机器学习第一次作业

2020年9月

1 线性回归扩展

课堂上，老师讲解了如何求$J(\mathbf{w}) = ||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2$的最优解，其中$\mathbf{X} \in \mathrm{R}^{n\times d}, \mathbf{y} \in \mathrm{R}^n, \mathbf{w} \in \mathrm{R}^d$。这个问题的最优解可以通过求解最优性方程$\nabla_{\mathbf{w}} J(\mathbf{w}) = \mathbf{0}$得到。

（1）该问题还可以通过牛顿法求解。请写出$J(\mathbf{w})$对$\mathbf{w}$的海森矩阵H的表达式，并证明$\mathbf{w} - H^{-1}\nabla_{\mathbf{w}} J(\mathbf{w})$就是最优解。

（2）在课堂上，我们仅考虑了目标变量y是标量值的情况。下面我们进行拓展，拓展到多元最小二乘法的求解。给定一个数据集$\left\{(x^{(i)}, y^{(i)})\right\}_{i=1}^n$，其中，$x^{(i)} \in \mathrm{R}^d$，$y^{(i)} \in \mathrm{R}^p$，$y^{(i)}$为一个向量，损失函数如下：

$$J(\Theta) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{p}\left(\left(\Theta^T x^{(i)}\right)_j - y_j^{(i)}\right)^2.$$

其中$\Theta \in \mathrm{R}^{d\times p}$，$\Theta$为所求解参数矩阵。通过求解$\nabla_\Theta J(\Theta) = \mathbf{0}$得到$\Theta$的解析解。

2 L1正则化与L2正则化比较

给定数据矩阵$\mathbf{X} \in \mathrm{R}^{n\times d}$，其中$n$为样本数，$d$为每条数据的维度。$\mathbf{y} \in \mathrm{R}^n$是每条数据的标注。为了简化推导过程，假设$\mathbf{X}^T\mathbf{X} = nI$，且在回归过程中不考虑偏差项。$\mathbf{X}$的第$i$列可以简记为$\mathbf{X}_{*i}$。

（1）使用L1正则化时，损失函数为$J_1(\mathbf{w}) = ||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2 + \lambda||\mathbf{w}||_1$ $(\lambda > 0)$。请将目标函数$J_1(\mathbf{w})$改写成$J_1(\mathbf{w}) = ||\mathbf{y}||_2^2 + \sum_{i=1}^d f(\mathbf{X}_{*i}, \mathbf{w}_i)$的形式。（即给出$f$的表达式，它的第一个参数是向量，第二个参数是标量。表达式中可以使用$\mathbf{y}$, 但不能使用其它列$\mathbf{X}_{*i}$）

（2）利用上一问的答案，给出$f(\mathbf{X}_{*i}, \mathbf{w}_i)$的最优解$\mathbf{w}_i^* = 0$的充要条件。

（3）将第一问中的L1正则化换成L2正则化，损失函数为$J_2(\mathbf{w}) = ||\mathbf{X}\mathbf{w} - \mathbf{y}||_2^2 + \lambda||\mathbf{w}||_2^2$ $(\lambda > 0)$。加入L2正则化后，最优解记为$\mathbf{w}^\#$。仿照前两问的流程，给出$\mathbf{w}^\#$的第i个元素为0($\mathbf{w}_i^\# = 0$)的充要条件。

（4）如果一个向量的大部分元素都为0，那么它就是稀疏向量。根据前三问的回答，L1正则化的最优解$\mathbf{w}^*$和L2正则化的最优解$\mathbf{w}^\#$，哪一个更可能是稀疏的？为什么？

3 局部加权逻辑斯特回归

在这个问题中，我们研究的是局部加权形式的逻辑斯特回归。给定n个样本$\left\{\left(x^{(i)}, y^{(i)}\right)\right\}_{i=1}^{n}$，其中，$x^{(i)} \in \mathrm{R}^d$。我们对每个训练样本给予不同的权重，对下面的损失函数$\ell(\theta)$利用**牛顿法**进行优化（给出伪代码即可，牛顿法每次步长为1）：

$$\ell(\theta)=\frac{\lambda}{2} \theta^T \theta-\sum_{i=1}^n w^{(i)}\left[y^{(i)} \log h_\theta\left(x^{(i)}\right)+\left(1-y^{(i)}\right) \log \left(1-h_\theta\left(x^{(i)}\right)\right)\right]$$

其中$w^{(i)}=\exp \left(-\frac{\left\|x-x^{(i)}\right\|^2}{2 \tau^2}\right)$，$\theta$为模型参数，$h_\theta(x)=g\left(\theta^T x\right)=\frac{1}{1+e^{-\theta^T x}}$。