习题二

151220023, 段建辉, djhbarca@163.com

2017 年 4 月 13 日

1 [10pts] Lagrange Multiplier Methods

请通过拉格朗日乘子法(可参见教材附录B.1)证明《机器学习》教材中式(3.36)与式(3.37)等价。即下面公式(1.1)与(1.2)等价。

$$\begin{aligned} \min_{\mathbf{w}} \quad & -\mathbf{w}^{\mathrm{T}}\mathbf{S}_b\mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}^{\mathrm{T}}\mathbf{S}_w\mathbf{w} = 1 \end{aligned} \tag{1.1}$$

$$\mathbf{S}_b\mathbf{w} = \lambda\mathbf{S}_w\mathbf{w} \tag{1.2}$$

Proof. 现在假设函数 $\varphi_1 = -\mathbf{w}^{\mathrm{T}}\mathbf{S}_b\mathbf{w} \quad \varphi_2 = \mathbf{w}^{\mathrm{T}}\mathbf{S}_w\mathbf{w} - 1$
那么再继续假设函数 $\phi = \varphi_1 + \lambda\varphi_2 = -\mathbf{w}^{\mathrm{T}}\mathbf{S}_b\mathbf{w} + \lambda(\mathbf{w}^{\mathrm{T}}\mathbf{S}_w\mathbf{w} - 1)$
那么对该函数针对w进行求导，并令导数为0得到：

$$\bigtriangledown\phi = -\mathbf{S}_b\mathbf{w} + \lambda\mathbf{S}_w\mathbf{w} = 0$$

即为：

$$\mathbf{S}_b\mathbf{w} = \lambda\mathbf{S}_w\mathbf{w}$$

在这个时候，是在函数ϕ取得最小值的时候倒数为0，即为(1.1) 与 (1.2)等价

□

2 [20pts] Multi-Class Logistic Regression

教材的章节3.3介绍了对数几率回归解决二分类问题的具体做法。假定现在的任务不再是二分类问题，而是多分类问题，其中$y \in \{1, 2 \ldots, K\}$。请将对数几率回归算法拓展到该多分类问题。

(1) **[10pts]** 给出该对率回归模型的“对数似然”(log-likelihood);

(2) **[10pts]** 计算出该“对数似然”的梯度。

Solution. (1)

根据课本上的二分类对数似然我们可以得到：

$$\ell(\mathbf{w_i}, b_i) = \sum_{i=1}^{m} \ln p(y_i|\mathbf{x}_i; \mathbf{w_i}, b_i), \text{其中} i \in \{1, 2\ldots, K\}$$

假设该多分类问题满足如下$K-1$个对数几率，

$$\begin{aligned}
\ln \frac{p(y=1|\mathbf{x})}{p(y=K|\mathbf{x})} &= \mathbf{w}_1^{\mathrm{T}}\mathbf{x} + b_1 \\
\ln \frac{p(y=2|\mathbf{x})}{p(y=K|\mathbf{x})} &= \mathbf{w}_2^{\mathrm{T}}\mathbf{x} + b_2 \\
&\cdots \\
\ln \frac{p(y=K-1|\mathbf{x})}{p(y=K|\mathbf{x})} &= \mathbf{w}_{K-1}^{\mathrm{T}}\mathbf{x} + b_{K-1}
\end{aligned}$$

那么我们有:

$$p(y=i|\mathbf{x}) = p(y=K|\mathbf{x})e^{\mathbf{w}_i^{\mathrm{T}}\mathbf{x}+b_i} \quad \text{其中} i \in \{1, 2\ldots, K-1\}$$

令 $\beta = (\mathbf{w_i};b_i)$, $\alpha = (\mathbf{x};\, 1)$,则$\mathbf{w}_i^{\mathrm{T}}\mathbf{x} + b_i$ 可以简写为 $\beta_i^{\mathrm{T}}\alpha$ 其中$i \in \{1, 2\ldots, K-1\}$

再令$p_K(\alpha;\beta) = p(y = K|\alpha;\beta)$ 同样令$p_i(\alpha;\beta) = p(y = i|\alpha;\beta) = p(y = K|\mathbf{x})e^{\mathbf{w}_i^{\mathrm{T}}\mathbf{x}+b_i}$,其中$i \in \{1, 2\ldots, K-1\}$ 因为为多分类问题，那么如果样例分类到一个类别中，其余类别的y值均为0，那么定义指示函数$\mathbb{I}(\cdot)$:

$$\mathbb{I}(y=j) = \begin{cases} 1 & \text{若}y\text{等于}j \\ 0 & \text{若}y\text{不等于}j \end{cases}$$

则将原始的似然函数重写为：

$$p(y_j|x_i; \mathbf{w_i}, b_i) = \sum_{i=1}^{m}\sum_{j=1}^{K} \mathbb{I}(y_i = j)p_j(y_i = j|\alpha_i;\beta)$$

最后根据似然函数以及对率几率回归可以得到：

$$\begin{aligned}
\ell(\beta) &= \sum_{i=1}^{m} \ln\left(\sum_{j=1}^{K} \mathbb{I}(y_i = j)p_j(y_i = j|\alpha_i;\beta)\right) \\
&= \sum_{i=1}^{m} \ln\left(\sum_{j=1}^{K-1} \mathbb{I}(y = j)p(y = K|\mathbf{x})e^{\beta^{\mathrm{T}}\alpha_i} + \mathbb{I}(y = K)p(y = K|\mathbf{x})\right)
\end{aligned}$$

即最小化：

$$\ell(\beta) = \sum_{i=1}^{m}\left(-\sum_{j=1}^{K-1} \mathbb{I}(y_i = j)\beta^{\mathrm{T}}\alpha_i + \ln(1 + e^{\beta^{\mathrm{T}}\alpha_K})\right)$$

(2)

求梯度即为令上述似然函数对β求导，得到：

$$\nabla_\beta \ell(\beta) = -\sum_{i=1}^{m}\left[\alpha_i\left(\mathbb{I}(y_i = j) - p_j(y_i = j|\alpha_i;\beta)\right)\right]$$

另一种形式为:

$$\nabla_\beta \ell(\beta) = -\sum_{i=1}^{m}\left[\alpha_i\left(\mathbb{I}(y_i=j) - \frac{e^{\beta^{\mathbf{T}}\alpha_i}}{1+e^{\beta^{\mathbf{T}}\alpha_i}}\right)\right]$$

即为最后的梯度表达式。

3 [35pts] Logistic Regression in Practice

对数几率回归(Logistic Regression, 简称LR)是实际应用中非常常用的分类学习算法。

(1) **[30pts]** 请编程实现二分类的LR, 要求采用牛顿法进行优化求解, 其更新公式可参考《机器学习》教材公式(3.29)。详细编程题指南请参见链接：`http://lamda.nju.edu.cn/ml2017/PS2/ML2_programming.html`

(2) **[5pts]** 请简要谈谈你对本次编程实践的感想(如过程中遇到哪些障碍以及如何解决, 对编程实践作业的建议与意见等)。

Solution. (1)
见代码。
(2)
遇到的问题：
1、最大的问题就是数据的归一化，因为数据差距很大，直接进行迭代就会在p函数中溢出,因此需要进行统一，但是在归一化的过程中很容易将某些值变得非常非常小，从而出现二阶导矩阵不可逆的情况，所以一开始的分类很差劲，准确率一直在60%左右
2、建议：希望能够再给一些提示以及可能出现的问题...毕竟大二的学弟还是有点吃力的...谢谢助教学长!
3、代码借鉴：`https://sanlo.github.io/2016/11/13/Introduction-to-quantitative-investment-and-`
中的Newton迭代，因为一开始写的牛顿迭代法总是有问题，而且会overflow很快，因此借鉴了此处的牛顿迭代法。

4 [35pts] Linear Regression with Regularization Term

给定数据集$D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_m, y_m)\}$, 其中$\mathbf{x}_i = (x_{i1}; x_{i2}; \cdots; x_{id}) \in \mathbb{R}^d$,

$y_i \in \mathbb{R}$, 当我们采用线性回归模型求解时, 实际上是在求解下述优化问题：

$$\hat{\mathbf{w}}^*_{\mathbf{LS}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2, \tag{4.1}$$

其中, $\mathbf{y} = [y_1, \cdots, y_m]^{\mathrm{T}} \in \mathbb{R}^m, \mathbf{X} = [\mathbf{x}_1^{\mathrm{T}}; \mathbf{x}_2^{\mathrm{T}}; \cdots; \mathbf{x}_m^{\mathrm{T}}] \in \mathbb{R}^{m \times d}$, 下面的问题中, 为简化求解过程, 我们暂不考虑线性回归中的截距(intercept)。

在实际问题中, 我们常常不会直接利用线性回归对数据进行拟合, 这是因为当样本特征很多, 而样本数相对较少时, 直接线性回归很容易陷入过拟合。为缓解过拟合问题, 常对公式(4.1)引入正则化项, 通常形式如下：

$$\hat{\mathbf{w}}^*_{\mathbf{reg}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\Omega(\mathbf{w}), \tag{4.2}$$

其中, $\lambda > 0$为正则化参数, $\Omega(\mathbf{w})$是正则化项, 根据模型偏好选择不同的Ω。

下面, 假设样本特征矩阵$\mathbf{X}$满足列正交性质, 即$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{I}$, 其中$\mathbf{I} \in \mathbb{R}^{d \times d}$是单位矩阵, 请回答下面的问题(需要给出详细的求解过程)：

(1) **[5pts]** 考虑线性回归问题, 即对应于公式(4.1), 请给出最优解$\hat{\mathbf{w}}^*_{\mathbf{LS}}$的闭式解表达式;

(2) **[10pts]** 考虑岭回归(ridge regression)问题, 即对应于公式(4.2)中$\Omega(\mathbf{w}) = \|\mathbf{w}\|_2^2 = \sum_{i=1}^d w_i^2$时, 请给出最优解$\hat{\mathbf{w}}^*_{\mathbf{Ridge}}$的闭式解表达式;

(3) **[10pts]** 考虑LASSO问题, 即对应于公式(4.2)中$\Omega(\mathbf{w}) = \|\mathbf{w}\|_1 = \sum_{i=1}^d |w_i|$时, 请给出最优解$\hat{\mathbf{w}}^*_{\mathbf{LASSO}}$的闭式解表达式;

(4) **[10pts]** 考虑ℓ_0-范数正则化问题,

$$\hat{\mathbf{w}}^*_{\ell_0} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_0, \tag{4.3}$$

其中, $\|\mathbf{w}\|_0 = \sum_{i=1}^d \mathbb{I}[w_i \neq 0]$,即$\|\mathbf{w}\|_0$表示$\mathbf{w}$中非零项的个数。通常来说, 上述问题是NP-Hard问题, 且是非凸问题, 很难进行有效地优化得到最优解。实际上, 问题(3)中的LASSO可以视为是近些年研究者求解ℓ_0-范数正则化的凸松弛问题。

但当假设样本特征矩阵$\mathbf{X}$满足列正交性质, 即$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{I}$时, ℓ_0-范数正则化问题存在闭式解。请给出最优解$\hat{\mathbf{w}}^*_{\ell_0}$的闭式解表达式, 并简要说明若去除列正交性质假设后, 为什么问题会变得非常困难？

Solution. (1)

设$\mathbf{E}_{(w)} = \frac{1}{2}\|\mathbf{y_i} - \mathbf{X}\mathbf{w}\|_2^2$,因为为$L_2$范数，那么对$\mathbf{w}$求导得到：

$$\begin{aligned}
\frac{\partial \mathbf{E}_{(w)}}{\partial w} &= \frac{\partial \mathbf{E}_{(w)}}{\partial(\mathbf{y} - \mathbf{X}\mathbf{w})}\frac{\partial(\mathbf{y} - \mathbf{X}\mathbf{w})}{\partial(\mathbf{w})} \\
&= -\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w})
\end{aligned}$$

令该式为零即可得到闭式解表达式为：$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

(2)

设$\mathbf{E}_{(w)} = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2$,因为为$L_2$范数，那么对$\mathbf{w}$求导得到：

$$\begin{aligned}
\frac{\partial \mathbf{E}_{(w)}}{\partial w} &= \frac{\partial \mathbf{E}_{(w)}}{\partial(\mathbf{y} - \mathbf{X}\mathbf{w})}\frac{\partial(\mathbf{y} - \mathbf{X}\mathbf{w})}{\partial(\mathbf{w})} + 2\lambda\mathbf{w} \\
&= -\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + 2\lambda\mathbf{w}
\end{aligned}$$

令该式为零即可得到闭式解表达式为：$\mathbf{w} = (2\lambda\mathbf{I} + \mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$

(3)

设$\mathbf{E}_{(w)} = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1$，因为为$L_1$范数，$L_1$范数不可微，但是存在次微分，那么有：

$$\frac{\partial}{\partial \mathbf{w}}\|\mathbf{w}\|_1 = sign(\mathbf{w})$$

其中$sign(\mathbf{w})$的表示如下：

$$sign(\mathbf{w}) = \begin{cases} +1 & \mathbf{w} > 0 \\ -1 & \mathbf{w} < 0 \\ \ \ 0 & \mathbf{w} = 0 \end{cases}$$

那么对$\mathbf{w}$求梯度得到：

$$\begin{aligned} \frac{\partial \mathbf{E}_{(w)}}{\partial w} &= \frac{\partial \mathbf{E}_{(w)}}{\partial(\mathbf{y} - \mathbf{X}\mathbf{w})}\frac{\partial(\mathbf{y} - \mathbf{X}\mathbf{w})}{\partial(\mathbf{w})} + sign(\mathbf{w}) \\ &= -\mathbf{X}^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + sign(\mathbf{w}) \end{aligned}$$

则令该式为零即可得到闭式解表达式为：$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{y} - sign(\mathbf{w}))$

(4)

设$\mathbf{E}_{(w)} = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_0$