习题二

141220132, 银琦, 141220132@smail.nju.edu.cn

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.
令

$$f(\mathbf{w}) = -\mathbf{w}^{\mathrm{T}}\mathbf{S}_b\mathbf{w}$$

$$g(\mathbf{w}) = \mathbf{w}^{\mathrm{T}}\mathbf{S}_w\mathbf{w} - 1$$

要使$f(\mathbf{w})$最小且同时满足$g(\mathbf{w}) = 0$的约束，由拉格朗日乘子法，可定义拉格朗日函数为：

$$L(\mathbf{w}, \lambda) = f(\mathbf{w}) + \lambda g(\mathbf{w})$$

对其$\mathbf{w}$求偏导数$\nabla_{\mathbf{w}}L(\mathbf{w}, \lambda)$可得到：

$$\nabla_{\mathbf{w}}L(\mathbf{w}, \lambda) = -(\mathbf{S}_b + \mathbf{S}_b^{\mathrm{T}})\mathbf{w} + \lambda(\mathbf{S}_w + \mathbf{S}_w^{\mathrm{T}})\mathbf{w}$$

根据$\mathbf{S}_b$和$\mathbf{S}_w$的定义可知$\mathbf{S}_b$与$\mathbf{S}_w$均为对称矩阵，所以

$$\mathbf{S}_b + \mathbf{S}_b^{\mathrm{T}} = \mathbf{S}_b + \mathbf{S}_b = 2\mathbf{S}_b$$

$$\mathbf{S}_w + \mathbf{S}_w^{\mathrm{T}} = \mathbf{S}_w + \mathbf{S}_w = 2\mathbf{S}_w$$

即

$$\nabla_{\mathbf{w}}L(\mathbf{w}, \lambda) = -2\mathbf{S}_b\mathbf{w} + 2\lambda\mathbf{S}_w\mathbf{w}$$

将其对$\mathbf{w}$的偏导数$\nabla_{\mathbf{w}}L(\mathbf{w}, \lambda)$置零可得到

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

即

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

□

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$y \in \{1, 2 \dots, K\}$。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.

(1)根据提示1，可以求出$p(y=t|\mathbf{x})(t=1,2,\cdots,K-1)$和$p(y=K|\mathbf{x})$如下：

$$p(y=t|\mathbf{x}) = \frac{e^{\mathbf{w}_t^{\mathrm{T}}\mathbf{x}+b_t}}{1+\sum_{k=1}^{K-1} e^{\mathbf{w}_k^{\mathrm{T}}\mathbf{x}+b_k}} (t=1,2,\cdots,K-1)$$

$$p(y=K|\mathbf{x}) = \frac{1}{1+\sum_{k=1}^{K-1} e^{\mathbf{w}_k^{\mathrm{T}}\mathbf{x}+b_k}}$$

令$\beta_i = (\mathbf{w}_i; b_i), \hat{\mathbf{x}} = (\mathbf{x}; 1)$，则$\mathbf{w}_i^{\mathrm{T}}\mathbf{x} + b_i$可简写为$\beta_i^{\mathrm{T}}\hat{\mathbf{x}}$，由课本公式3.25可写出：

$$\ell(\beta_1, \beta_2, \cdots, \beta_K) = \sum_{i=1}^{m} \sum_{j=1}^{K} \mathbb{I}(y=j) ln p(y_i|\mathbf{x}_i; \beta_j)$$

将$p(y=t|\mathbf{x})(t=1,2,\cdots,K-1)$和$p(y=K|\mathbf{x})$代入上式可得：

$$\ell(\beta_1, \beta_2, \cdots, \beta_K) = \sum_{i=1}^{m} \sum_{j=1}^{K-1} \mathbb{I}(y=j) ln \frac{e^{\beta_j^{\mathrm{T}}\hat{\mathbf{x}}_i}}{1+\sum_{k=1}^{K-1} e^{\beta_k^{\mathrm{T}}\hat{\mathbf{x}}_i}} + \sum_{i=1}^{m} \mathbb{I}(y=K) ln \frac{1}{1+\sum_{k=1}^{K-1} e^{\beta_k^{\mathrm{T}}\hat{\mathbf{x}}_i}}$$

化简可得：

$$\ell(\beta_1, \beta_2, \cdots, \beta_K) = \sum_{i=1}^{m} \sum_{j=1}^{K-1} \mathbb{I}(y=j) \beta_j^{\mathrm{T}}\hat{\mathbf{x}}_i - \sum_{i=1}^{m} \sum_{j=1}^{K} \mathbb{I}(y=j) ln(1+\sum_{k=1}^{K-1} e^{\beta_k^{\mathrm{T}}\hat{\mathbf{x}}_i})$$

最大化上式，即为最小化：

$$\ell(\beta_1, \beta_2, \cdots, \beta_K) = -\sum_{i=1}^{m} \sum_{j=1}^{K-1} \mathbb{I}(y=j) \beta_j^{\mathrm{T}}\hat{\mathbf{x}}_i + \sum_{i=1}^{m} \sum_{j=1}^{K} \mathbb{I}(y=j) ln(1+\sum_{k=1}^{K-1} e^{\beta_k^{\mathrm{T}}\hat{\mathbf{x}}_i})$$

(2)对每一个β_i求偏导，即：

$$\nabla_{\beta_i}\ell(\beta_1,\beta_2,\cdots,\beta_K)=\frac{\partial}{\partial\beta_i}[-\sum_{i=1}^{m}\sum_{j=1}^{K-1}\mathbb{I}(y=j)\beta_j^{\mathrm{T}}\hat{\mathbf{x}}_i+\sum_{i=1}^{m}\sum_{j=1}^{K}\mathbb{I}(y=j)ln(1+\sum_{k=1}^{K-1}e^{\beta_k^{\mathrm{T}}\hat{\mathbf{x}}_i})]$$

$$=-\sum_{i=1}^{m}\mathbb{I}(y=j)\hat{\mathbf{x}}_i+\sum_{i=1}^{m}\mathbb{I}(y=j)\frac{\hat{\mathbf{x}}_i\sum_{k=1}^{K-1}e^{\beta_k^{\mathrm{T}}\hat{\mathbf{x}}_i}}{1+\sum_{k=1}^{K-1}e^{\beta_k^{\mathrm{T}}\hat{\mathbf{x}}_i}}$$

$$=-\sum_{i=1}^{m}\mathbb{I}(y=j)\hat{\mathbf{x}}_i+\sum_{i=1}^{m}\frac{\hat{\mathbf{x}}_i e^{\beta_j^{\mathrm{T}}\hat{\mathbf{x}}_i}}{1+\sum_{k=1}^{K-1}e^{\beta_k^{\mathrm{T}}\hat{\mathbf{x}}_i}}$$

$$=-\sum_{i=1}^{m}\hat{\mathbf{x}}_i(\mathbb{I}(y=j)-p(y=j|x_i))$$

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. 一开始觉得毫无头绪，加上对matlab不熟悉，无从下手，于是在网上查找了相关资料，进行学习，主要的参考网站为：`http://blog.csdn.net/icefire_tyh/article/details/52068844`。

在编写过程中，仍然遇到了不少问题：

(1)开始先实现牛顿迭代法的算法，没有进行交叉验证，按照参考资料，将迭代终止条件设置为当前的ℓ与上一次ℓ之差小于一个很小的数，但是会导致二阶导矩阵不可逆，然后将终止条件改成了对迭代次数的控制；在求逆矩阵时，了解到pinv函数比inv函数的精度要高，于是使用pinv函数求逆矩阵同样可以避免出现上述情况。

(2)当牛顿迭代法实现后要划分数据集，进行10折交叉验证，一开始忘记了在每次验证前将迭代次数重新置0，导致精度很低，修改后精度还是很低，进行数据集划分是调用了函数crossvalind，我认为fold中的输出顺序不影响精度，所以暂时没找到原因，猜测可能是data与targets没有对应。后来进行了手动划分数据集，将data中的数据平均分成10份，第一次循环中第一份为测试集，后九份为训练集，后九次循环同理，经测试后，迭代轮数为5时可以取得精度最高：0.9625。

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)令$E_{\mathbf{w}} = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$，对$\mathbf{w}$求导得到

$$\frac{\partial E_{\mathbf{w}}}{\partial \mathbf{w}} = \mathbf{X}^\mathrm{T}\mathbf{X}\mathbf{w} - \mathbf{X}^\mathrm{T}\mathbf{y}$$

令上式为零可得$\hat{\mathbf{w}}^*_{\mathbf{LS}}$的最优解。因为题中条件征矩阵$\mathbf{X}$满足列正交性质, 即$\mathbf{X}^\mathrm{T}\mathbf{X} = \mathbf{I}$，所以可求得：

$$\hat{\mathbf{w}}^*_{\mathbf{LS}} = (\mathbf{X}^\mathrm{T}\mathbf{X})^{-1}\mathbf{X}^\mathrm{T}\mathbf{y} = \mathbf{X}^\mathrm{T}\mathbf{y}$$

(2)当$\Omega(\mathbf{w}) = \|\mathbf{w}\|_2^2 = \sum_{i=1}^d w_i^2$时，原式可化为

$$\hat{\mathbf{w}}^*_{\mathbf{Ridge}} = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2$$

令$E_{\mathbf{w}}=\frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2+\lambda\|\mathbf{w}\|_2^2$，对$\mathbf{w}$求导得到

$$\frac{\partial E_{\mathbf{w}}}{\partial \mathbf{w}}=\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}-\mathbf{X}^{\mathrm{T}}\mathbf{y}+2\lambda\mathbf{w}$$

令上式为零可得$\hat{\mathbf{w}}^*_{\mathbf{Ridge}}$的最优解：

$$\hat{\mathbf{w}}^*_{\mathbf{Ridge}}=(\mathbf{X}^{\mathrm{T}}\mathbf{X}+2\lambda)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}=(\mathbf{I}+2\lambda\mathbf{I})\mathbf{X}^{\mathrm{T}}\mathbf{y}=(1+2\lambda)\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

(3)当$\Omega(\mathbf{w})=\|\mathbf{w}\|_1=\sum_{i=1}^d|w_i|$时,原式可化为

$$\hat{\mathbf{w}}^*_{\mathbf{LASSO}}=\arg\min_{\mathbf{w}}\frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2+\lambda\|\mathbf{w}\|_1$$

令$E_{\mathbf{w}}=\frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2+\lambda\|\mathbf{w}\|_1=\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{y}-\mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbf{w}+\frac{1}{2}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}^{\mathrm{T}}\mathbf{w}+\lambda\|\mathbf{w}\|_1$ 由(1)可得$\hat{\mathbf{w}}^*_{\mathbf{LS}}=\mathbf{X}^{\mathrm{T}}\mathbf{y}$，所以$E_{\mathbf{w}}$可化为：

$$E_{\mathbf{w}}=\sum_{i=1}^d-\hat{w}^*_{LS\_i}w_i+\frac{1}{2}w_i^2+\lambda|w_i|+\frac{1}{2}y_i^2$$

要使得$E_{\mathbf{w}}$取最小值，$\hat{w}^*_{LS_i}$和w_i一定是同正负的(w_i可以等于0)，下面分情况讨论：

(a)若$\hat{w}^*_{LS_i}>0$且$w_i\geq 0$，则去掉绝对值符号可得

$$E_{\mathbf{w}}=\sum_{i=1}^d-\hat{w}^*_{LS\_i}w_i+\frac{1}{2}w_i^2+\lambda w_i+\frac{1}{2}y_i^2$$

对其求导可得：

$$\frac{\partial E_{\mathbf{w}}}{\partial w_i}=-\hat{w}^*_{LS\_i}+w_i+\lambda$$

令上式等于零，可得到$w_i=\hat{w}^*_{LS_i}-\lambda$

(b)若$\hat{w}^*_{LS_i}<0$且$w_i\leq 0$，则去掉绝对值符号可得

$$E_{\mathbf{w}}=\sum_{i=1}^d-\hat{w}^*_{LS\_i}w_i+\frac{1}{2}w_i^2-\lambda w_i+\frac{1}{2}y_i^2$$

对其求导可得：

$$\frac{\partial E_{\mathbf{w}}}{\partial w_i}=-\hat{w}^*_{LS\_i}+w_i-\lambda$$

令上式等于零，可得到$w_i=\hat{w}^*_{LS_i}+\lambda$

综上可得

$$w_i=\begin{cases}\hat{w}^*_{LS\_i}-\lambda & \hat{w}^*_{LS\_i}>\lambda\\ 0 & |\hat{w}^*_{LS\_i}|\leq\lambda\\ \hat{w}^*_{LS\_i}+\lambda & \hat{w}^*_{LS\_i}<-\lambda\end{cases}$$

说明：课本254页有使用近端梯度下降求解ℓ_1的闭式解的过程。

(4)最后一项若为0，则与(1)中结果一样；若不为0，应该为一常数，所以求导计算最小值时扔则与(1)中结果一样，即：

$$\hat{\mathbf{w}}^*_{\ell_0}=\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

去除列正交性质后，ℓ_0正则会导致函数不光滑，不连续，优化方法不容易计算。