对解空间进行搜索的方法。其基本思想如下：

对于一个函数 f(x)，如果我们要求函数值为 0 时的 x，如图所示：



我们先随机选一个点，然后求出该点的切线，即导数，延长它使之与 x 轴相交，以相交时的 x 的值作为下一次迭代的值。

更新规则为：

$$\theta^{(t+1)} = \theta^{(t)} - \frac{f(\theta^{(t)})}{f'(\theta^{(t)})} \tag{1}$$

那么如何将牛顿方法应用到机器学习问题求解中呢？

对于机器学习问题，我们优化的目标函数为极大似然估计 L，当极大似然估计函数取值最大时，其导数为 0，这样就和上面函数 f 取 0 的问题一致了。极大似然函数的求解更新规则是：

$$\theta^{(t+1)} = \theta^{(t)} - \frac{f'(\theta^{(t)})}{f''(\theta^{(t)})} \tag{2}$$

上面是当参数θ为实数时的情况，当参数为向量时，更新规则变为如下所示：

$$\theta^{(t+1)} = \theta^{(t)} - H^{-1}\nabla_{\theta} L \tag{3}$$

其中，H 是一个 n*n 的矩阵，n 为参数向量的长度，H 是函数的二次导数矩阵，被成为 Hessian 矩阵。其某个元素 H_{ij} 计算公式如下：

$$H_{ij} = \frac{\partial L}{\partial \theta_i \partial \theta_j} \tag{4}$$

牛顿方法相对于梯度下降的优点是收敛速度快，通常十几次迭代就可以收敛。它也被称为二次收敛（quadratic convergence），因为当迭代到距离收敛值比较近的时候，每次迭代都能使误差变为原来的平方。缺点是当参数向量较大的时候，每次迭代都需要计算一次 Hessian 矩阵的逆，比较耗时。

如果目标函数求得是最小值的话，那么更新规则不会改变，那如何判断得到的参数使目标函数达到极大值还是极小值呢？可以通过判断二阶导数的值来确定，当二阶导数小于 0 时，即为最大值，当二阶导数大于 0 时，即为最小值。

指数分布族

指数分布族是指可以表示为指数形式的概率分布。指数分布的形式如下：

$$P(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta)) \tag{5}$$

其中，η成为分布的自然参数（nature parameter）；T(y)是充分统计量（sufficient statistic），通常 T(y)=y。当参数 a、b、T 都固定的时候，就定义了一个以η为参数的函数族。

实际上，大多数概率分布都可以表示成公式 5 的形式。比如：

1) 伯努利分布（Bernoulli）：对 0、1 问题进行建模；
2) 多项式分布（Multinomial）：多有 K 个离散结果的事件建模；
3) 泊松分布（Poisson）：对计数过程进行建模，比如网站访问量的计数问题，放射性衰变的数目，商店顾客数量等问题；
4) 伽马分布（gamma）与指数分布（exponential）：对有间隔的正数进行建模，比如公交车的到站时间问题；
5) β 分布：对小数建模；
6) Dirichlet 分布：对概率分布进建模；
7) Wishart 分布：协方差矩阵的分布；
8) 高斯分布（Gaussian）；

现在，我们将高斯分布与伯努利分布表示成为指数分布族的形式。

伯努利分布是对 0,1 问题进行建模的分布，它可以用如下形式表示：

$$P(y;\varphi) = \varphi^y(1-\varphi)^{1-y} \quad y\epsilon\{0,1\} \tag{6}$$

这个形式，我们在上一篇笔记中见过，就不再详述了。我们将其转换形式，推导如下：

$$\begin{aligned} \mathrm{P}(\mathrm{y};\varphi) &= \varphi^{y}(1-\varphi)^{1-y} = \exp(\log\varphi^{y}(1-\varphi)^{1-y}) \\ &= \exp(\mathrm{y}\log\varphi + (1-\mathrm{y})\log(1-\varphi)) \\ &= \exp(\mathrm{y}\log\frac{\varphi}{1-\varphi} + \log(1-\varphi)) \end{aligned} \tag{7}$$

由公式 7，我们就将伯努利分布表示成公式 5 的形式；其中：

$$\mathrm{b}(\mathrm{y}) = 1$$

$$\mathrm{T}(\mathrm{y}) = \mathrm{y}$$

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

$$\mathrm{a}(\eta) = -\log(1-\varphi) = 1+e^{-\eta}$$

可以看到，η的形式与上一篇笔记中的 logistic 函数一致，这是因为 logistic 模型对问题的前置概率估计是伯努利分布的缘故。

由高斯分布可以推导出线性模型，由线性模型的假设函数可知，高斯分布的方差与假设函数无关，因而为了简便计算，我们将方差设为 1，即便不这样做，最后的结果也是方差作为一个系数而已。高斯分布转换为指数分布族形式的推导过程如下：

$$\begin{aligned} \mathrm{N}(\mu,1) &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(y-\mu)^{2}\right) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y^{2}-\frac{1}{2}\mu^{2}+\mu y\right) \\ &= \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}y^{2})\exp(\mu y-\frac{1}{2}\mu^{2}) \end{aligned} \tag{8}$$

由公式 8 可知：

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

$$\mathrm{T}(\mathrm{y}) = \mathrm{y}$$

$$\eta = \mu$$

$$\mathrm{a}(\eta) = \frac{1}{2}\mu^{2}$$

推导的关键在于将指数内部的纯 y 项移到外面，纯非 y 项作为函数 a，混杂项为$\eta^{T}T(y)$。

广义线性模型

定义了指数分布族后有什么用呢？我们可以通过指数分布族引出广义线性模型（Generalized Linear Model，GLM）。注意到上述公式 7 与公式 8 的η变量，在公式 7 中，η与伯努利分布中的参数φ的关系是 logistic 函数，再通过推导可以

得到逻辑斯蒂回归（推导会在下面）；在公式 8 中，η与正态分布的参数 u 的关系是相等，我们可以推导出最小二乘模型（Ordinary Least Squares）。通过这两个例子，我们大致可以得到结论，η以不同的映射函数与其他概率分布函数中的参数发生联系，从而得到不同的模型，广义线性模型正是将指数分布族中的所有成员（每个成员正好有一个这样的联系）都作为线性模型的扩展，通过各种非线性的连接函数将线性函数映射到其他空间从而大大扩大了线性模型可解决的问题。

下面我们看 GLM 的形式化定义，GLM 有三个假设：

1) y|x; θ~ExpFamily(η)；给定样本 x 与参数θ，样本分类 y 服从指数分布族中的某个分布；
2) 给定一个 x，我们需要的目标函数为$h_\theta(x) = E[T(y)|x]$
3) $\eta = \theta^T x$

依据这三个假设，我们可以推导出 logistic 模型与最小二乘模型。Logistic 模型的推导过程如下：

$$h_\theta(x) = E[T(y)|x] = \mathrm{E}[y|x] = \mathrm{p}(y = 1|x; \theta) = \varphi$$
$$= \frac{1}{1 + e^{-\eta}} = \frac{1}{1 + e^{-\theta^T x}} \quad (9)$$

公式 9 中，第一行是伯努利分布的性质，第二行由假设二与假设三推出。

同样的，对于最小二乘模型，推导过程如下：

$$h_\theta(x) = E[T(y)|x] = \mathrm{E}[y|x] = \mu = \eta = \theta^T x \quad (10)$$

其中，将η与原始概率分布中的参数联系起来的函数成为正则相应函数（canonical response function），如$\varphi = \frac{1}{1+e^{-\eta}}$、$\mu = \eta$即是正则响应函数。正则响应函数的逆成为正则关联函数（canonical link function）。

所以，对于广义线性模型，需要决策的是选用什么样的分布，当选取高斯分布时，我们就得到最小二乘模型，当选取伯努利分布时，我们得到 logistic 模型，这里所说的模型是假设函数 h 的形式。

所以总结一下，广义线性模型通过假设一个概率分布，得到不同的模型，而之前所讨论的梯度下降、牛顿方法都是为了求取模型中的线性部分（$\theta^T x$）的参数θ的。

GLM 举例-多项式分布

多项式分布推导出的 GLM 可以解决多类分类问题，是 logistic 模型的扩展。应用的问题比如邮件分类、预测病人疾病等。

多项式分布的目标值 $y \in \{1,2,3,\dots,k\}$；其概率分布为：

$$P(y = i) = \varphi_i \tag{11}$$

其中，因为$\sum \varphi_i = 1$，所以我们可以只保留 k-1 个参数，使得：

$$\varphi_k = 1 - \sum_{i=1}^{k-1} \varphi_i \tag{12}$$

为了使多项式分布能够写成指数分布族的形式，我们首先定义 T(y)，如下所示：

$$T(1) = \begin{bmatrix}1\\0\\0\\\vdots\\0\end{bmatrix}, T(2) = \begin{bmatrix}0\\1\\0\\\vdots\\0\end{bmatrix}, \dots, T(k-1) = \begin{bmatrix}0\\0\\0\\\vdots\\1\end{bmatrix}, T(k) = \begin{bmatrix}0\\0\\0\\\vdots\\0\end{bmatrix} \tag{13}$$

这样，我们还可以引入指示函数 I，使得

$$I(\text{True}) = 1, I(\text{False}) = 0 \tag{14}$$

这样，T(y)向量中的某个元素还可以表示成：

$$T(y)_i = I(y = i) \tag{15}$$

举例来说，当 y=2 时，$T(2)_2=I(2=2)=1$，$T(2)_3=I(2=3)=0$。根据公式 15，我们还可以得到：

$$E[T(y)_i] = \sum_{y=1}^{k} T(y)_i \varphi_i = \sum_{y=1}^{k} I(y = i)\varphi_i = \varphi_i \tag{16}$$

$$\sum_{i=1}^{k} I(y = i) = 1 \tag{17}$$

于是，二项分布转变为指数分布族的推导如下：

$$\begin{aligned} P(y;\varphi) &= \varphi_1^{I\{y=1\}}\varphi_2^{I\{y=2\}} \dots \varphi_{k-1}^{I\{y=k-1\}}\varphi_k^{I\{y=k\}} \\ &= \varphi_1^{I\{y=1\}}\varphi_2^{I\{y=2\}} \dots \varphi_{k-1}^{I\{y=k-1\}}\varphi_k^{1-\sum_{i=1}^{k-1} I(y=i)} \\ &= \exp(\log\varphi_1^{I\{y=1\}}\varphi_2^{I\{y=2\}} \dots \varphi_{k-1}^{I\{y=k-1\}}\varphi_k^{1-\sum_{i=1}^{k-1} I(y=i)}) \end{aligned}$$

$$= \exp(\sum_{i=1}^{k-1} I(y=i)log\varphi_i + (1-\sum_{i=1}^{k-1} I(y=i))log\varphi_k)$$

$$= \exp(\sum_{i=1}^{k-1} I(y=i)\log\left(\frac{\varphi_i}{\varphi_k}\right) + log\varphi_k)$$

$$= \exp(\sum_{i=1}^{k-1} T(y)_i \log\left(\frac{\varphi_i}{\varphi_k}\right) + log\varphi_k)$$

$$= \exp(\eta^T \mathrm{T(y)} - \mathrm{a}(\eta)) \quad (18)$$

其中，公式 18 中的 T(y)已不再等于 y，而是一个向量。公式的推导过程中，第一步代入公式 17；第二步指数对数变换；第三步将对数内的乘积变为对数外的相加；第四步将 Y 的部分归入到指数内的第一项；第五步代入公式 15；第六步将连加和转变为向量相乘的形式。

公式 18 最后一步的各个变量分别如下：

$$\eta = \begin{bmatrix} log\,\varphi_1/\varphi_k \\ log\,\varphi_2/\varphi_k \\ \vdots \\ log\,\varphi_{k-1}/\varphi_k \end{bmatrix}$$

$$\mathrm{b(y)} = 1$$

$$\mathrm{a}(\eta) = -\log\varphi_k$$

由η表达式可知：

$$\eta_i = log\,\varphi_i/\varphi_k \Longrightarrow \varphi_i = \varphi_k e^{\eta_i} \quad (19)$$

为了方便表示，再定义：

$$\eta_k = log\,\varphi_k/\varphi_k = 0 \quad (20)$$

于是，我们可以得到：

$$\sum_{j=1}^{k} \varphi_i = \sum_{j=1}^{k} \varphi_k e^{\eta_i} = 1 \Longrightarrow \varphi_k = \frac{1}{\sum_{j=1}^{k} e^{\eta_i}} \quad (21)$$

代入公式 19，得到：

$$\varphi_i = \frac{e^{\eta_i}}{\sum_{j=1}^{k} e^{\eta_i}} = \frac{e^{\eta_i}}{1+\sum_{j=1}^{k-1} e^{\eta_i}} \quad (22)$$

从而，我们就得到了连接函数，有了连接函数后，就可以把多项式分布的概率表达出来，即将公式 22 代入公式 11：

$$\mathrm{P(y=i)}=\varphi_i=\frac{e^{\eta_i}}{1+\sum_{j=1}^{k-1}e^{\eta_i}}=\frac{e^{\theta_i^T x}}{1+\sum_{j=1}^{k-1}e^{\theta_j^T x}} \tag{23}$$

注意到，上式中的每个参数η_i都是一个可用线性向量$\theta_i^T x$ 表示出来的，因而这里的θ其实是一个二维矩阵。

于是，我们可以得到假设函数 h 如下：

$$h_\theta(x)=E[T(y)|x;\theta]=E\begin{bmatrix}I\{\mathrm{y}=1\}\\I\{y=2\}\\\vdots\\I\{y=k-1\}\end{bmatrix}|x;\theta\Big]=\begin{bmatrix}\varphi_1\\\varphi_2\\\vdots\\\varphi_{k-1}\end{bmatrix} \tag{24}$$

公式 24 中代入公式 23 即可。

那么如何根据假设函数 h 求得参数θ，当然还是最大似然函数的方法，最大似然函数如下：

$$\ell(\theta)=\prod_{i=1}^{m}p(y^{(i)}|x^{(i)};\theta)=\prod_{i=1}^{m}\prod_{j=1}^{k}\varphi_j^{I\{y^{(i)}=j\}} \tag{25}$$

对公式 25 取对数，我们得到如下最大似然函数：

$$\mathrm{L}(\theta)=\sum_{i=1}^{m}\sum_{j=1}^{k}I\{y^{(i)}=j\}log\varphi_j \tag{26}$$

然后，将公式 22 代入公式 26 即可得到最大似然函数的对数，依此使用梯度下降算法或者牛顿方法求得参数后，使用假设函数 h 对新的样例进行预测，即可完成多类分类任务。这种多种分类问题的解法被称为 softmax regression。