C1-3

Machine Learning
by Andrew Ng, Stanford Engineering

Xiaojie Zhou

szxjzhou@163.com

2016.8.8

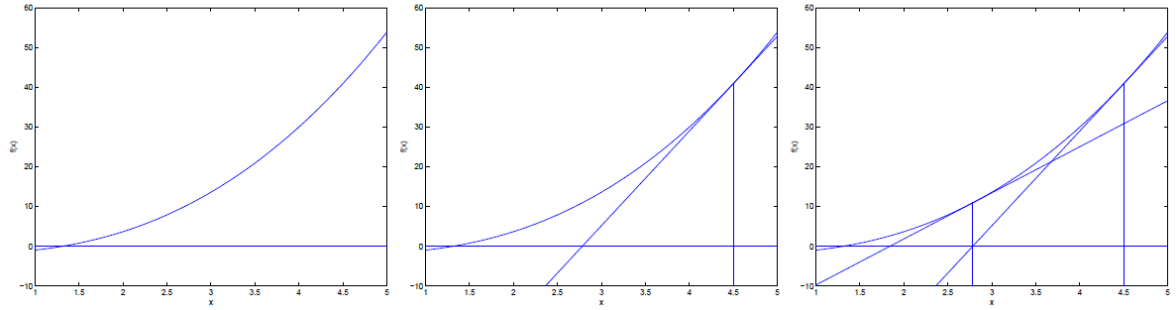
第四集：牛顿方法

- 牛顿方法(Newton’s Method)
- 指数分布族(Exponential Family)
- 广义线性模型(Generalized Linear Models(GLMs))

第四集：牛顿方法

- 牛顿方法
  - 之前我们讲到用梯度上升法解决Logistic回归的最优化问题（极大似然），下面将讲述另外一种方法
    - 对于某个实数域上的函数f，我们需要求f(θ)=0时对应的θ值，可以通过牛顿方法进行求解
      - 在牛顿方法中首先随机初始化θ的值θ0，此时我们会得到对应的函数值f(θ0)
      - 接下来我们在f(θ0)这一点处做切线，切线会与y=0交于某一点θ1，此时又会得到新的函数值f(θ1)。再在f(θ1)这一点处做切线...
      - 如此迭代下去直至收敛，即可得到f(θ)=0时对应的θ值

Here's a picture of the Newton's method in action:



$$\theta := \theta - \frac{f(\theta)}{f'(\theta)}$$

第四集：牛顿方法

- 牛顿方法
  - 牛顿方法给了我们一个求f(θ)=0时对应的θ值的工具
    - 回到我们的问题，我们需要求ℓ(θ)的极大值（ℓ(θ)作为一个凹函数，其极大值等于最大值），也就相当于求ℓ(θ)的导数为0时对应的θ值，由此我们可以通过下面这个更新方法迭代得到θ的值

$$\theta := \theta - \frac{\ell'(\theta)}{\ell''(\theta)}$$

    - 牛顿方法是一种收敛速度很快的方法（二次收敛，因为考虑了二阶导数即梯度的梯度，因此通常下一次迭代的误差率接近于上一次迭代误差率的平方），对于Logistic回归、广义线性模型(GLM)作用很大
    - 上面所讲述的更新办法针对θ为一维参数的情况，其实对于一般意义上的牛顿方法(Newton-Raphson Method)来说更新办法如下（该更新办法中的θ是一个向量）

$$\theta := \theta - H^{-1}\nabla_\theta \ell(\theta)$$

其中∇θℓ(θ)为ℓ(θ)的梯度，H为Hessian矩阵

$$H_{ij} = \frac{\partial^2 \ell(\theta)}{\partial \theta_i \partial \theta_j}$$

第四集：牛顿方法

- 牛顿方法
  - 比较牛顿方法和梯度下降法可以得出下面的结论
    - 两者在形式上相似，其中Hessian矩阵的逆就好比梯度下降法的学习率参数α
    - 牛顿法收敛速度相比梯度下降法更快（梯度下降法是一阶收敛，牛顿方法是二阶收敛）。这是因为牛顿方法中考虑了二阶导数即梯度的梯度，而梯度下降法只考虑了梯度，因此通常下一次迭代的误差率接近于上一次迭代误差率的平方
    - 由于海森矩阵的的逆在迭代中不断减小，起到自动控制逐渐缩小步长的效果
    - 牛顿法存在的缺点是每次迭代都要重新计算一次Hessian矩阵的逆，而计算Hessian矩阵的逆是比较困难的。因此不宜用于特征很多的建模优化（因为在这些问题中θ值的维数很大）

第四集：牛顿方法

- 指数分布族
  - 之前讲到了以线性回归的方法和Logistic回归的方法，这两种方法都涉及到对于y关于x的概率密度函数的定义方法
    - 在线性回归中我们假设的是$y=h(x)+\varepsilon$，其中$h(x)$为$\theta^T x$，ε符合正态分布，最终得到了最小二乘衡量的结果；而在Logistic回归中我们假设的是$y=h(x)$，符合二项分布，最终得到了Logistic回归的目标函数
  - 实际上这两个算法都是广义线性模型这一类算法的特例
    - 上面所说到的正态分布、二项分布这些概率分布都同属于指数分布族，对于指数分布族来说具有下面的形式

$$p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$$

其中y为自变量（也就是之前我们常用的x），η称为该分布的自然参数(natural parameter/canonical parameter)，T(y)被称为充分统计量(sufficient statistic)一般情况下等于y，由于η为参数，因此$\eta^T T(y)$一般也为实数。当我们固定下函数a,b,T后，我们可以将这个表达式看成以η为参数，y为自变量的一类概率分布。这里面改变了η以后就会得到一组新的概率分布。

第四集：牛顿方法

- 指数分布族
  - 现在来看通过指数分布族的通项公式如何得到正态分布、二项分布这些概率分布
    - 先来看二项分布，在二项分布中参数为φ，因变量y的取值要么为0要么为1
      - $p(y = 1; \varphi) = \varphi; p(y = 0; \varphi) = 1 - \varphi \rightarrow p(y; \varphi) = \varphi^y(1 - \varphi)^{1-y}$

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

第一个式子到第二个式子不是那么显然，其实对于两个式子同时取log很容易发现是相等的

      - 对比指数分布族的通项公式可得$\varphi = 1/(1 + e^{-\eta})$，这也就是之前的Sigmoid函数，这也就印证了之前选择Sigmoid函数作为Logistic回归中h(x)定义方法的合理性。同时我们还可得到：

$$\begin{aligned} T(y) &= y \\ a(\eta) &= -\log(1-\phi) \\ &= \log(1+e^{\eta}) \\ b(y) &= 1 \end{aligned}$$

$$\eta = \log\frac{\varphi}{1-\varphi} \Leftrightarrow -\eta = \log\frac{1-\varphi}{\varphi} = \log\left(\frac{1}{\varphi} - 1\right)$$

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

第四集：牛顿方法

- 指数分布族
  - 接下来我们来看一下高斯分布如何表达成类似的形式
    - 由于我们在线性回归里面得到最小二乘估计的过程中发现方差的取值与原优化问题无关，因此在这里我们也可以简单地将方差看成某个常数项（因此在下面我们将方差简单地设置成1，得到了如下的关系式）。此时发现这也是指数分布族的特例

$$
\begin{aligned}
p(y;\mu) &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(y-\mu)^2\right) \\
&= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y^2\right)\cdot\exp\left(\mu y-\frac{1}{2}\mu^2\right)
\end{aligned}
$$

$$
\begin{aligned}
\eta &= \mu \\
T(y) &= y \\
a(\eta) &= \mu^2/2 \\
&= \eta^2/2 \\
b(y) &= (1/\sqrt{2\pi})\exp(-y^2/2)
\end{aligned}
$$

    - 实际上就算加上方差σ也可以对应到指数分布族。由于涉及到两个参数，因此需要更加广义的指数分布族的定义$p(y;\eta,\tau) = b(a,\tau)\ \mathrm{Exp}[(\eta^T T(y) - a(\eta)) / c(\tau)]$，其中τ称为dispersion parameter，在这里对应到σ^2

第四集：牛顿方法

- 指数分布族
  - 其实还有很多常见的分布都可以对应到指数分布族
    - 多项式(Multinomial)分布：二项分布好比扔硬币要么为0要么为1，多项式分布相当于扔骰子，取值范围虽然也是有限数量元素集合，但不止于{0,1}
    - 泊松(Poisson)分布：主要用于描述单位时间、面积、体积等单位内事件个数的相对频率（比如说一年内中彩票的概率、一定时间内顾客的数量）
    - Gamma分布、指数(Exponential)分布：主要用于对连续非负随机变量进行建模，和泊松分布有一定类似性（比如说泊松过程的事件间隔时间服从指数分布，从头开始到第n次事件发生的间隔时间服从Gamma分布），只不过这个描述的是连续的变量。比如说在公交车站上等到下一辆车的时间的概率。
    - Beta分布、狄利克雷(Dirichlet)分布：主要用于对事件成功的概率进行建模（即对概率分布进行建模）。对于硬币或者骰子这样的简单实验，我们事先能很准确地掌握系统成功的概率。然而通常情况下，系统成功的概率是未知的。比如说特殊的硬币，我们需要知道这枚硬币正面朝上的概率。然后抛10次硬币，出现5次正面，于是我们认为硬币出现正面的概率p最可能是0.5。但是即使硬币出现正面的概率p为0.4，也会出现抛10次出现5次正面的情况。因此我们并不能完全确定硬币出现正面的概率就是0.5，所以p也是一个随机变量，它服从Beta分布。如果我们要确定的不只是硬币这种{0,1}二元取值的，而是骰子这种多元的，对应的事件成功的概率p服从狄利克雷分布

第四集：牛顿方法

- 广义线性模型
  - 下面来看如何在指数分布族的基础上得到对应的线性模型，这些模型都属于广义线性模型。对于广义线性模型来说需要满足下面三个前提条件
    - $y \mid x; \theta \sim ExponentialFamily(\eta)$：说明y关于参数η的分布满足指数分布族的定义（其中η由自变量x和参数θ通过某种运算得到，该运算在广义线性模型中是有一定条件的，该条件详见下面第三条）
    - 广义线性模型的目标是在给定x的情况下需要得到T(y)的期望，即E[T(y)|x]的值（T(y)为充分统计量，在统计学的角度看可以提供参数θ的全部信息，在通常情况下等于y值，因此我们相当于在给定x的情况下得到E[y|x]的值，也就是原分布的期望值）
      - 对照我们原来的线性回归和Logistic回归，我们发现在线性回归中我们要预测的h(θ)就是对应正态分布的均值$\theta^T x$，在Logistic回归中我们要预测的h(θ)就是E[y|x]。由此线性回归和Logistic回归都是满足这个的定义的
        - $p(y = 1|x; \theta) = 0 \cdot p(y = 0|x; \theta) + 1 \cdot p(y = 1|x; \theta) = E[y|x; \theta]$
    - 广义线性模型的自然参数η与自变量x满足线性关系，即$\eta=\theta^T x$
      - 如果η为向量，则$\eta_i=\theta_i^T x$

第四集：牛顿方法

- 广义线性模型
  - 满足上述三个条件的模型都可统称为广义线性模型，在广义线性模型中有很多好的性质（比如说学习起来速度快，比较高效）
  - 下面来看之前的Logistic回归是怎样对应到广义线性模型上的
    - 先来看Logistic回归是如何满足广义线性模型的三个条件的

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

$$
\begin{aligned}
T(y) &= y \\
a(\eta) &= -\log(1-\phi) \\
&= \log(1+e^{\eta}) \\
b(y) &= 1
\end{aligned}
\qquad
\begin{aligned}
h_\theta(x) &= E[y|x;\theta] \\
&= \phi \\
&= 1/(1+e^{-\eta}) \\
&= 1/(1+e^{-\theta^T x})
\end{aligned}
$$

在之前的例子中我们已经看过Logistic回归对应的二项分布属于指数分布族；左下图展示了Logistic回归定义的h(x)是符合E[y|x]的；同时在Logistic回归中$\eta=\theta^T x$，符合第三个条件

左下图用η代入其中的φ后得到的函数g(η)称为正则响应函数(canonical response function)，而g^{-1}称为正则关联函数(canonical link function)

第四集：牛顿方法

- 广义线性模型
  - 接下来看一下之前的线性回归是怎样对应到广义线性模型上的

$$
\begin{aligned}
p(y;\mu) &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(y-\mu)^2\right) \\
&= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}y^2\right)\cdot\exp\left(\mu y-\frac{1}{2}\mu^2\right)
\end{aligned}
\qquad
\begin{aligned}
\eta &= \mu \\
T(y) &= y \\
a(\eta) &= \mu^2/2 \\
&= \eta^2/2 \\
b(y) &= (1/\sqrt{2\pi})\exp(-y^2/2)
\end{aligned}
\qquad
\begin{aligned}
h_\theta(x) &= E[y|x;\theta] \\
&= \mu \\
&= \eta \\
&= \theta^T x.
\end{aligned}
$$

  - 由此我们可以总结出机器学习的一个简单的自动化方法
    - **先根据数据的特征选择对应的符合指数分布族的分布，然后根据这个分布通过计算E[T(y)|x]得到对应的式子f(φ)，用η代入其中的φ后得到正则响应函数g(η)，再将式子中的η替换为$\theta^T x$，即可得到对应的广义线性模型h(x)。接下来按照之前的办法写出对数似然方程，再通过梯度上升或者牛顿方法即可完成对于系数θ的求解**

第四集：牛顿方法

- 广义线性模型
  - 下面我们用这个思想来看一个复杂一点的例子，考虑一个多个离散变量的分类问题
    - 首先看到这个是个分类问题，说明最后结果的取值不是连续的而是离散的；而条件又告诉我们，最后的取值不是二元的，因此不能用二项分布。由于是多元的变量，因此我们选定我们的模型符合多项式分布
    - 接下来我们需要将多项式分布写成指数分布族的形式
      - 首先先看多项式分布是什么样子的。我们可以假设有k个离散的结果为{y1,y2,...,yk}，而这些结果发生的概率分别为φ1,φ2,...,φk-1,1-(φ1+φ2+...+φk-1)（避免过度参数化，多出一个多余参数φk），从而我们的参数集中只有k-1个元素{φ1,φ2,...,φk-1}，但是为了后面简单表示，仍然保留了φk这一写法。
      - 然后我们定义一个记号1{·}，在这个中括号里面是一个命题，当这个命题为真时该值为1，否则该值为0，即1{True} = 1, 1{False} = 0，例子： 1{2 = 3} = 0, 1{5-2 = 3} = 1

第四集：牛顿方法

- 广义线性模型
  - 续上需要将多项式分布写成指数分布族的形式
    - 借助上面的记号，我们可以将p(y;φ)表达成下面的通项形式

$$
\begin{aligned}
p(y;\phi) &= \phi_1^{1\{y=1\}}\phi_2^{1\{y=2\}}\cdots\phi_k^{1\{y=k\}} \\
&= \phi_1^{1\{y=1\}}\phi_2^{1\{y=2\}}\cdots\phi_k^{1-\sum_{i=1}^{k-1}1\{y=i\}}
\end{aligned}
$$

    - 接下来的问题是怎样将1{·}表示成可解释的函数形式，因此我们可以引入一个k-1*k维的矩阵M，其定义如下（仅mii上的元素为1，否则为0（i=1~k-1））

$$
\mathrm{M} = [m(1) \quad m(2) \quad \ldots \quad m(k-1) \quad m(k)] = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & & 0 & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & & 1 & 0 \end{bmatrix}
$$

从而我们可以将该矩阵与这一记号结合起来，得到(m(y))i = 1{y = i},

第四集：牛顿方法

- 广义线性模型
  - 续上需要将多项式分布写成指数分布族的形式

$$P(y;\phi) = \phi_1^{1\{y=1\}}\phi_2^{1\{y=2\}}\dots\phi_k^{1\{y=k\}} = \phi_1^{1\{y=1\}}\phi_2^{1\{y=2\}}\dots\phi_k^{1-\sum_{i=1}^{k-1}1\{y=i\}} = \phi_1^{(m(y))1}\phi_2^{(m(y))2}\dots\phi_k^{1-\sum_{i=1}^{k-1}(m(y))i}$$

$$= \exp\left[(m(y))1\log(\phi 1) + (m(y))2\log(\phi 2) + \dots + \left(1 - \sum_{i=1}^{k-1}(m(y))i\right)\log(\phi k)\right]$$

$$= \exp\left[(m(y))1\log\left(\phi 1/\phi k\right) + (m(y))2\log\left(\phi 2/\phi k\right) + \dots + (m(y))k-1\log\left(\phi k-1/\phi k\right) + \log(\phi k)\right]$$

$$= \exp\left[\begin{bmatrix}\log\left(\phi 1/\phi k\right) & \log\left(\phi 2/\phi k\right) & \dots & \log\left(\phi k-1/\phi k\right)\end{bmatrix}\begin{bmatrix}(m(y))1\\(m(y))2\\\vdots\\(m(y))k-1\end{bmatrix} + \log(\phi k)\right]$$

$$p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$$

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

对比指数分布族可以发现m(y)就是这里的T(y)，并且可以得到如右图所示的关系

第四集：牛顿方法

- 广义线性模型
  - 接下来我们需要计算E[T(y)|x]

$$
\begin{aligned}
h_\theta(x) &= \mathrm{E}[T(y)|x;\theta] \\
&= \mathrm{E}\left[\left.\begin{array}{c} 1\{y=1\} \\ 1\{y=2\} \\ \vdots \\ 1\{y=k-1\} \end{array}\right| x;\theta\right] \\
&= \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{k-1} \end{bmatrix}
\end{aligned}
$$

这里面利用了前面定义的记号1{·}和性质(m(y))i = 1{y = i}（由于在前面已经证实(m(y))i=(T(y))i，因此这也就等价于(T(y))i = 1{y = i}），可以不难推出E[(T(y))i] = P(y = i) = φi

接下来的问题就是将φ表示成η经过某种运算的结果，得到正则响应函数g(η)

$$
\eta = \begin{bmatrix} \log(\phi_1/\phi_k) \\ \log(\phi_2/\phi_k) \\ \vdots \\ \log(\phi_{k-1}/\phi_k) \end{bmatrix}
$$

第四集：牛顿方法

- 广义线性模型
  - 下面需要得到正则响应函数g(η)
    - 不失一般性地，我们可以作如下的假设

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

这里面的i=1,2,⋯,k，此时我们可以看到ηk=log1=0，不会影响结果的正确性

    - 从而可以得到

$$\begin{aligned} e^{\eta_i} &= \frac{\phi_i}{\phi_k} \\ \phi_k e^{\eta_i} &= \phi_i \\ \phi_k \sum_{i=1}^{k} e^{\eta_i} &= \sum_{i=1}^{k} \phi_i = 1 \end{aligned} \quad \longrightarrow \quad \phi_i = \frac{e^{\eta_i}}{\sum_{j=1}^{k} e^{\eta_j}}$$

第四集：牛顿方法

- 广义线性模型
  - 由此我们得到了正则响应函数，通过将式子中的η替换为$\theta^T x$，即可得到对应的广义线性模型h(x)和对数似然方程ℓ(θ)

$$
\begin{aligned}
h_\theta(x) &= \mathrm{E}[T(y)|x;\theta] \\
&= \mathrm{E}\left[\begin{array}{c} 1\{y=1\} \\ 1\{y=2\} \\ \vdots \\ 1\{y=k-1\} \end{array} \middle| x;\theta \right] \\
&= \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{k-1} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\exp(\theta_1^T x)}{\sum_{j=1}^k \exp(\theta_j^T x)} \\ \frac{\exp(\theta_2^T x)}{\sum_{j=1}^k \exp(\theta_j^T x)} \\ \vdots \\ \frac{\exp(\theta_{k-1}^T x)}{\sum_{j=1}^k \exp(\theta_j^T x)} \end{bmatrix}.
\end{aligned}
$$

在这里面只定义了φ1,φ2,⋯,φk-1的求解方法，而根据之前所说，φk可以通过1-(φ1+φ2+⋯+φk-1)得到

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

这种方法也被称为Softmax回归(Softmax Regression)，可以看成是Logistic回归的一种推广形式