习题二

参考解答

2017 年 4 月 25 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.

记优化目标为$f(\mathbf{w}) = -\mathbf{w}^{\mathrm{T}}\mathbf{S}_b\mathbf{w}$，约束为$g(\mathbf{w}) = \mathbf{w}^{\mathrm{T}}\mathbf{S}_w\mathbf{w} - 1$，则公式(1.1)等价于一个等式约束的优化问题：

寻找$\mathbf{w}$的最优取值$\mathbf{w}^*$，使目标函数$f(\mathbf{w})$最小且同时满足$g(\mathbf{w}) = 0$的等式约束。

这类问题可以采用标准的拉格朗日乘子法来求解，将等式约束的优化问题转化为一个无约束的优化问题：

寻找最优点$\mathbf{w}^*$，使得梯度$\nabla f(\mathbf{w}^*)$和$\nabla g(\mathbf{w}^*)$方向平行。即存在λ使得

$$\nabla f(\mathbf{w}^*) + \lambda\nabla g(\mathbf{w}^*) = 0 \tag{1.3}$$

代入展开得：

$$-2\mathbf{S}_b\mathbf{w} + 2\lambda\mathbf{S}_w\mathbf{w} = 0 \tag{1.4}$$

即得到公式(1.2) □

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$\mathbb{I}(\cdot)$，

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

Solution. 由提示可知，

$$p(y=k|\mathbf{x})=\begin{cases}\dfrac{e^{\mathbf{w}_k^{\mathrm{T}}\mathbf{x}+b_k}}{1+\sum_{i=1}^{K-1}e^{\mathbf{w}_i^{\mathrm{T}}\mathbf{x}+b_i}}, \text{ if } k\le K-1 \\ \dfrac{1}{1+\sum_{i=1}^{K-1}e^{\mathbf{w}_i^{\mathrm{T}}\mathbf{x}+b_i}}, \text{ if } k=K\end{cases} \tag{2.1}$$

由此可得对数似然如下，

$$\begin{aligned}
\ell(\mathbf{w},\mathbf{b}) &= \sum_{i=1}^{m}\sum_{j=1}^{K}\mathbb{I}(y_i=j)\ln p(y_i=j|\mathbf{x}_i) && (2.2)\\
&= \sum_{i=1}^{m}\sum_{j=1}^{K-1}\mathbb{I}(y_i=j)(\mathbf{w}_j^{\mathrm{T}}\mathbf{x}+b_j)-\sum_{i=1}^{m}\ln(1+\sum_{i=1}^{K-1}e^{\mathbf{w}_i^{\mathrm{T}}\mathbf{x}+b_i}) && (2.3)
\end{aligned}$$

为便于讨论，另$\boldsymbol{\beta}_j=(\mathbf{w}_j;b_j)$，$\hat{\mathbf{x}}=(\mathbf{x};1)$，从而计算出该对数似然的梯度如下，

$$\frac{\partial\ell(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}_j}=\sum_{i=1}^{m}(\mathbb{I}(y_i=j)-p(y_i=j|\hat{\mathbf{x}}_i))\hat{\mathbf{x}}_i \tag{2.4}$$

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.

关于编程题，一些常见的问题和回答(FAQ)将更新在网站中，请参看链接：`http://lamda.nju.edu.cn/ml2017/ml_faq.html`

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) 由《机器学习》书第三章公式(3.11)可知, $\hat{\mathbf{w}}^*_{\mathbf{LS}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$.

(2) $\hat{\mathbf{w}}^*_{\mathbf{Ridge}} = \arg\min_{\mathbf{w}} F(\mathbf{w}) = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2$, 由于$F(\mathbf{w})$关于$\mathbf{w}$是凸的且可微，因此只需要取$\mathbf{w} = \hat{\mathbf{w}}^*_{\mathbf{Ridge}}$时，使得

$$\frac{\partial F(\mathbf{w})}{\partial \mathbf{w}} = -\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{w}) + 2\lambda\mathbf{w} = 0.$$

因此可知,

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

(3) $\hat{\mathbf{w}}^*_{\mathbf{LASSO}} = \arg\min_{\mathbf{w}} F(\mathbf{w}) = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_1$,由于$F(\mathbf{w})$关于$\mathbf{w}$是凸的,但不可微。

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

记$f(\mathbf{w}) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} - \mathbf{U}\mathbf{w} + \lambda\|\mathbf{w}\|_1 = \sum_{i=1}^d (\frac{1}{2}w_i^2 - u_i w_i + \lambda|w_i|) = \sum_{i=1}^d f_i(w_i)$, 其中$\mathbf{U} = \mathbf{y}^{\mathrm{T}}\mathbf{X} = [u_1, \cdots, u_d] \in \mathbb{R}^{1\times d}$, 对于任意$i = 1, \cdots, d$, 通过对$w_i$的符号进行讨论可知:

$$[\hat{\mathbf{w}}^*_{\mathbf{LASSO}}]_i = \mathbf{Shrink}_\lambda(u_i) = \begin{cases} u_i - \lambda & u_i > \lambda, \\ 0 & u_i \in [-\lambda, \lambda], \\ u_i + \lambda & u_i < -\lambda. \end{cases} \tag{4.4}$$

(4) $\hat{\mathbf{w}}^*_{\ell_0} = \arg\min_{\mathbf{w}} F(\mathbf{w}) = \arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_0$,由于$F(\mathbf{w})$关于$\mathbf{w}$是非凸的，且不连续，不可微。

$$\begin{aligned} F(\mathbf{w}) &= \frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_0 \\ &= \frac{1}{2}(\mathbf{y}^{\mathrm{T}}\mathbf{y} - 2\mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbf{w} + \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w}) + \lambda\|\mathbf{w}\|_0 \\ &= \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\mathbf{X}\mathbf{w} + \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + \lambda\|\mathbf{w}\|_0 \end{aligned}$$

记$f(\mathbf{w}) = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} - \mathbf{U}\mathbf{w} + \lambda\|\mathbf{w}\|_0 = \sum_{i=1}^d (\frac{1}{2}w_i^2 - u_i w_i + \lambda\mathbb{I}[w_i \neq 0]) = \sum_{i=1}^d f_i(w_i)$, 其中$\mathbf{U} = \mathbf{y}^{\mathrm{T}}\mathbf{X} = [u_1, \cdots, u_d] \in \mathbb{R}^{1\times d}$, 对于任意$i = 1, \cdots, d$, 通过对$w_i$是否为0讨论可知:

$$[\hat{\mathbf{w}}^*_{\ell_0}]_i = \begin{cases} u_i & u_i > \sqrt{2\lambda}, \\ 0 & u_i \in [-\sqrt{2\lambda}, \sqrt{2\lambda}], \\ u_i & u_i < -\sqrt{2\lambda}. \end{cases} \tag{4.5}$$

如果去掉列正交性质之后, $\mathbf{w}$展开后会出现类似$w_i w_j$的耦合项, 会导致无法对w_i拆开进行逐项最小化, 因此使得整体优化难度骤增。(言之有理即可)

Remark.

机器学习是一门实践和理论并重的学科，对于数学的要求很高。本题有一定的难度，比较考察数学积累，尤其考察了矩阵运算以及基本优化技巧。从本题中，依次由(1)-(4)，分别是简单凸优化问题，可微凸优化，不可微凸优化以及非凸优化。

对于凸优化问题，如果可微，直接利用梯度(gradient)信息进行优化；如果不可微，则可以使用次梯度(sub-gradient)信息。因此，对于(4)中非凸优化问题，标准的做法是将其拆开为d项进行逐项优化，这是建立在列正交性质上的。而对于一般的非凸优化，目前还没有很好的解决手段。