习题二

学号：141130077
作者姓名：邱梓豪
邮箱：2957606241@qq.com

2017 年 4 月 12 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. 此处用于写证明(中英文均可)
由拉格朗日乘子法可知，（1.1）的优化问题可以转化为下面的拉格朗日函数的优化问题：

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

对$\mathbf{w}$求偏导，可得：

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

又因为 $\mathbf{S}_b$=$(\mu_0 - \mu_1)(\mu_0 - \mu_1)^{\mathrm{T}}$，所以$\mathbf{S}_b$=$\mathbf{S}_b^{\mathrm{T}}$
同理 $\mathbf{S}_w$=$\sum_{x\in x_0}(x-\mu_0)(x-\mu_0)^{\mathrm{T}}$+$\sum_{x\in x_1}(x-\mu_1)(x-\mu_1)^{\mathrm{T}}$，所以$\mathbf{S}_w$=$\mathbf{S}_w^{\mathrm{T}}$
所以（1.4）可以化为：

$$0 = -2\mathbf{S}_b\mathbf{w} + 2\lambda\mathbf{S}_w\mathbf{w} \tag{1.5}$$

即：

$$\mathbf{S}_b\mathbf{w} = \lambda\mathbf{S}_w\mathbf{w} \tag{1.6}$$

□

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$y \in \{1, 2 \dots, K\}$。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，

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

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. 此处用于写解答(中英文均可)

（1）假设这个问题的训练集为$\{(\mathbf{x}_i, y_i) : i = 1, 2 \dots, N\}$，$y_i \in \{1, 2 \dots, K\}$。为了方便起见，这里我假设该多分类问题满足如下K-1个对数几率（与提示不同）：

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

则有：

$$p(y=k|x) = \frac{e^{\mathbf{w}_K^{\mathrm{T}}\mathbf{x}}}{\sum_i e^{\mathbf{w}_i^{\mathrm{T}}\mathbf{x}}} \tag{2.1}$$

所以其对数似然函数为:

$$\begin{aligned}
L_i(\mathbf{w}_1, \mathbf{w}_1, \dots, \mathbf{w}_K) &= \sum_{i=1}^{N} \ln p(y_i|\mathbf{x}, \mathbf{w}) \\
&= \sum_{i=1}^{N} \ln \prod_{j=1}^{K} (\frac{e^{\mathbf{w}_j^{\mathrm{T}}\mathbf{x}_i}}{\sum_k e^{\mathbf{w}_k^{\mathrm{T}}\mathbf{x}_i}})^{\mathbb{I}(y_i=j)} \\
&= \sum_{i=1}^{N} \sum_{j=1}^{K} \mathbb{I}(y_i=j) \ln \frac{e^{\mathbf{w}_j^{\mathrm{T}}\mathbf{x}_i}}{\sum_k e^{\mathbf{w}_k^{\mathrm{T}}\mathbf{x}_i}}
\end{aligned}$$

这就是该模型的对数似然，其中：

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

（2）令$\frac{e^{\mathbf{w}_j^{\mathrm{T}}\mathbf{x}_i}}{\sum_k e^{\mathbf{w}_k^{\mathrm{T}}\mathbf{x}_i}}$=$t_j(x_i)$，则有：

$$\frac{\partial \ln t_j(x_i)}{\partial \mathbf{w}_k}=\begin{cases}[1-t_k(x)]x & 若j等于k\\ -t_k(x)x & 若j不等于k\end{cases}$$

所以有：

$$\frac{\sum_{j=1}^{K}\mathbb{I}(y=j)\ln t_j(x)}{\partial \mathbf{w}_k}=(\mathbb{I}(y=k)-t_k(x))x \tag{2.2}$$

所以该对数似然的在$\mathbf{w}_k$上的导数为：

$$\frac{\partial L_i(\mathbf{w}_1,\mathbf{w}_1,\ldots,\mathbf{w}_k)}{\partial \mathbf{w}_k}=\sum_{i=1}^{N}(\mathbb{I}(y_i=k)-t_k(x_i))x_i \tag{2.3}$$

故该对数似然的梯度为：

$$[\sum_{i=1}^{N}(\mathbb{I}(y_i=1)-t_1(x_i))x_i,\sum_{i=1}^{N}(\mathbb{I}(y_i=2)-t_2(x_i))x_i,\ldots,\sum_{i=1}^{N}(\mathbb{I}(y_i=N)-t_N(x_i))x_i] \tag{2.4}$$

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. 此处用于写解答(中英文均可)

（1）本题代码见压缩包中

（2）问题：在本次作业中我使用的是python语言。在处理向量的相加和数乘时，由于对python语言一些数据结构操作特性不太了解，导致了一些问题，比如我开始使用python内置的list表示一个向量，我原来认为list1+list2就可以表示两个向量相加，但最后发现这原来是将list2接到list1之后。后来我将list转换成了numpy中提供的向量类型，解决了这个问题。

建议：我建议在出编程题时可以明确代码的输入输出接口，比如本题中可以说输入是当前文件夹下的2个csv文件，输出是10个csv文件。

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. 此处用于写解答(中英文均可)

（1）原式可化为：

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

上式对$\mathbf{w}$求导得：

$$\frac{dE}{d\mathbf{w}} = 2\mathbf{X}^T(\mathbf{X}\mathbf{w} - \mathbf{y}) \tag{4.4}$$

令上式为0，可得：

$$\mathbf{w}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{4.5}$$

又因为$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{I}$，所以上式可以写成：

$$\mathbf{w}^* = \mathbf{X}^T\mathbf{y} \tag{4.6}$$

（2）对于岭回归，其要优化的问题形式如下：

$$E(\mathbf{w}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2 = \frac{1}{2}(\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y}) + \lambda\mathbf{w}^T\mathbf{w} \tag{4.7}$$

为了求得$\mathbf{w}$的最优解，上式对$\mathbf{w}$求微分得：

$$\begin{aligned}
\frac{dE}{d\mathbf{w}} &= \frac{1}{2}[\frac{d}{d\mathbf{w}}(\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y})] + \lambda\frac{d}{d\mathbf{w}}\mathbf{w}^T\mathbf{w} \\
&= \frac{1}{2}[\mathbf{X}^T(\mathbf{X}\mathbf{w} - \mathbf{y}) + \mathbf{X}^T(\mathbf{X}\mathbf{w} - \mathbf{y})] + \lambda\mathbf{w} \\
&= \mathbf{X}^T(\mathbf{X}\mathbf{w} - \mathbf{y}) + \lambda\mathbf{w} \\
&= \mathbf{X}^T\mathbf{X}\mathbf{w} - \mathbf{X}^T\mathbf{y} + \lambda\mathbf{w} = 0
\end{aligned}$$

于是可得：

$$\mathbf{w}^* = (\mathbf{X}^T\mathbf{X} + \lambda I)^{-1}\mathbf{X}^T\mathbf{y} \tag{4.8}$$

（3）在LASSO问题中，其要优化的问题形式如下：

$$\begin{aligned}
E(\mathbf{w}) &= \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1 \\
&= \frac{1}{2}(\mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{X}\mathbf{w} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}) + \lambda\|\mathbf{w}\|_1 \\
&= \frac{1}{2}\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\mathbf{w} + \frac{1}{2}\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}) + \lambda|\mathbf{w}|
\end{aligned}$$

为了求得$\mathbf{w}$的最优解，上式应该对$\mathbf{w}$求微分，此时$\frac{1}{2}\mathbf{y}^T\mathbf{y}$可忽略，又因为假设样本特征矩阵$\mathbf{X}$满足列正交性质, 即$\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{I}$，所以上式可写成如下形式：

$$E(\mathbf{w}) = -\mathbf{y}^T\mathbf{X}\mathbf{w} + \frac{1}{2}\mathbf{w}^2 + \lambda|\mathbf{w}| \tag{4.9}$$

注意到用最小二乘法得到的结果为$\hat{\mathbf{w}}^*_{\mathbf{LS}} = \mathbf{X}^T\mathbf{y}$，所以上式又可以写为：

$$E(\mathbf{w}) = -\sum_{i=1}^{m}\hat{\mathbf{w}}^*_{\mathbf{LS}i}\mathbf{w}_i + \frac{1}{2}\mathbf{w}_i^2 + \lambda|\mathbf{w}_i| \tag{4.10}$$

固定一个i，目标就转变为最小化下式：

$$E(\mathbf{w}_i) = -\hat{\mathbf{w}}^*_{\mathbf{LS}i}\mathbf{w}_i + \frac{1}{2}\mathbf{w}_i^2 + \lambda|\mathbf{w}_i| \tag{4.11}$$

显然为了使上式获得最小值，$\hat{\mathbf{w}}^*_{\mathbf{LS}i}$和$\mathbf{w}_i$必须同号，所以可以分以下两种情况进行讨论：
Case 1: $\hat{\mathbf{w}}^*_{\mathbf{LS}i}$>0，则此时$\mathbf{w}_i$>0，对（4.11）中的$\mathbf{w}_i$求导，并令其为0可得：

$$\mathbf{w}_i = \hat{\mathbf{w}}^*_{\mathbf{LSi}} - \lambda \tag{4.12}$$

所以有：

$$\hat{\mathbf{w}}^*_{\mathbf{lasso}i} = |\hat{\mathbf{w}}^*_{\mathbf{LSi}}| - \lambda \tag{4.13}$$

Case2: $\hat{\mathbf{w}}^*_{\mathbf{LS}i} \leq 0$，则此时$\mathbf{w}_i \leq 0$，对（4.11）中的$\mathbf{w}_i$求导，并令其为0可得：

$$\mathbf{w}_i = \hat{\mathbf{w}}^*_{\mathbf{LSi}} + \lambda \tag{4.14}$$

所以有：

$$\hat{\mathbf{w}}^*_{\mathbf{lasso}i} = -(|\hat{\mathbf{w}}^*_{\mathbf{LSi}}| - \lambda) \tag{4.15}$$

综上所述，最后的解可写成如下形式：

$$\hat{\mathbf{w}}^*_{\mathbf{lasso}i} = sgn(\hat{\mathbf{w}}^*_{\mathbf{LS}i})(|\hat{\mathbf{w}}^*_{\mathbf{LSi}}| - \lambda) \tag{4.16}$$

其中sgn(x)为符号函数。
（4）