机器学习导论 习题四

221300079, 王俊童, 221300079@smail.nju.edu.cn

2024 年 5 月 21 日

作业提交注意事项

1. 作业所需的 LaTeX 及 Python 环境配置要求请参考: [Link];
2. 请在 LaTeX 模板中第一页填写个人的学号、姓名、邮箱;
3. 本次作业需提交的文件为:
   (a) 作答后的 LaTeX 代码 — `HW4.tex`;
   (b) 由 (a) 编译得到的 PDF 文件 — `HW4.pdf`;
   (c) 题目 2.(2) 的求解代码文件 — `svm_qp_dual.py`
   (d) 题目 3 的求解代码文件 — `svm_kernel_solution.py`

   其他文件 (如其他代码、图片等) 无需提交. 请将以上文件**打包为 `学号_姓名.zip`** (例如 `221300001_张三.zip`) 后提交;
3. 若多次提交作业, 则在命名 .zip 文件时加上版本号, 例如 `221300001_张三_v1.zip`" (批改时以版本号最高的文件为准);
4. 本次作业提交截止时间为 **5 月 28 日 23:59:59**. 未按照要求提交作业, 提交作业格式不正确, **作业命名不规范**, 将会被扣除部分作业分数; 除特殊原因 (如因病缓交, 需出示医院假条) 逾期未交作业, 本次作业记 0 分; **如发现抄袭, 抄袭和被抄袭双方成绩全部取消**;
5. 学习过程中, 允许参考 ChatGPT 等生成式语言模型的生成结果, 但必须在可信的信息源处核实信息的真实性; **不允许直接使用模型的生成结果作为作业的回答内容**, 否则将视为作业非本人完成并取消成绩;
6. 本次作业提交地址为 [Link], 请大家预留时间提前上交, 以防在临近截止日期时, 因网络等原因无法按时提交作业.

1 [35pts] Soft Margin

考虑软间隔 SVM 问题, 其原问题形式如下:

$$
\begin{aligned}
\min_{\boldsymbol{w},b,\xi_i} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i^p \\
\text{s.t.} \quad & y_i\left(\boldsymbol{w}^\top\boldsymbol{x}_i + b\right) \geq 1 - \xi_i \\
& \xi_i \geq 0, i \in [m].
\end{aligned} \tag{1.1}
$$

其中, 松弛变量 $\boldsymbol{\xi} = \{\xi_i\}_{i=1}^m, \xi_i > 0$ 表示样本 $\boldsymbol{x}_i$ 对应的间隔约束不满足的程度, 在优化问题中加入惩罚 $C\sum_{i=1}^m \xi_i^p, C > 0, p \geq 1$ 使得不满足约束的程度尽量小（$\xi_i \to 0$）. 课本式 (6.35) 即为 $p = 1$ 时对应的情况, 此时, 所有违反约束的样本都会受到相同比例的惩罚，而不考虑它们违反约束的程度. 这可能导致模型对较大偏差的样本不够敏感，不足以强调更严重的违规情况. 下面将考虑一些该问题的变式:

(1) **[2+7pts]** 我们首先考虑 $p = 2$ 的情况, 它对于违反约束程度较大的样本提供了更大的惩罚.

  (a) 如课本式 (6.34)-(6.35) 所述, $p = 1$ 的情况对应了 hinge 损失 $\ell_{hinge} : x \to \max(0, 1 - x)$. 请直接写出 $p = 2$ 的情况下对应的损失函数.

  (b) 请推导 $p = 2$ 情况下软间隔 SVM 的对偶问题.

(2) **[14pts]** $p = 1$ 的情况下, 相当于对向量 $\boldsymbol{\xi}$ 使用 L_1 范数惩罚: $\|\boldsymbol{\xi}\|_1 = \sum_i |\xi_i|$. 现在, 我们考虑使用 L_∞ 范数惩罚: $\|\boldsymbol{\xi}\|_\infty = \max_i \xi_i$, 这会使得模型着重控制最大的违背约束的程度，从而促使模型在最坏情况下的表现尽可能好. 请推导使用 L_∞ 范数惩罚的原问题和对偶问题.

(3) **[4+8pts]** 在(1.1)中, 正例和负例在目标函数中分类错误的“惩罚”是相同的. 然而在实际场景中, 很多时候正例和负例错分的“惩罚”代价是不同的 (参考教材 2.3.4 节). 比如考虑癌症诊断问题, 将一个确实患有癌症的人误分类为健康人, 以及将健康人误分类为患有癌症, 产生的错误影响以及代价不应该认为是等同的. 所以我们考虑对负例分类错误的样本施加 $k > 0$ 倍于正例中被分错的样本的“惩罚”.

  (a) 令(1.1)中 $p = 1$, 并令所有正例样本的集合为 D_+, 负例样本的集合为 D_-. 请给出相应的 SVM 优化问题.

  (b) 请给出相应的对偶问题.

Solution. 现在给出这个题的解答:

(1).p = 2

(a). 因为此时总的间隔是 4，我们可以得到:

$$
r = \frac{4}{\|\boldsymbol{w}\|_2^2}
$$

所以对于需要优化的函数:

$$
\min_{\boldsymbol{w},b,\xi_i} \quad \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i^2
$$

其中 ξ 是松弛变量。通过这个式子可以推导损失函数:

$$\ell_{hinge}: x \rightarrow \max(0, (1 - y_i(\boldsymbol{w}^\top x_i + b)))^2$$

(b). 下面给出对偶问题的具体推导，使用拉格朗日乘子法，首先可以得到原问题如下：

$$\begin{aligned}
\min_{\boldsymbol{w}, b, \xi_i} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i^2 \\
\text{s.t.} \quad & y_i\left(\boldsymbol{w}^\top \boldsymbol{x}_i + b\right) \geq 1 - \xi_i \\
& \xi_i \geq 0, i \in [m].
\end{aligned} \tag{1.2}$$

所以可以得到拉格朗日函数如下:

$$L(\boldsymbol{w}, b, \xi, \alpha, \beta) = \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^{m}\xi_i^2 - \sum_{i=1}^{m}\alpha_i(y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b) - 1 + \xi_i) - \sum_{i=1}^{m}\beta_i\xi_i$$

对 w,b,ξ 分别求偏导:

$$\frac{\partial L}{\partial \boldsymbol{w}} = w - \sum_{i=1}^{m}\alpha_i y_i x_i = 0$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{m}\alpha_i y_i = 0$$

$$\frac{\partial L}{\partial \xi} = 2C\xi_i - \mu_i - \alpha_i = 0$$

带入原式子化简可以得到对偶问题:

$$\begin{aligned}
\min_{\alpha} \quad & \sum_{i=1}^{m}\alpha_i - \sum_{i=1}^{m}\frac{(\alpha_i + \beta_i)^2}{4C} \\
\text{s.t.} \quad & \sum_{i=1}^{m}\alpha_i y_i = 0 \\
& 0 \leq \alpha_i \leq 2C, i \in [m].
\end{aligned} \tag{1.3}$$

(2) 若损失函数换成 L_∞ 范数作为惩罚，令 $||\xi||_\infty = max|\xi_i|$, 我们可以得到原问题形式:

$$\begin{aligned}
\min_{\boldsymbol{w}, b, \xi_i} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C||\xi||_\infty \\
\text{s.t.} \quad & y_i\left(\boldsymbol{w}^\top \boldsymbol{x}_i + b\right) \geq 1 - \xi_i \\
& \xi_i \geq 0, i \in [m].
\end{aligned} \tag{1.4}$$

这个原问题的等价问题等于:

$$\begin{aligned}
\min_{\boldsymbol{w}, b, \eta} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\eta \\
\text{s.t.} \quad & y_i\left(\boldsymbol{w}^\top \boldsymbol{x}_i + b\right) \geq 1 - \xi_i \\
& 0 \leq \xi_i \leq \eta, i \in [m].
\end{aligned} \tag{1.5}$$

所以可以得到这个函数的 lagrange 函数如下形式:

$$L(\boldsymbol{w}, b, \xi, \eta, \alpha, r, \beta)=\frac{1}{2}\|\boldsymbol{w}\|_2^2+C\eta-\sum_{i=1}^{m}\alpha_i(y_i(\boldsymbol{w}^\top\boldsymbol{x}_i+b)-1+\xi_i)-\sum_{i=1}^{m}r_i\xi_i-\sum_{i=1}^{m}\beta_i(\xi_i-\eta)$$

对 w,b,ξ,η 分别求偏导:

$$\frac{\partial L}{\partial \boldsymbol{w}}=w-\sum_{i=1}^{m}\alpha_i y_i x_i=0$$
$$\frac{\partial L}{\partial b}=\sum_{i=1}^{m}\alpha_i y_i=0$$
$$\frac{\partial L}{\partial \xi}=\beta_i-\alpha_i-r_i=0$$
$$\frac{\partial L}{\partial \eta}=C-\sum_{i=1}^{m}\beta_i=0$$

带入原问题可以得到这个原问题的对偶问题如下所示:

$$\begin{aligned}
\min_{\alpha} \quad & \sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^\top x_j \\
\text{s.t.} \quad & \sum_{i=1}^{m}\alpha_i y_i=0 \\
& \sum_{i=1}^{m}\alpha_i\le C \\
& 0\le\alpha_i, i\in[m].
\end{aligned} \tag{1.6}$$

(3) 对于这个问题，我们假设分类的正确样本和错误样本有数量如下：$D_+=$ n, $D_-=$ m. 同时我们引入不同的惩罚系数 C_+,C_-，且根据题意具有以下关系 $C_-=kC_+$.

(a). 所以根据上面所示，这个问题的原问题可以表达如下:

$$\begin{aligned}
\min_{\boldsymbol{w}, b, \xi_i^+, \xi_j^-} \quad & \frac{1}{2}\|\boldsymbol{w}\|_2^2+C_+\sum_{i}^{n}\xi_i^+ + C_-\sum_{j}^{m}\xi_i^- \\
\text{s.t.} \quad & y_i\left(\boldsymbol{w}^\top\boldsymbol{x}_i+b\right)\ge 1-\xi_i^+ \\
& y_j\left(\boldsymbol{w}^\top\boldsymbol{x}_j+b\right)\ge 1-\xi_j^- \\
& \xi_i^+\ge 0, i\in[n]. \\
& \xi_j^-\ge 0, j\in[m]. \\
& C_-=kC_+
\end{aligned} \tag{1.7}$$

(b). 所以可以根据原问题得到这个问题的 lagrange 函数如下所示:

$$L(\boldsymbol{w}, b, \xi_i^+, \xi_j^-, \alpha_1, \alpha_2, \beta_1, \beta_2)=\frac{1}{2}\|\boldsymbol{w}\|_2^2+C_+\sum_{i=1}^{n}\xi_i^+ + C_-\sum_{j=1}^{m}\xi_j^- - \sum_{i=1}^{n}\alpha_{1i}(y_i(\boldsymbol{w}^\top\boldsymbol{x}_i+b)-1+\xi_i^+)$$
$$-\sum_{j=1}^{m}\alpha_{2j}(y_j(\boldsymbol{w}^\top\boldsymbol{x}_j+b)-1+\xi_j^-)-\sum_{i=1}^{n}\beta_{1i}\xi_i^+-\sum_{j=1}^{m}\beta_{2j}\xi_j^-$$

对 w,b,ξ^+,ξ^- 分别求偏导:

$$\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{n} \alpha_{1i} y_i x_i - \sum_{j=1}^{m} \alpha_{2j} y_j x_j = 0$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n} \alpha_{1i} y_i + \sum_{j=1}^{m} \alpha_{2j} y_j = 0$$

$$\frac{\partial L}{\partial \xi^+} = C_+ - \alpha_{1i} - \beta_{1i} = 0$$

$$\frac{\partial L}{\partial \xi^-} = C_- - \alpha_{2j} - \beta_{2j} = 0$$

所以把这个偏导带入原问题可以得到如下的对偶问题:

$$\begin{aligned}
\min_{\alpha_{1i},\alpha_{2j}} \quad & \sum_{i=1}^{n} \alpha_{1i} + \sum_{j=1}^{m} \alpha_{2j} - \frac{1}{2}\sum_{i=1}^{n}\sum_{i=1}^{n} \alpha_i \alpha_j y_i y_j x_i^\top x_j - \frac{1}{2}\sum_{j=1}^{m}\sum_{j=1}^{m} \alpha_j \alpha_j y_j y_j x_j^\top x_j \\
& - \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^\top x_j \\
\text{s.t.} \quad & \sum_{i=1}^{n} \alpha_{1i} y_i + \sum_{j=1}^{m} \alpha_{2j} y_j = 0 \\
& 0 \le \alpha_{1i} \le C_+, i \in [n]. \\
& 0 \le \alpha_{2j} \le C_-, j \in [m]. \\
& C_- = kC_+
\end{aligned} \tag{1.8}$$

2 [20pts] Primal and Dual Problem

给定一个包含 m 个样本的数据集 $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$, 其中每个样本的特征维度为 d, 即 $\boldsymbol{x}_i \in \mathbb{R}^d$. 软间隔 SVM 的原问题和对偶问题可以表示为:

原问题:

$$
\begin{aligned}
\min_{\boldsymbol{w}, b, \xi} \quad & \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} + C\sum_{i=1}^n \xi_i \\
\text{s.t.} \quad & y_i\left(\boldsymbol{w}^\top \boldsymbol{x}_i + b\right) \geq 1 - \xi_i, \forall i \in [m] \\
& \xi_i \geq 0, \forall i \in [m]
\end{aligned}
$$

对偶问题:

$$
\begin{aligned}
\min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\boldsymbol{\alpha}^\top \boldsymbol{Q}\boldsymbol{\alpha} - \mathbf{1}_m^\top \boldsymbol{\alpha} \\
\text{s.t.} \quad & \boldsymbol{y}^\top \boldsymbol{\alpha} = 0 \\
& 0 \leq \alpha_i \leq C, \forall i \in [m]
\end{aligned}
$$

其中, 对于任意 $i, j \in [m]$ 有 $Q_{ij} \equiv y_i y_j \boldsymbol{x}_i^\top \boldsymbol{x}_j$.

上述的原问题和对偶问题都是都是二次规划 (Quadratic Programming) 问题, 都可以使用相关软件包求解. 本题目中我们将通过实践来学习凸优化软件包的使用, 并以软间隔 SVM 为例了解原问题、对偶问题在优化方面的特性.

(1) **[2pts]** 请直接写出原问题和对偶问题的参数量 (注意参数只包含分类器所保存的参数, 不包含中间变量).

(2) **[10pts]** 请参考 `lab2/svm_qp.py` 中对于原问题的求解代码, 编写对偶问题的求解代码 `lab2/svm_qp_dual.py`. (这里使用了 CVXPY 求解 QP 问题.) 请将代码提交至下方的解答处.

(3) **[8pts]** 设特征维度和样例数量的比值 $r = \frac{d}{m}$, 请绘制原问题和对偶问题的求解速度随着这个比值变化的曲线图. 并简述: 何时适合求解原问题, 何时适合求解对偶问题?

Solution. 此处用于写解答 (中英文均可)

(1) 原问题参数量为: $\boldsymbol{w}, b, \xi$, 对偶问题为:α

(2) 对偶问题的求解代码为:

```python
import cvxpy as cp
import numpy as np

def solve_dual(X, y, C):
    '''
    :参数 X: ndarray, 形状为(m, d), 样例矩阵
    :参数 y: ndarray, 形状为(m), 样例标签向量
    :参数 C: 标量, 含义与教材式(6.35)中C相同
    :返回: alpha, SVM的对偶变量
    '''
    m, d = X.shape
    y = y.reshape(-1, 1) * 1.0

    alpha = cp.Variable((m, 1),pos=True)
    Q = np.matmul(y * X, (y * X).T)
    prob = cp.Problem(cp.Minimize(0.5 * cp.quad_form(alpha, Q) - cp.sum(alpha))
                    ,[alpha <= C,
                      alpha >= 0,
```

```
                              cp.sum(cp.multiply(alpha,y)) == 0])
    prob.solve()
    return alpha.value
```

(3) 曲线图为:

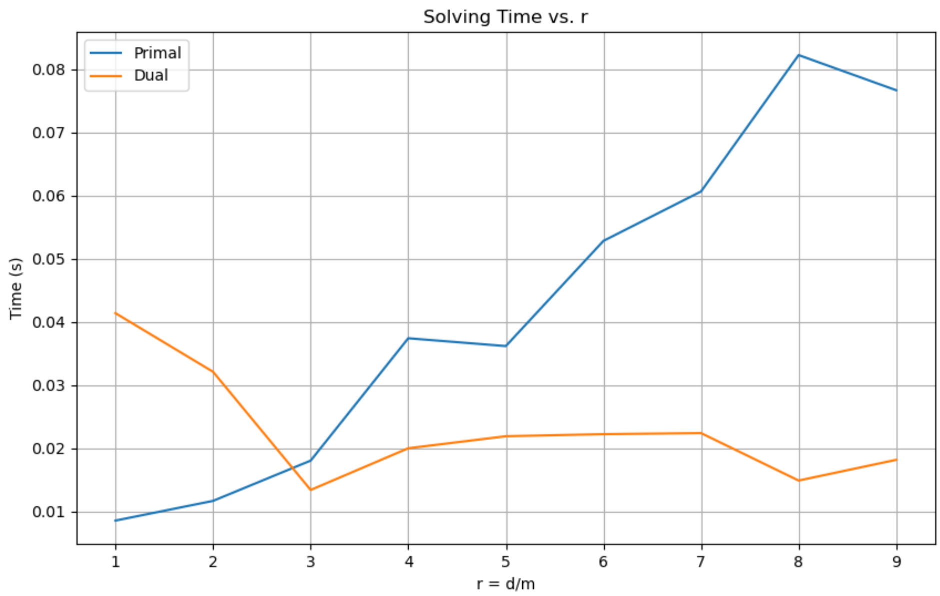


图 1: Primal-dual Problem solving time from 1 to 10

简述题:
这个 r 的比值我从 1 取值到了 10，可以看到，特征维度和样例数量的比值较低的时候，primal 问题明显消耗时间少一些，所以此时 r 小，更适合解决原问题。而当 r 的值上去了之后我们可以发现，primal 问题的解决时间明显慢下来了，所以此时适合解决对偶问题。

3 [15pts] Kernel Function in Practice

`lab3/svm_kernel.py` 中构造了异或 (XOR) 问题, 如图 2 所示. 该问题是线性不可分的.

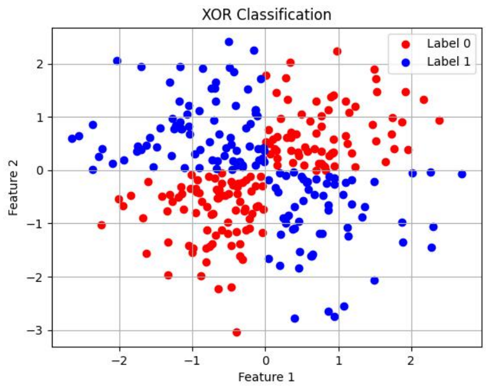


图 2: 异或 (XOR) 问题

本题中我们将通过实验了解核函数的选择对于 SVM 解决非线性问题的影响. 请使用 sklearn 包中的 SVM 分类器完成下述实验:

(1) **[6pts]** 请分别训练线性核 SVM 分类器和高斯核 (RBF 核) SVM 分类器, 并绘制出各自的决策边界.

(2) **[6pts]** sklearn 还允许自定义核函数, 参考 `lab3/svm_kernel_custom.py` 的用法, 编写核函数 $\kappa(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{1+\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2}$, 训练该核函数的 SVM 分类器, 并绘制出决策边界.

具体的实验要求可以参考 `lab3/svm_kernel.py` 的 `main` 部分. 请将 `lab3/svm_kernel_solution.py` 中的代码和三个核函数分别对应的决策边界图提交至下方的解答处.

最后, 请直接回答 **[3pts]**: 三个核函数, 各自能够解决异或 (XOR) 分类问题吗?

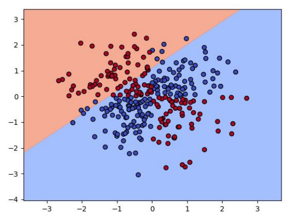
Solution. 此处用于写解答 (中英文均可)

求解代码为:

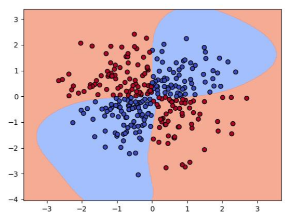
```python
from sklearn import svm
import numpy as np

def svm_kernel_linear(X, Y):
    '''
    :参数 X: ndarray, 形状(m, d), 样例矩阵
    :参数 Y: ndarray, 形状(m), 样例标签向量
    :返回: clf_linear, 训练好的分类器
    '''
    clf_linear = svm.SVC(kernel='linear', C=1.0)
    clf_linear.fit(X, Y)
    return clf_linear

def svm_kernel_rbf(X, Y):
    '''
    :参数 X: ndarray, 形状(m, d), 样例矩阵
    :参数 Y: ndarray, 形状(m), 样例标签向量
```

```
    :返回: clf_rbf, 训练好的分类器
    '''
    clf_rbf = svm.SVC(kernel='rbf', C=1.0)
    clf_rbf.fit(X, Y)
    return clf_rbf

def custom_kernel(X1, X2):
    '''
    :参数 X1: ndarray, 形状(m, d)
    :参数 X2: ndarray, 形状(n, d)
    :返回: 形状为(m, n)的Gram矩阵, 第(i,j)个元素为X1[i]和X2[j]之间的核函数值
    '''
    dist_squared = np.sum((X1[:, np.newaxis] - X2)**2, axis=2)
    K = 1 / (1 + dist_squared)
    return K

def svm_kernel_custom(X, Y):
    '''
    :参数 X: ndarray, 形状(m, d), 样例矩阵
    :参数 Y: ndarray, 形状(m), 样例标签向量
    :返回: clf_custom, 训练好的分类器
    '''
    clf_custom = svm.SVC(kernel=lambda X1, X2: custom_kernel(X1, X2), C=1.0)
    clf_custom.fit(X, Y)
    return clf_custom
```

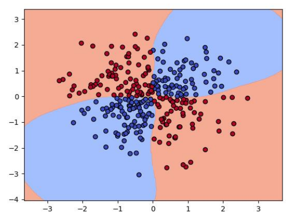
决策边界为:



(a) clf_linear



(b) clf_rbf



(c) clf_custom

能否解决异或 (XOR) 分类问题：明显第一个效果不好，后面两个可以解决 xor 问题。

4 [30pts] Maximum Likelihood Estimation

给定由 m 个样本组成的训练集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_m, y_m)\}$, 其中 $\boldsymbol{x}_i \in \mathbb{R}^d$ 是第 i 个示例, $y_i \in \mathbb{R}$ 是对应的实值标记. 令 $\boldsymbol{X} \in \mathbb{R}^{m\times d}$ 表示整个训练集中所有样本特征构成的矩阵, 并令 $\boldsymbol{y} \in \mathbb{R}^m$ 表示训练集中所有样本标记构成的向量. 线性回归的目标是寻找一个参数向量 $\boldsymbol{w} \in \mathbb{R}^d$, 使得在训练集上模型预测的结果和真实标记之间的差距最小. 对于一个样本 $\boldsymbol{x}$, 线性回归给出的预测为 $\hat{y} = \boldsymbol{w}^\top \boldsymbol{x}$,[1] 它与真实标记 y 之间的差距可以用平方损失 $(\hat{y} - y)^2$ 来描述. 因此, 在整个训练集上最小化损失函数的过程可以写作如下的优化问题:

$$\boldsymbol{w}^\star = \underset{\boldsymbol{w}}{\arg\min} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 \tag{4.1}$$

(1) **[8pts]** 考虑这样一种概率观点: 样本 $\boldsymbol{x}$ 的标记 y 是从一个高斯分布 $\mathcal{N}(\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2)$ 中采样得到的. 这个高斯分布的均值由样本特征 $\boldsymbol{x}$ 和模型参数 $\boldsymbol{w}$ 共同决定, 而方差是一个额外的参数 σ^2. 基于这种概率观点, 我们可以基于观测数据对高斯分布中的参数 $\boldsymbol{w}$ 做极大似然估计. 请证明: $\boldsymbol{w}$ 的极大似然估计结果 $\boldsymbol{w}_{\mathrm{MLE}}$ 与式 (4.1) 中的 $\boldsymbol{w}^\star$ 相等;

(2) **[9pts]** 极大似然估计容易过拟合, 一种常见的解决办法是采用最大后验估计: 沿着上一小问的思路, 现在我们希望在概率建模下对参数 $\boldsymbol{w}$ 做最大后验估计. 为此, 引入参数 $\boldsymbol{w}$ 上的先验 $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{0}, \lambda \boldsymbol{I})$. 其中, 均值 $\boldsymbol{0}$ 是 d 维的全 0 向量, $\boldsymbol{I}$ 是 d 维单位矩阵, $\lambda > 0$ 是一个控制方差的超参数. 现在, 请推导对 $\boldsymbol{w}$ 做最大后验估计的目标函数, 并讨论一下该结果与 "带有 L_2 范数正则项的线性回归" 之间的关系;

(3) **[9pts]** 沿着上一小问的思路, 我们尝试给参数 $\boldsymbol{w}$ 施加一个拉普拉斯先验. 简便起见, 我们假设参数 $\boldsymbol{w}$ 的 d 个维度之间是独立的, 且每一维都服从 0 均值的一元拉普拉斯分布, 即:

$$\begin{aligned} p(\boldsymbol{w}) &= \prod_{j=1}^{d} p(w_j)\ , \\ p(w_j) &= \mathrm{Lap}(w_j \mid 0, \lambda),\ j = 1, 2, \ldots, d\ . \end{aligned} \tag{4.2}$$

请推导对 $\boldsymbol{w}$ 做最大后验估计的目标函数, 并讨论一下该结果与 "带有 L_1 范数正则项的线性回归" 之间的关系;

Note: 由参数 μ, λ 确定的一元拉普拉斯分布的概率密度函数为:

$$\mathrm{Lap}(w \mid \mu, \lambda) = \frac{1}{2\lambda} \exp\left(-\frac{|w-\mu|}{\lambda}\right)\ . \tag{4.3}$$

(4) **[4pts]** 基于 (2) 和 (3) 的结果, 从概率角度讨论为什么 L_1 范数能使模型参数更稀疏.

Solution. (1) 我们对这个做一个 guass 分布:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

带入这个 guass 分布 $N(\boldsymbol{w}^\top \boldsymbol{x}, \sigma^2)$, 所以可以得到极大似然函数如下:

$$L(\boldsymbol{w}) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(y_i - \boldsymbol{w}^\top x_i)^2}{2\sigma^2})$$

[1] 本题不考虑偏移 b, 可参考教材第 3 章将偏移 b 吸收进 $\boldsymbol{w}$.

做对数极大似然:

$$\log L(\boldsymbol{w}) = -\frac{m}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i - \boldsymbol{w}^\top x_i)$$

由于第一个项是个无关项，然后我们可以得到这个式子的最小化的形式:

$$\begin{aligned}\boldsymbol{w}MLE &= \arg\min_{\boldsymbol{w}}\{\frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i - \boldsymbol{w}^\top x_i)\} \\ &= \arg\min_{\boldsymbol{w}}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2\end{aligned}$$

所以可以得到这个式子于式子 4.1 中的 $\boldsymbol{w}^\star$ 相等.

(2) 首先由于要做最大后验分布, 可以有贝叶斯公式得到:

$$p(\boldsymbol{w}|X,y) \propto p(y|X,\boldsymbol{w})p(\boldsymbol{w})$$

我们可以针对这个做最大后验，取一个负对数函数:

$$\arg\min_{\boldsymbol{w}}\log p(\boldsymbol{w}|X,y) \propto -\log p(y|X,\boldsymbol{w}) - \log p(\boldsymbol{w})$$

所以我们可以得到，由这个第一问知道，我们的 $p(y|X,w)$ 是很容易求出来的，这个等价于 $\frac{1}{2\sigma^2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2$, 所以我们只需要计算 $p(w)$ 即可, 由于这个东西符合多维 guass 分布，所以可以得到:

$$\begin{aligned}p(\boldsymbol{w}) &= \frac{1}{(2\pi\lambda)^{d/2}}exp(-\frac{1}{2\lambda}\|\boldsymbol{w}\|^2) \\ -\log p(\boldsymbol{w}) &= \frac{1}{2\lambda}\|\boldsymbol{w}\|^2 + \frac{d}{2}\log(2\pi\lambda) \propto \frac{1}{2\lambda}\|\boldsymbol{w}\|^2\end{aligned}$$

所以综上我们可以得到这个:

$$\arg\min_{\boldsymbol{w}}\log p(\boldsymbol{w}|X,y) \propto \frac{1}{2\sigma^2}\|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \frac{1}{2\lambda}\|\boldsymbol{w}\|^2$$

这个形式是等价于 L_2 正则化项的，由 L_2 正则化项的标准表达可以得到:

$$L_2 = L_{data} + \sigma\|\boldsymbol{w}\|_2^2$$

所以这个结果跟正则化这个是等价的，可以得到 $\sigma = \frac{1}{2\lambda}$. 说明在这种先验函数的前提下，这个式子符合一种 “带有 L_2 范数正则项的线性回归” 之间的关系。
(3) 对于这个题的 laplace 先验，我们只需要对 $p(\boldsymbol{w})$ 进行重新计算即可, 整个方法同上，我们仍然取负对数:

$$\arg\min_{\boldsymbol{w}}\log p(\boldsymbol{w}|X,y) \propto -\log p(y|X,\boldsymbol{w}) - \log p(\boldsymbol{w})$$

对于 $p(\boldsymbol{w})$

$$p(\boldsymbol{w}) = \prod_{j=1}^{d} \frac{1}{2\lambda} exp(-\frac{|w_j|}{\lambda})$$

$$p(\boldsymbol{w}) \propto \sum_{i=1}^{d} \frac{|w_j|}{\lambda}$$

所以我们可以得到这个结果:

$$\arg\min_{\boldsymbol{w}} \log p(\boldsymbol{w}|X, y) \propto \frac{1}{2\sigma^2} \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \sum_{i=1}^{d} \frac{|w_j|}{\lambda}$$

这个形式是等价于 L_1 正则化项的，由 L_1 正则化项的标准表达可以得到:

$$L_1 = L_{data} + \sigma \|\boldsymbol{w}\|_1$$

所以这里的 $\sigma = \frac{1}{\lambda}$

(4). 模型的参数的稀疏程度跟所选取的函数有密切关系, 拉普拉斯先验分布的特点是它在零点处有一个尖峰，并且随着远离零点，概率密度迅速下降。这意味着在贝叶斯推断中，模型更倾向于选择接近零的参数值，因为这些值在先验分布中具有更高的概率。当我们在后验分布中最大化时，这种先验的偏好会导致许多参数的估计值趋向于零，从而产生稀疏解

而且,L1 正则化在参数空间中引入了非平滑的约束，这导致目标函数在参数空间中形成“尖角”，即在某些方向上，即使是很小的移动也会导致正则化项的显著增加。这些“尖角”恰好对应于参数为零的点，因此在优化过程中，解往往会“卡”在这些尖角上，使得相应的参数为零。这就是为什么 L_1 范数能使模型参数更稀疏.

Acknowledgments

允许与其他同样未完成作业的同学讨论作业的内容, 但需在此注明并加以致谢; 如在作业过程中, 参考了互联网上的资料 (包括生成式模型的结果), 且对完成作业有帮助的, 亦需注明并致谢.