习题二

141180016, 丁俊峰, 141180016@smail.nju.edu.cn

2017 年 4 月 17 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.
根据拉格朗日乘子法，可将带约束的式（1.1）转换为不带约束的:

$$L(\mathbf{w}, \lambda) = -\mathbf{w}^T\mathbf{S}_b\mathbf{w} + \lambda(\mathbf{w}^T\mathbf{S}_w\mathbf{w} - 1) \tag{1.3}$$

然后对(1.3)式对于$\mathbf{w}$和λ偏导，得到极值点:

$$\frac{\partial L}{\partial \mathbf{w}} = -\frac{\partial(\mathbf{w}^T\mathbf{S}_b\mathbf{w} - \lambda\mathbf{w}^T\mathbf{S}_w\mathbf{w})}{\partial \mathbf{w}} = 0 \tag{1.4}$$

根据The Matrix Cookbook中公式(81):$\frac{\partial \mathbf{x}^T\mathbf{B}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{B} + \mathbf{B}^T)\mathbf{x}$，式(1.4)：

$$\frac{\partial(\mathbf{w}^T\mathbf{S}_b\mathbf{w} - \lambda\mathbf{w}^T\mathbf{S}_w\mathbf{w})}{\partial \mathbf{w}} = (\mathbf{S}_b + \mathbf{S}_b^T)\mathbf{w} - \lambda(\mathbf{S}_w + \mathbf{S}_w^T)\mathbf{w} = 2(\mathbf{S}_b\mathbf{w} - \lambda\mathbf{S}_w\mathbf{w}) = 0 \tag{1.5}$$

由(1.5)可以得到$\mathbf{S}_b\mathbf{w} = \lambda\mathbf{S}_w\mathbf{w}$

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。

提示1：假设该多分类问题满足如下$K-1$个对数几率，

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

提示2：定义指示函数$\mathbb{I}(\cdot)$，

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

Solution.

(1) 由提示（1）可知该模型的概率分布为

$$p(y=1|\mathbf{x})=\frac{e^{\beta_1^T\mathbf{x}}}{1+\sum\limits_{k=1}^{K-1}e^{\beta_k^T\mathbf{x}}}$$

$$p(y=2|\mathbf{x})=\frac{e^{\beta_2^T\mathbf{x}}}{1+\sum\limits_{k=1}^{K-1}e^{\beta_k^T\mathbf{x}}}$$

$$\cdots$$

$$p(y=K-1|\mathbf{x})=\frac{e^{\beta_K^T-1\mathbf{x}}}{1+\sum\limits_{k=1}^{K-1}e^{\beta_k^T\mathbf{x}}}$$

$$p(y=K|\mathbf{x})=\frac{1}{1+\sum\limits_{k=1}^{K-1}e^{\beta_k^T\mathbf{x}}}$$

所以对数似然为

$$\begin{aligned}
\ell(\beta)&=\sum_{i=1}^{m}ln(p(y_i|\mathbf{x}_i;\beta))\\
&=\sum_{i=1}^{m}ln\left(\left(\frac{1}{1+\sum\limits_{k=1}^{K-1}e^{\beta_k^T\mathbf{x}_i}}\right)^{\mathbb{I}(y_i=K)}\times\prod_{j=1}^{K-1}\left(\frac{e^{\beta_j^T\mathbf{x}_i}}{1+\sum\limits_{k=1}^{K-1}e^{\beta_k^T\mathbf{x}_i}}\right)^{\mathbb{I}(y_i=j)}\right)\\
&=\sum_{i=1}^{m}\left(\sum_{j=1}^{K-1}\left(\mathbb{I}(y_i=j)[\beta_j^T\mathbf{x}_i-ln(1+\sum_{k=1}^{K-1}e^{\beta_k^T\mathbf{x}_i})]\right)-\mathbb{I}(y_i=K)ln(1+\sum_{k=1}^{K-1}e^{\beta_k^T\mathbf{x}_i})\right)\\
&=\sum_{i=1}^{m}\left(\sum_{j=1}^{K-1}\mathbb{I}(y_i=j)\beta_j^T\mathbf{x}_i-ln(1+\sum_{k=1}^{K-1}e^{\beta_k^T\mathbf{x}_i})\right)
\end{aligned}$$

(2) 梯度为

$$\frac{\partial \ell(\beta)}{\partial \beta_l} = \sum_{i=1}^{m} \left(\mathbb{I}(y_i = l)\mathbf{x}_i - \frac{\mathbf{x}_i e^{\beta_l^T \mathbf{x}_i}}{1 + \sum\limits_{k=1}^{K-1} e^{\beta_k^T \mathbf{x}_i}} \right) (l = 1, 2, 3...K-1)$$

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) 代码见压缩包内main.py

(2) 这次编程作业中遇到了牛顿法迭代时出现奇异矩阵无法求逆的问题，通过断点调试，发现问题出现在数据没有预处理进行归一化上，导致指数系数过大出现0，使得矩阵不可逆。后来通过对训练数据归一化解决了这个问题。

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)对于式(4.1)关于$\mathbf{w}$求导:

$$\begin{aligned}
\bigtriangledown_{\mathbf{w}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 &= \frac{1}{2} \bigtriangledown_{\mathbf{w}} (\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y}) \\
&= \frac{1}{2} \bigtriangledown_{\mathbf{w}} (\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} - \mathbf{y}^T\mathbf{X}\mathbf{w} - \mathbf{w}^T\mathbf{X}^T\mathbf{y} + \mathbf{y}^T\mathbf{y}) \\
&= \frac{1}{2} \bigtriangledown_{\mathbf{w}} (\mathbf{w}^T\mathbf{w} - \mathbf{y}^T\mathbf{X}\mathbf{w} - \mathbf{w}^T\mathbf{X}^T\mathbf{y}) \\
&= \frac{1}{2} \bigtriangledown_{\mathbf{w}} (tr(\mathbf{w}^T\mathbf{w}) - tr(\mathbf{y}^T\mathbf{X}\mathbf{w}) - tr(\mathbf{w}^T\mathbf{X}^T\mathbf{y})) \\
&= \frac{1}{2}(2\mathbf{w} - \bigtriangledown_{\mathbf{w}} tr(\mathbf{w}^T\mathbf{X}^T\mathbf{y}) - \bigtriangledown_{\mathbf{w}} tr(\mathbf{w}^T\mathbf{X}^T\mathbf{y})) \\
&= \mathbf{w} - \bigtriangledown_{\mathbf{w}} tr(\mathbf{w}^T\mathbf{X}^T\mathbf{y}) \\
&= \mathbf{w} - \mathbf{X}^T\mathbf{y} \\
&= 0
\end{aligned}$$

得到$\hat{\mathbf{w}}^*_{\mathbf{LS}} = \mathbf{X}^T\mathbf{y}$

(2))对于式（4.2）关于$\mathbf{w}$求导:

$$\begin{aligned}
\bigtriangledown_{\mathbf{w}}(\frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\mathbf{w}^T\mathbf{w}) &= \frac{1}{2}(2\mathbf{w} - 2\mathbf{X}^T\mathbf{y}) + \lambda \bigtriangledown_{\mathbf{w}} tr(\mathbf{w}^T\mathbf{w}) \\
&= \mathbf{w} - \mathbf{X}^T\mathbf{y} + \lambda 2\mathbf{w} \\
&= (2\lambda + 1)\mathbf{w} - \mathbf{X}^T\mathbf{y} \\
&= 0
\end{aligned}$$

得到$\hat{\mathbf{w}}^*_{\mathbf{Ridge}} = \frac{\mathbf{X}^T\mathbf{y}}{2\lambda+1}$

(3)

$$
\begin{aligned}
\bigtriangledown_{\mathbf{w}}(\frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2+\lambda\|\mathbf{w}\|_1) &= \bigtriangledown_{\mathbf{w}}(\frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2+\lambda\sum_{i=1}^{d}|w_i|) \\
&= \frac{1}{2}(2\mathbf{w}-2\mathbf{X}^T\mathbf{y})+\lambda\bigtriangledown_{\mathbf{w}}(\sum_{i=1}^{d}|w_i|) \\
&= \mathbf{w}-\mathbf{X}^T\mathbf{y}+\lambda sign(\mathbf{w}) \\
&= 0
\end{aligned}
$$

得到$\hat{\mathbf{w}}^*_{\mathbf{LASSO}}=\mathbf{X}^T\mathbf{y}-\lambda sign(\mathbf{w})$

令$\mathbf{m}=\mathbf{X}^T\mathbf{y}$,则$w_i=m_i-\lambda sign(w_i)$

因为

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

所以$w_i>0$时，即$w_i=m_i-\lambda>0$，即$m_i>\lambda$;

$w_i<0$时，即$w_i=m_i+\lambda<0$，即$m_i<-\lambda$;

当$|m_i|\leq\lambda$时，$w_i=0$

所以最终

$$
\hat{w_i}^*(i=1,2,3,...,d)=\begin{cases} m_i-\lambda & (m_i>\lambda) \\ m_i+\lambda & (m_i<\lambda) \\ 0 & (|m_i|\leq\lambda) \end{cases}
$$

(4)

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

令$m=X^Ty$

$$
(E_{w_i})=\begin{cases} \lambda+\frac{1}{2}{w_i}^2-w_im_i & (w_i\neq 0) \\ 0 & (w_i=0) \end{cases}
$$

$w_i\neq 0$时，E_{w_i}是一个二次函数，$w_i=m_i$时最小值$(E_{w_i})_{min}=\lambda-\frac{1}{2}(m_i)^2$ 所以

$$
(w_i)^*_{l0}(i=1,2,3,...,d)\begin{cases} m_i & (m_i\geqslant\sqrt{2\lambda}) \\ 0 & (m_i<\sqrt{2\lambda}) \end{cases}
$$

当$\mathbf{X}^{\mathrm{T}}\mathbf{X}\neq\mathbf{I}$时，$E=\frac{1}{2}(y^Ty-y^TXw-w^TX^Ty+wX^TXw)+\lambda\|\mathbf{w}\|_0$，此时无法将$wX^TXw$中的$w_i$分离出来，使得求解困难。