# 习题二

151242041, 王昊庭, hatsuyukiw@gmail.com

2017 年 4 月 11 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 Lagrange 乘子法, 上述优化问题可以转化为无约束优化问题,

$$L(\mathbf{w}, \lambda) = \min_{\mathbf{w}} \quad -\mathbf{w}^{\mathrm{T}}\mathbf{S}_b\mathbf{w} + \lambda(\mathbf{w}^{\mathrm{T}}\mathbf{S}_w\mathbf{w} - 1)\,.$$

上式对 $\mathbf{w}$ 和 $\lambda$ 求梯度, 有,

$$\begin{aligned} \frac{\partial L(\mathbf{w}, \lambda)}{\partial \mathbf{w}} &= -2\mathbf{S}_b\mathbf{w} + 2\lambda\mathbf{S}_w\mathbf{w} = 0\,, \\ \frac{\partial L(\mathbf{w}, \lambda)}{\partial \lambda} &= \mathbf{w}^{\mathrm{T}}\mathbf{S}_w\mathbf{w} - 1 = 0\,. \end{aligned}$$

注意到上述两等式的解就是原约束问题的解, 并且第一式等价于公式 (3.37). 故而现在只需要证明第一式蕴含第二式.

反设当 $\mathbf{S}_b\mathbf{w} = \lambda\mathbf{S}_w\mathbf{w}$ 时, $\mathbf{w}^{\mathrm{T}}\mathbf{S}_w\mathbf{w} - 1 \neq 0$, 那么选取足够大 (小) 的 $\lambda$ 可以使该优化问题无界. 又由于只有当 $\mathbf{S}_b\mathbf{w} = \lambda\mathbf{S}_w\mathbf{w}$ 时才可能达到该优化问题的最优解, 故而该问题没有最优解, 但是我们已知问题有解, 矛盾.

综上所述, 原优化问题与 $\mathbf{S}_b\mathbf{w} = \lambda\mathbf{S}_w\mathbf{w}$ 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。

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.** 定义 $\beta_i=(\mathbf{w}_i\ b_i)$, $\mathbf{x}'=(\mathbf{x}\ 1)$. 以下推导从一个与提示 1 略有不同但是更加对称的假设出发.

假设该问题满足如下的 $K$ 个对数几率, 其中 $c$ 为待定的系数.

$$\begin{aligned}
\ln p(y=1|\mathbf{x}') &= c\beta_1^{\mathrm{T}}\mathbf{x}' \\
\ln p(y=2|\mathbf{x}') &= c\beta_2^{\mathrm{T}}\mathbf{x}' \\
&\cdots \\
\ln p(y=K|\mathbf{x}') &= c\beta_K^{\mathrm{T}}\mathbf{x}'
\end{aligned}$$

由于概率的实际意义存在如下归一化条件.

$$\sum_{i=1}^{K} p(y=i|\mathbf{x})=1\,.$$

由此可以解出各概率,

$$p(y=i|\mathbf{x}')=\frac{e^{\beta_i^{\mathrm{T}}\mathbf{x}'}}{\sum_{k=1}^{K} e^{\beta_k^{\mathrm{T}}\mathbf{x}'}}\,.$$

该模型的对数似然为,

$$\begin{aligned}
l(\beta,\mathbf{x}') &= \sum_{i=1}^{m}\ln p(y_i|\mathbf{x}_\mathbf{i}') \\
&= \sum_{i=1}^{m}\ln\left(\frac{e^{\beta_{y_i}^{\mathrm{T}}\mathbf{x}_\mathbf{i}'}}{\sum_{k=1}^{K} e^{\beta_k^{\mathrm{T}}\mathbf{x}'}}\right)
\end{aligned}$$

该模型的对数似然的梯度为,

$$\frac{\partial l(\beta,\mathbf{x}')}{\beta}_{\ j}=\sum_{i=1}^{m}\left(x_i(\mathbb{I}(y_i==j)-p(y_i=j|\mathbf{x}_\mathbf{i}'))\right)\,.$$

注意到这个作者给出的模型的对数似然和对数似然的梯度, 与教材中的二分类模型的对应的函数是形式一致的.

## 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.** (2) 遇到的障碍是运算中容易出现 Underflow 和 Overflow. 最容易出现问题的地方是 Sigmoid 函数和 Hessian 矩阵求逆.

我采取的办法是频繁手动检测. 一个更好的办法是在训练 (优化) 过程中进行检测, 当准确率开始降低时停止训练 (优化). 众所周知, 牛顿法是以平方级别向最优解收敛的, 我的经验是当优化过程出现数值问题时, 一般在之前就已经达到了最优解.

我在 macOS 10.12, python 3.6.1, numpy 1.12.1, scikit-learn 0.18.1 上进行的实验, `evaluate.py` 似乎对`\n` 的支持不好. 我使用`\r\n` 换行之后 `evaluate.py` 才可以正确运行.

## 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 $\Omega$。

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 $\ell_0$-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 $\ell_0$-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, $\ell_0$-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. 根据教材公式 (3.11), $\hat{\mathbf{w}}^*_{\mathbf{LS}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$.

2. 改写目标函数为

$$\frac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda\mathbf{w}^{\mathrm{T}}\mathbf{w}\,,$$

对 $\mathbf{w}$ 求导可以得到

$$\begin{aligned}\hat{\mathbf{w}}^{\mathbf{Ridge}} &= (\mathbf{X}^{\mathrm{T}}\mathbf{X} + 2\lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \\ &= ((2\lambda+1)\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \\ &= \frac{1}{2\lambda+1}\mathbf{X}^{\mathrm{T}}\mathbf{y}\,.\end{aligned}$$

3. 事实上此时目标函数是不可导的, 但是我们忽略 $\mathbf{w}_i = 0$ 的情况, 规定 $\frac{\mathrm{d}|x|}{\mathrm{d}x} = \mathrm{sgn}(x)$. 对目标函数求导并使其等于 0, 得

$$\hat{\mathbf{w}}_j^{\mathbf{LASSO}} + \lambda\mathrm{sgn}(\hat{\mathbf{w}}_j^{\mathbf{LASSO}}) + \sum_{i=1}^{M} x_{ij}y_i = 0\,.$$

对 $\hat{\mathbf{w}}_j^{\mathbf{LASSO}}$ 的正负性分类讨论, 易见

$$\hat{\mathbf{w}}_j^{\mathbf{LASSO}} = \hat{\mathbf{w}}_j^{\mathbf{LS}} \max\left(0, 1 - \frac{\lambda}{|\hat{\mathbf{w}}_j^{\mathbf{LS}}|}\right)\,,$$

其中 $\hat{\mathbf{w}}^{\mathbf{LS}}$ 表示 (1) 中的解, 即朴素线性回归的解.

这个解的另一个更清晰的等价形式是,

$$\hat{\mathbf{w}}_j^{\mathbf{LASSO}} = \begin{cases}\hat{\mathbf{w}}_j^{\mathbf{LS}} - \lambda\mathrm{sgn}(\hat{\mathbf{w}}_j^{\mathbf{LS}}) & \lambda < |\hat{\mathbf{w}}_j^{\mathbf{LS}}| \\ 0 & \lambda \geq |\hat{\mathbf{w}}_j^{\mathbf{LS}}|\end{cases}$$

4. 将目标函数展开, 有

$$\begin{aligned}J(\mathbf{w}) &= \frac{1}{2}(\mathbf{y} - \mathbf{X}\mathbf{w})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda\|\mathbf{w}\|_0 \\ &= \frac{1}{2}\left(\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\mathbf{w} - 2\mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{y}\right) + \lambda\|\mathbf{w}\|_0 \\ &= \frac{1}{2}\left(\mathbf{w}^{\mathrm{T}}\mathbf{w} - 2\mathbf{w}^{\mathrm{T}}\hat{\mathbf{w}}^{LS} + \mathbf{y}^{\mathrm{T}}\mathbf{y}\right) + \lambda\|\mathbf{w}\|_0\,.\end{aligned}$$

注意到最后一步的推导使用了 (1) 中的结论.

将目标函数按照 $\mathbf{w}$ 的分量分离如下,

$$J(\mathbf{w}_j) = \frac{1}{2}\mathbf{w}_j^2 - \mathbf{w}_j\hat{\mathbf{w}}_j^{LS} + \lambda\mathbb{I}(\mathbf{w}_j \neq 0)\,.$$

注意到 $J(\mathbf{w}) = \sum_{j=1}^{d} J(\mathbf{w}_j) + \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{y}$, 我们只需要分别优化 $J(\mathbf{w}_j)$.

每一个 $J(\mathbf{w}_j)$ 是一个可能存在一个断点的二次函数, 最优解如下,

$$\hat{\mathbf{w}}_j^{\ell_0} = \begin{cases} \hat{\mathbf{w}}_j^{\mathbf{LS}} & \lambda < \frac{1}{2}(\hat{\mathbf{w}}_j^{\mathbf{LS}})^2 \\ 0 & \lambda \geq \frac{1}{2}(\hat{\mathbf{w}}_j^{\mathbf{LS}})^2 \end{cases}$$

5. 当 $\mathbf{X}$ 满足列正交性时, 设 $\mathbf{X}$ 的第 $i$ 列为 $\mathbf{X}_i$, 则目标函数为,

$$\begin{aligned} \|\mathbf{y} - \sum_{i=1}^{d} w_i\mathbf{X}_i\|^2 &= \|\mathbf{y}\|^2 - 2\mathbf{y}\cdot(\sum_{i=1}^{d} w_i\mathbf{X}_i) + \|\sum_{i=1}^{d} w_i\mathbf{X}_i\|^2 \\ &= \|\mathbf{y}\|^2 - 2\mathbf{y}\cdot(\sum_{i=1}^{d} w_i\mathbf{X}_i) + \sum_{i=1}^{d}\|w_i\mathbf{X}_i\|^2 \end{aligned}$$

注意最后一步推导使用了 $\mathbf{X}_i$ 之间的正交性. 此时目标函数没有交错项, 所以可以将目标函数按照 $\mathbf{w}$ 的各个分量拆开, 分别求解, 此时这个多变量最优化问题转化为若干个单变量最优化问题, 因此更容易求解.

直观地说, 列正交性保证了 $\mathbf{w}$ 的各个分量产生的 loss 互相不影响.

(按照这个思路, 四小题均可以使用统一的形式求解, 不过这个作者做到第四小题才意识到列正交性的作用, 所以也就不改了.)