习题二

141210016, 刘冰楠, bingnliu@outlook.com

2017 年 4 月 19 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)与(2)等价。

$$\begin{aligned} \min_{\mathbf{w}} \quad & -\mathbf{w}^{\mathrm{T}}\mathbf{S}_b\mathbf{w} \\ \text{s.t.} \quad & \mathbf{w}^{\mathrm{T}}\mathbf{S}_w\mathbf{w} = 1 \end{aligned} \tag{1}$$

$$\mathbf{S}_b\mathbf{w} = \lambda\mathbf{S}_w\mathbf{w} \tag{2}$$

Proof.

1.1 Identification of the Problem

(1) is the standard form of a optimization problem—minimize a objective function under constraints.

We use *KKT Multipliers Method* to find local minimums of a general non-linear optimization problem with equality and inequality constraints. In this problem, there are only **equality constraints**, therefore, we can just use *Lagrange Multipliers Method*.

Note Lagrange Multipliers Method requires **differentiability** of objective and constraint functions, which is obvious in this problem.

1.2 Write Lagrangian

$$\begin{aligned} L(\mathbf{w}, \lambda) =& f(\mathbf{w}) + \lambda g(\mathbf{w}) \\ =& -\mathbf{w}^{\mathrm{T}}\mathbf{S}_b\mathbf{w} + \lambda(\mathbf{w}^{\mathrm{T}}\mathbf{S}_w\mathbf{w} - 1) \end{aligned} \tag{3}$$

1.3 Conditions of Local Minimum

Sufficient and necessary conditions of local minimum contain the semi-definiteness of *Hessian matrices*, which is too complicated. Here we only consider **necessary** conditions, i.e. **first-order** conditions:

$$\begin{cases} \nabla_{\mathbf{w}} L(\mathbf{w}, \lambda) = 0 \\ \nabla_{\lambda} L(\mathbf{w}, \lambda) = 0 \end{cases} \tag{4}$$

Recall how to take derivative by a vector as a whole (Appendix A.3), we get:

$$\begin{cases} -2\mathbf{S}_b\mathbf{w} + 2\lambda\mathbf{S}_w\mathbf{w} = 0 \\ \mathbf{w}^{\mathrm{T}}\mathbf{S}_w\mathbf{w} - 1 = 0 \end{cases} \tag{5}$$

After simplification, we have:

$$\begin{cases} \mathbf{S}_b\mathbf{w} = \lambda\mathbf{S}_w\mathbf{w} \\ \mathbf{w}^{\mathrm{T}}\mathbf{S}_w\mathbf{w} = 1 \end{cases} \tag{6}$$

□

1.4 Discussion on Convexity of The Problem

Lagrange Multipliers Method can only find **local minimum**. For *convex optimization problems*, local minimums are also **global minimum**. Besides, we have good technique for convex optimization problems. So usually, after we write out the standard form of a optimization problem, we will consider its convexity.

1.4.1 Definition of Convex Optimization Problems

optimize (minimize) an objective function on convex sets.

1.4.2 Convexity of the Objective Function

$\mathbf{w}\mathbf{S}_b\mathbf{w}$ is a *quadratic form* of $\mathbf{w}$, so its convexity is equivalent to the positive semi-definiteness of matrix $\mathbf{S}_b$.

$\mathbf{S}_b$ is defined as $(\boldsymbol{\mu} - \boldsymbol{\mu}_0)(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^{\mathrm{T}}$, for convenience, let

$$\mathbf{a} = (a_1, a_2, \ldots, a_k)^{\mathrm{T}} = \boldsymbol{\mu} - \boldsymbol{\mu}_0 \tag{7}$$

then $\mathbf{S}_b = \mathbf{a}\mathbf{a}^{\mathrm{T}}$ is a matrix of rank 1, we can calculate its *eigen-polynomial*:

$$f(\lambda) = -(a_1^2 + a_2^2 + \cdots + a_k^2 - \lambda)\lambda^{(k-1)}. \tag{8}$$

So *eigenvalues* of the matrix are $\sum_{i=1}^{k} a_i^2$ and 0. Therefore the matrix is **positive semi-definiteness**. i.e. the objective function is convex.

1.4.3 Convexity of the domain set

$\mathbf{w}^{\mathrm{T}}\mathbf{S}_w\mathbf{w} = 1$ is also a quadratic form of $\mathbf{w}$, it defines a *hypersurface* in $\mathbb{R}^n$. We know a hypersurface is convex iff it is a hyperplane. Therefore the domain set is not convex.

In conclusion, the problem is not a convex optimization problem. □

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.

2.1 Problem (1)

$\forall j = 1, 2, 3, \ldots K$, we already know:

$$\ln \frac{p(y=j|\mathbf{x})}{p(y=K|\mathbf{x})} = \mathbf{w}_j^{\mathrm{T}}\mathbf{x} + b_j \tag{9}$$

take natural exponential on both sides:

$$p(y=j \mid \mathbf{x}) = \exp(\mathbf{w}_j^{\mathrm{T}}\mathbf{x} + b_j) * p(y=K \mid \mathbf{x}), \quad \forall j = 1, 2, 3, \ldots K-1 \tag{10}$$

Using normalization condition of probability

$$\sum_{k=1}^{K} p(y=k \mid \mathbf{x}) = 1, \quad \forall k = 1, 2, 3, \ldots K \tag{11}$$

we have:

$$\begin{aligned}
p(y=K \mid \mathbf{x}) &= \frac{1}{1+\sum_{k=1}^{K}\exp(\mathbf{w}_k^{\mathrm{T}}\mathbf{x} + b_k)} \\
p(y=j \mid \mathbf{x}) &= \frac{\exp(\mathbf{w}_j^{\mathrm{T}}\mathbf{x} + b_j)}{1+\sum_{k=1}^{K}\exp(\mathbf{w}_k^{\mathrm{T}}\mathbf{x} + b_k)},
\end{aligned} \tag{12}$$

$\forall j = 1, 2, 3, \ldots K-1$

Now we can calculate the *log-likelihood* function $\ell(\mathbf{w}, b)$:

$$\begin{aligned}\ell(\mathbf{w}, b) &= \sum_{i=1} m \ln p(y_i \mid \mathbf{x}_i;\ \mathbf{w}, b) \\ &= \sum_{i=1}^{m} \sum_{j=1}^{K} \mathbb{I}(y_i = j) * \ln p(y_i = j \mid \mathbf{x}_i;\ \mathbf{w}, b)\end{aligned} \tag{13}$$

From eqn(12) we know:

$$\begin{aligned}\ln p(y_i = K \mid \mathbf{x}) &= -\ln(1 + \sum_{k=1}^{K} \exp(\mathbf{w}_k^{\mathrm{T}} \mathbf{x} + b_k)) \\ \ln p(y_i = j \mid \mathbf{x}) &= (\mathbf{w}_j^{\mathrm{T}} \mathbf{x} + b_j) - \ln(1 + \sum_{k=1}^{K} \exp(\mathbf{w}_k^{\mathrm{T}} \mathbf{x} + b_k)) \\ \forall j = 1, 2, 3, \ldots K - 1\end{aligned} \tag{14}$$

Then insert (14) into (13):

$$\begin{aligned}\ell(\mathbf{w}, b) = &\sum_{i=1}^{m} \{ \sum_{j=1}^{K-1} \mathbb{I}(y_i = j) * [(\mathbf{w}_j^{\mathrm{T}} \mathbf{x}_i + b_j) - \ln(1 + \sum_{k=1}^{K} \exp(\mathbf{w}_k^{\mathrm{T}} \mathbf{x}_i + b_k))] \\ &- \mathbb{I}(y_i = K) \ln(1 + \sum_{k=1}^{K} \exp(\mathbf{w}_k^{\mathrm{T}} \mathbf{x}_i + b_k)) \} \\ = &\sum_{i=1}^{m} \{ \sum_{j=1}^{K-1} \mathbb{I}(y_i = j) * (\mathbf{w}_j^{\mathrm{T}} \mathbf{x}_i + b_j) - \sum_{j=1}^{K} \mathbb{I}(y_i = j) * \ln(1 + \sum_{k=1}^{K} \exp(\mathbf{w}_k^{\mathrm{T}} \mathbf{x}_i + b_k)) \} \\ = &\sum_{i=1}^{m} \{ \sum_{j=1}^{K-1} \mathbb{I}(y_i = j) * (\mathbf{w}_j^{\mathrm{T}} \mathbf{x}_i + b_j) - \ln(1 + \sum_{k=1}^{K} \exp(\mathbf{w}_k^{\mathrm{T}} \mathbf{x}_i + b_k)) \}\end{aligned}$$

2.2 Problem (2)

该“对数似然”为多元函数，其梯度为一向量。要求梯度，只需分别求出“对数似然”对每一个因变量的偏导数，即分别求 $\partial \ell(\mathbf{w}, b) / \partial \mathbf{w}_j$ 及 $\partial \ell(\mathbf{w}, b_j) / \partial b$.

这里为了简化推导，并充分利用向量化符号表示的优势，我们令 $\boldsymbol{\beta}_j^{\mathrm{T}} = (\mathbf{w}; b_j)$, $\mathbf{X}_i = (\mathbf{x}_i, 1)$, 则 $\ell(\mathbf{w}, b)$ 可写作:

$$\ell(\mathbf{w}, b) = \sum_{i=1}^{m} \{ \sum_{j=1}^{K-1} \mathbb{I}(y_i = j) * (\boldsymbol{\beta}_j^{\mathrm{T}} \mathbf{X}_i) - \ln(1 + \sum_{k=1}^{K} \exp(\boldsymbol{\beta}_k^{\mathrm{T}} \mathbf{X}_i)) \} \tag{15}$$

求 $\ell(\mathbf{w}, b)$ 对 $\boldsymbol{\beta}_j$ 的偏导，所得向量的各分量即为 $\ell(\mathbf{w}, b)$ 分别对 $\mathbf{w}_{j1}, \mathbf{w}_{j2}, \ldots$ 和 b_j 的偏导

数：

$$\begin{aligned}
\frac{\partial \ell(\mathbf{w}, b)}{\partial \boldsymbol{\beta}_j} &= \sum_{i=1}^{m}\{\frac{\partial}{\partial \boldsymbol{\beta}_j}\sum_{j=1}^{K-1}\mathbb{I}(y_i = j) * (\boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{X}_i) - \frac{\partial}{\partial \boldsymbol{\beta}_j}\ln(1 + \sum_{k=1}^{K}\exp(\boldsymbol{\beta}_k^{\mathrm{T}}\mathbf{X}_i))\} \\
&= \sum_{i=1}^{m}\{\mathbb{I}(y_i = j) * \mathbf{X}_i - \frac{\exp(\boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{X}_i) * \mathbf{X}_i}{1 + \sum_{k=1}^{K}\exp(\boldsymbol{\beta}_k^{\mathrm{T}}\mathbf{X}_i)}\} \\
&= \sum_{i=1}^{m}\mathbf{X}_i\{\mathbb{I}(y_i = j) - \frac{\exp(\boldsymbol{\beta}_j^{\mathrm{T}}\mathbf{X}_i)}{1 + \sum_{k=1}^{K}\exp(\boldsymbol{\beta}_k^{\mathrm{T}}\mathbf{X}_i)}\} \\
&= \sum_{i=1}^{m}\mathbf{X}_i\{\mathbb{I}(y_i = j) - p(y = j \mid \mathbf{x})\} \\
&\forall j = 1, 2, \ldots, K-1
\end{aligned} \tag{16}$$

综上可得对数似然的梯度为：

$$\nabla \ell(\mathbf{w}, b) = (\frac{\partial \ell(\mathbf{w}, b)}{\partial \boldsymbol{\beta}_1}, \frac{\partial \ell(\mathbf{w}, b)}{\partial \boldsymbol{\beta}_2}, \ldots, \frac{\partial \ell(\mathbf{w}, b)}{\partial \boldsymbol{\beta}_{K-1}}) \tag{17}$$

其中 $\partial \ell(\mathbf{w}, b)/\partial \boldsymbol{\beta}_j \quad (j = 1, 2, \ldots, K-1)$ 由式(16)确定。

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.

3.1 Problem (1)

代码见附件 main.py.

代码充分利用了 numpy 向量化的写法，效率较高；同时实现了使用 sklearn 中的 logistic 回归作为 benchmark 对比。

我实现的 logistic 回归中，数值解停止条件有二：

- 系数 $\hat{\beta}$ 改变比例足够小
- Hessian 矩阵的 condition number 足够大 (以至于近似奇异)

3.2 Problem (2)

遇到的主要问题是矩阵可逆性及数值不稳定问题，针对 exp 函数实现了 safe_exp 函数以保证不会 blow up，针对矩阵则使用 SVD 的方法来解决。

我认为编程作业理应有 3-8 小时的工作量，而且有一定难度，这是非常好的。但是希望能够提供足够的 reference，以给同学们方向性指引。

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：

$$\hat{\mathbf{w}}^*_{\mathbf{LS}}=\arg\min_{\mathbf{w}}\frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2, \tag{18}$$

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(18)引入正则化项, 通常形式如下：

$$\hat{\mathbf{w}}^*_{\mathbf{reg}}=\arg\min_{\mathbf{w}}\frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2+\lambda\Omega(\mathbf{w}), \tag{19}$$

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(18), 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(19)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(19)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,

$$\hat{\mathbf{w}}^*_{\ell_0}=\arg\min_{\mathbf{w}}\frac{1}{2}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|_2^2+\lambda\|\mathbf{w}\|_0, \tag{20}$$

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.

4.1 Problem (1)

Let $E_{\hat{\mathbf{w}}}$ be our objective function to be minimized:

$$E_{\hat{\mathbf{w}}}=(\mathbf{y}-\mathbf{X}\hat{\mathbf{w}})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\hat{\mathbf{w}}) \tag{21}$$

Take derivative $\partial E_{\hat{\mathbf{w}}}/\partial\hat{\mathbf{w}} = 0$ to find its minimum:

$$-2\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}) = 0 \tag{22}$$

Simplify and we have the *normal equation*:

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X})\hat{\mathbf{w}} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{23}$$

If $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is **invertible**, we can have a closed-form solution for $\hat{\mathbf{w}}$:

$$\hat{\mathbf{w}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{24}$$

The orthonormal property of $\mathbf{X}$ is **not needed** to derive this solution. It can just be used to further simplify the solution to:

$$\hat{\mathbf{w}} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{25}$$

For convenience, let $\hat{\mathbf{w}}^{\mathrm{LS}} = \mathbf{X}^{\mathrm{T}}\mathbf{y}$

4.2 Problem (2)

Let $E_{\hat{\mathbf{w}}}$ be our objective function to be minimized:

$$\begin{aligned} E_{\hat{\mathbf{w}}} =& \frac{1}{2}(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}) + \lambda\Omega(\hat{\mathbf{w}}) \\ =& \frac{1}{2}(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}) + \lambda\hat{\mathbf{w}}^{\mathrm{T}}\hat{\mathbf{w}} \end{aligned} \tag{26}$$

This is still differentiable, so we can take derivative $\partial E_{\hat{\mathbf{w}}}/\partial\hat{\mathbf{w}} = 0$ to find its minimum:

$$-\mathbf{X}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}) + 2\lambda\hat{\mathbf{w}} = 0 \tag{27}$$

Simplify and we have the *normal equation*:

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X} + 2\lambda\mathbf{I})\hat{\mathbf{w}} = \mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{28}$$

Even if $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is not invertible, since we add a constant, we can have a closed-form solution for $\hat{\mathbf{w}}$:

$$\hat{\mathbf{w}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X} + 2\lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \tag{29}$$

The orthonormal property of $\mathbf{X}$ is **not needed** to derive this solution. It can just be used to further simplify the solution to:

$$\begin{aligned} \hat{\mathbf{w}} =& (\mathbf{I} + 2\lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} \\ =& \frac{1}{1+2\lambda}\mathbf{X}^{\mathrm{T}}\mathbf{y} \\ =& \frac{1}{1+2\lambda}\hat{\mathbf{w}}^{\mathrm{LS}} \end{aligned} \tag{30}$$

Interpretation of this solution: Estimation of Ridge regression is a simple shrinkage of least square regression under orthonormal assumption.

4.3 Problem (3)

4.3.1 Characteristics of Problem (3) and How to Tackle It

We use **absolute loss** in LASSO regression, it cannot be expressed using vector $\hat{\mathbf{w}}$ and is **not directly differentiable**. Therefore, there is not closed-form solution for general LASSO regression problems.

However, with **orthonomal assumption**, equations of $\hat{\mathbf{w}}_i$ can be **decoupled** (variables can be solved separately). Then a closed-form solution can be derived.

4.3.2 Derivation of Closed-form Solution

Let $E_{\hat{\mathbf{w}}}$ be our objective function to be minimized:

$$\begin{aligned}
E_{\hat{\mathbf{w}}} =&\frac{1}{2}(\mathbf{y}-\mathbf{X}\hat{\mathbf{w}})^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\hat{\mathbf{w}})+\lambda\sum_i|\hat{\mathbf{w}}_i| \\
=&\frac{1}{2}(\mathbf{y}^{\mathrm{T}}\mathbf{y}-\mathbf{y}^{\mathrm{T}}\mathbf{X}\hat{\mathbf{w}}-\hat{\mathbf{w}}\mathbf{X}^{\mathrm{T}}\mathbf{y}+\hat{\mathbf{w}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\mathbf{w}})+\lambda\sum_i|\hat{\mathbf{w}}_i| \\
=&\frac{1}{2}(\mathbf{y}^{\mathrm{T}}\mathbf{y}-\mathbf{y}^{\mathrm{T}}\mathbf{X}\hat{\mathbf{w}}-\mathbf{y}^{\mathrm{T}}\mathbf{X}\hat{\mathbf{w}}+\hat{\mathbf{w}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\mathbf{w}})+\lambda\sum_i|\hat{\mathbf{w}}_i| \\
=&\frac{1}{2}(\mathbf{y}^{\mathrm{T}}\mathbf{y}-2\mathbf{y}^{\mathrm{T}}\mathbf{X}\hat{\mathbf{w}}+\hat{\mathbf{w}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\mathbf{w}})+\lambda\sum_i|\hat{\mathbf{w}}_i|
\end{aligned} \tag{31}$$

Since we only care about the value of $\hat{\mathbf{w}}$, we can discard constant term $\mathbf{y}^{\mathrm{T}}\mathbf{y}$. Let

$$\begin{aligned}
\tilde{E}_{\hat{\mathbf{w}}} =&E_{\hat{\mathbf{w}}}-\mathbf{y}^{\mathrm{T}}\mathbf{y} \\
=&-\mathbf{y}^{\mathrm{T}}\mathbf{X}\hat{\mathbf{w}}+\frac{1}{2}\hat{\mathbf{w}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\mathbf{w}}+\lambda\sum_i|\hat{\mathbf{w}}_i|
\end{aligned} \tag{32}$$

Using orthonormal assumption $\mathbf{X}^{\mathrm{T}}\mathbf{X}=\mathbf{I}$ in eqn(32), we have:

$$\begin{aligned}
\tilde{E}_{\hat{\mathbf{w}}} =&-\mathbf{y}^{\mathrm{T}}\mathbf{X}\hat{\mathbf{w}}+\frac{1}{2}\hat{\mathbf{w}}^{\mathrm{T}}\hat{\mathbf{w}}+\lambda\sum_i|\hat{\mathbf{w}}_i| \\
=&\sum_i[-(\mathbf{y}^{\mathrm{T}}\mathbf{X})_i\hat{\mathbf{w}}_i+\frac{1}{2}\hat{\mathbf{w}}_i^2+\lambda|\hat{\mathbf{w}}_i|] \\
=&\sum_i[-\hat{\mathbf{w}}_i^{\mathrm{LS}}\hat{\mathbf{w}}_i+\frac{1}{2}\hat{\mathbf{w}}_i^2+\lambda|\hat{\mathbf{w}}_i|]
\end{aligned} \tag{33}$$

We can see that after the orthonormal assumption is used, different $\hat{\mathbf{w}}_i$s are separated. We can minimize them independently. Define:

$$\begin{aligned}
\tilde{E}_{\hat{\mathbf{w}}i} =&-\hat{\mathbf{w}}_i^{\mathrm{LS}}\hat{\mathbf{w}}_i+\frac{1}{2}\hat{\mathbf{w}}_i^2+\lambda|\hat{\mathbf{w}}_i| \\
=&\begin{cases}\frac{1}{2}\hat{\mathbf{w}}_i[\hat{\mathbf{w}}_i-2(\hat{\mathbf{w}}_i^{\mathrm{LS}}-\lambda)], & \hat{\mathbf{w}}_i\geq 0 \\ \frac{1}{2}\hat{\mathbf{w}}_i[\hat{\mathbf{w}}_i-2(\hat{\mathbf{w}}_i^{\mathrm{LS}}+\lambda)], & \hat{\mathbf{w}}_i< 0\end{cases} \\
=&\frac{1}{2}\hat{\mathbf{w}}_i[\hat{\mathbf{w}}_i-2(\hat{\mathbf{w}}_i^{\mathrm{LS}}-\mathrm{sgn}(\hat{\mathbf{w}}_i)\lambda)]
\end{aligned} \tag{34}$$

Note that in eqn(34), $\tilde{E}_{\hat{\mathbf{w}}i}$ is the sum of two parabolas. Their axis of symmetry are:

$$\hat{\mathbf{w}}_i = \begin{cases} \hat{\mathbf{w}}_i^{\mathrm{LS}} - \lambda, & \hat{\mathbf{w}}_i \geq 0 \\ \hat{\mathbf{w}}_i^{\mathrm{LS}} + \lambda, & \hat{\mathbf{w}}_i < 0 \end{cases} \tag{35}$$

respectively. So when $\tilde{E}_{\hat{\mathbf{w}}i}$ reaches its minimum, we have:

$$\hat{\mathbf{w}}_i = \begin{cases} \hat{\mathbf{w}}_i = \max\{0, \hat{\mathbf{w}}_i^{\mathrm{LS}} - \lambda\} \\ \hat{\mathbf{w}}_i = \min\{0, \hat{\mathbf{w}}_i^{\mathrm{LS}} + \lambda\} \end{cases} \tag{36}$$

In conclusion, we have the closed-form solution of $\hat{\mathbf{w}}$ for L1 regularization under orthonormal assumption:

$$\hat{\mathbf{w}}_i = \begin{cases} \hat{\mathbf{w}}_i^{\mathrm{LS}} - \lambda, & \hat{\mathbf{w}}_i^{\mathrm{LS}} > \lambda \\ 0, & -\lambda \leq \hat{\mathbf{w}}_i^{\mathrm{LS}} \leq \lambda \\ \hat{\mathbf{w}}_i^{\mathrm{LS}} + \lambda, & \hat{\mathbf{w}}_i^{\mathrm{LS}} < -\lambda \end{cases} \tag{37}$$

or in a more compact form,

$$\begin{aligned} &\hat{\mathbf{w}}_i = \begin{cases} \mathrm{sgn}(\hat{\mathbf{w}}_i^{\mathrm{LS}})(|\hat{\mathbf{w}}_i^{\mathrm{LS}}| - \lambda), & |\hat{\mathbf{w}}_i^{\mathrm{LS}}| > \lambda \\ 0, & |\hat{\mathbf{w}}_i^{\mathrm{LS}}| \leq \lambda \end{cases} \\ &\forall i = 1, 2, 3, \ldots, k. \end{aligned} \tag{38}$$

4.4 Problem (4)

In this case, only the regularization term is changed, so we can still separate variables. Similarly, we have:

$$\tilde{E}_{\hat{\mathbf{w}}} = -\mathbf{y}^{\mathrm{T}}\mathbf{X}\hat{\mathbf{w}} + \frac{1}{2}\hat{\mathbf{w}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\mathbf{w}} + \lambda \sum_i \mathbb{I}(\hat{\mathbf{w}}_i \neq 0) \tag{39}$$

Using orthonormal assumption $\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{I}$ in eqn(32), we have:

$$\begin{aligned} \tilde{E}_{\hat{\mathbf{w}}} &= -\mathbf{y}^{\mathrm{T}}\mathbf{X}\hat{\mathbf{w}} + \frac{1}{2}\hat{\mathbf{w}}^{\mathrm{T}}\hat{\mathbf{w}} + \lambda \sum_i \mathbb{I}(\hat{\mathbf{w}}_i \neq 0) \\ &= \sum_i [-(\mathbf{y}^{\mathrm{T}}\mathbf{X})_i \hat{\mathbf{w}}_i + \frac{1}{2}\hat{\mathbf{w}}_i^2 + \lambda\mathbb{I}(\hat{\mathbf{w}}_i \neq 0)] \\ &= \sum_i [-\hat{\mathbf{w}}_i^{\mathrm{LS}}\hat{\mathbf{w}}_i + \frac{1}{2}\hat{\mathbf{w}}_i^2 + \lambda\mathbb{I}(\hat{\mathbf{w}}_i \neq 0)] \end{aligned} \tag{40}$$

Similarly we define:

$$\begin{aligned} \tilde{E}_{\hat{\mathbf{w}}i} &= -\hat{\mathbf{w}}_i^{\mathrm{LS}}\hat{\mathbf{w}}_i + \frac{1}{2}\hat{\mathbf{w}}_i^2 + \lambda\mathbb{I}(\hat{\mathbf{w}}_i \neq 0) \\ &= \begin{cases} -\hat{\mathbf{w}}_i^{\mathrm{LS}}\hat{\mathbf{w}}_i + \frac{1}{2}\hat{\mathbf{w}}_i^2 + \lambda, & \hat{\mathbf{w}}_i \neq 0 \\ 0, & \hat{\mathbf{w}}_i = 0 \end{cases} \\ &= \begin{cases} \frac{1}{2}\hat{\mathbf{w}}_i(\hat{\mathbf{w}}_i - 2\hat{\mathbf{w}}_i^{\mathrm{LS}}) + \lambda, & \hat{\mathbf{w}}_i \neq 0 \\ 0, & \hat{\mathbf{w}}_i = 0 \end{cases} \end{aligned} \tag{41}$$

Therefore

$$\arg\min_{\hat{\mathbf{w}}_i} \tilde{E}_{\hat{\mathbf{w}}i} = \min\left\{\arg\min_{\hat{\mathbf{w}}_i \neq 0} \frac{1}{2}\hat{\mathbf{w}}_i(\hat{\mathbf{w}}_i - 2\hat{\mathbf{w}}_i^{\mathrm{LS}}) + \lambda, \quad 0\right\} \tag{42}$$

We know that $\arg\min_{\hat{\mathbf{w}}_i \neq 0} \frac{1}{2}\hat{\mathbf{w}}_i(\hat{\mathbf{w}}_i - 2\hat{\mathbf{w}}_i^{\mathrm{LS}}) + \lambda = \lambda - \frac{1}{2}(\hat{\mathbf{w}}_i^{\mathrm{LS}})^2$, when $\hat{\mathbf{w}}_i = \hat{\mathbf{w}}_i^{\mathrm{LS}}$.

Therefore,

$$\arg\min_{\hat{\mathbf{w}}_i} \tilde{E}_{\hat{\mathbf{w}}i} = \begin{cases} \lambda - \frac{1}{2}(\hat{\mathbf{w}}_i^{\mathrm{LS}})^2, & |\hat{\mathbf{w}}_i^{\mathrm{LS}}| > \sqrt{2\lambda} \\ 0, & |\hat{\mathbf{w}}_i^{\mathrm{LS}}| \leq \sqrt{2\lambda} \end{cases} \tag{43}$$

and

$$\hat{\mathbf{w}}_i = \begin{cases} \hat{\mathbf{w}}_i^{\mathrm{LS}}, & |\hat{\mathbf{w}}_i^{\mathrm{LS}}| > \sqrt{2\lambda} \\ 0, & |\hat{\mathbf{w}}_i^{\mathrm{LS}}| \leq \sqrt{2\lambda} \end{cases} \tag{44}$$

4.4.1 去除列正交性质假设后, 为什么问题会变得非常困难

第三四小问在一般情况下都是没有闭式解的，这是因为误差表达式(39)关于不可导，而且各个 $\hat{\mathbf{w}}_i$ 之间是耦合在一起的，即

$$\frac{1}{2}\hat{\mathbf{w}}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}\hat{\mathbf{w}} \tag{45}$$

是一个关于 $\hat{\mathbf{w}}$ 的二次型，展开后存在各个 $\hat{\mathbf{w}}_i\hat{\mathbf{w}}_j$ 的交叉项，无法求解。

而在列正交性质前提下，二次型的矩阵 $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ 变为单位矩阵 $\mathbf{I}$，因而交叉项全部消失，进而可把误差 $\tilde{E}_{\hat{\mathbf{w}}}$ 写成每一个 $\hat{\mathbf{w}}_i$ 造成的误差的求和形式。进而可单独对每一个 $\hat{\mathbf{w}}_i$ 讨论、计算 (也就是第三、四题的推导过程)，这一过程在无正交假设时是无法进行的。