Homework 1

肖飞宇　2018210441

1 MATHEMATICS BASICS

1.1 Optimization

Use the Lagrange multiplier method to solve the following problem:

$$\begin{aligned} &\min_{x_1,x_2} \quad x_1^2+x_2^2-1 \\ &\text{s.t.} \quad x_1+x_2-1=0 \\ &\quad\quad x_1-2x_2\ge 0 \end{aligned} \tag{1.1}$$

Proof. 该问题的对偶问题为：

$$\begin{aligned} &\max \quad \theta(w,v) \\ &\text{s.t.} \quad w\ge 0 \end{aligned} \tag{1.2}$$

其中

$$\theta(w,v)=\inf\left\{x_1^2+x_2^2-1-w(x_1-2x_2)-v(x_1+x_2-1)\right\} \tag{1.3}$$

$$=\inf\left\{(x_1-\frac{2+v}{2})^2+(x_2+\frac{2w-v}{2})^2-\frac{(w+v)^2}{4}+(w-\frac{v}{2})^2+v-1\right\} \tag{1.4}$$

$$=-\frac{(w+v)^2}{4}+(w-\frac{v}{2})^2+v-1 \tag{1.5}$$

得到

$$\begin{aligned} &\max \quad -\tfrac{(w+v)^2}{4}+(w-\tfrac{v}{2})^2+v-1 \\ &\text{s.t.} \quad w\ge 0 \end{aligned} \tag{1.6}$$

并且有

$$x_1=\frac{2+v}{2} \tag{1.7}$$

$$x_2=-\frac{2w-v}{2} \tag{1.8}$$

计算对目标变量的偏导数

$$\frac{\partial\theta}{\partial w} = \frac{3}{2}w - \frac{3}{2} = 0 \tag{1.9}$$

$$\frac{\partial\theta}{\partial v} = -\frac{3}{2}v - 1 = 0 \tag{1.10}$$

解得

$$w = \frac{2}{3} \tag{1.11}$$

$$v = \frac{2}{3} \tag{1.12}$$

代入方程 1.7得到

$$x_1 = \frac{2}{3} \tag{1.13}$$

$$x_2 = \frac{1}{3} \tag{1.14}$$

最优值为 $-\frac{4}{9}$ □

1.2 Calculus

The gamma function is defined by (assuming x > 0)

$$\Gamma(x) = \int_0^{\infty} u^{x-1} e^{-u} du \tag{1.15}$$

(1)Prove that $\Gamma(x+1) = x\Gamma(x)$.

Proof.

$$\begin{aligned} x\Gamma(x) &= \int_0^{\infty} x u^{x-1} e^{-u} du && (1.16)\\ &= \int_0^{\infty} \frac{d}{du}(u^x) e^{-u} du && (1.17)\\ &= u^x e^{-x}|_0^{\infty} - \int_0^{\infty} u^x(-e^{-u}) du && (1.18)\\ &= \int_0^{\infty} u^x e^{-u} du && (1.19)\\ &= \Gamma(x+1) && (1.20) \end{aligned}$$

□

(2)Also show that

$$\int_0^1 \mu^{a-1}(1-\mu)^{b-1} d\mu = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{1.21}$$

Proof.

$$\begin{aligned} \Gamma(a)\Gamma(b) &= \int_0^{\infty} x^{a-1} e^{-x} dx \int_0^{\infty} y^{b-1} e^{-y} dy && (1.22)\\ &= \int_0^{\infty}\int_0^{\infty} x^{a-1} e^{-x} y^{b-1} e^{-y} dx dy && (1.23) \end{aligned}$$

取 $x+y=z$ 有

$$\Gamma(a)\Gamma(b)=\int_0^\infty x^{a-1}\left(\int_x^\infty (z-x)^{b-1}e^{-z}dz\right)dx \tag{1.24}$$

交换积分顺序有

$$\Gamma(a)\Gamma(b)=\int_0^\infty e^{-z}\left(\int_0^z x^{a-1}(z-x)^{b-1}dx\right)dz \tag{1.25}$$

再换元 $x=zu$有

$$\Gamma(a)\Gamma(b)=\int_0^\infty e^{-z}\left(\int_0^1 (zu)^{a-1}(z-zu)^{b-1}zdu\right)dz \tag{1.26}$$

$$=\int_0^\infty e^{-z}z^{a+b-1}dz\int_0^1 u^{a-1}(1-u)^{b-1}du \tag{1.27}$$

$$=\Gamma(a+b)\int_0^1 u^{a-1}(1-u)^{b-1}du \tag{1.28}$$

即有

$$\int_0^1 \mu^{a-1}(1-\mu)^{b-1}d\mu=\frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{1.29}$$

□

1.3 Probability

Suppose $\lambda\sim\Gamma(p|\alpha,\beta)$ and $x|\lambda\sim\text{Poisson}(x|\lambda)$. Show that $\lambda|x\sim\Gamma(\lambda|\alpha+x,\beta+1)$,which implies that the Gamma distribution can serve as a conjugate prior to the Poisson distribution.

Proof. 首先有

$$f(\lambda)=\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda} \tag{1.30}$$

$$f(x|\lambda)=\frac{\lambda^x}{x!}\cdot e^{-\lambda} \tag{1.31}$$

则其联合分布密度函数为

$$f(x,\lambda)=f(x|\lambda)f(\lambda)=\frac{\lambda^x}{x!}e^{-\lambda}\cdot\frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\beta-1}e^{-\beta\mu}=\frac{\lambda^{x+\alpha-1}\beta^\alpha}{x!\Gamma(\alpha)}e^{-\lambda(\beta+1)} \tag{1.32}$$

则可以求得 x 的边缘分布密度函数为

$$f(x)=\int_0^\infty f(x,\lambda)d\lambda=\frac{\Gamma(x+\alpha)\beta^\alpha}{x!(\beta+1)^{x+\alpha}\Gamma(\alpha)} \tag{1.33}$$

根据贝叶斯公式有

$$f(\lambda|x)=\frac{f(x,\lambda)f(x)}{=}\frac{\frac{\lambda^{x+\alpha-1}\beta^\alpha}{x!\Gamma(\alpha)}e^{-\lambda(\beta+1)}}{\frac{\Gamma(x+\alpha)\beta^\alpha}{x!(\beta+1)^{x+\alpha}\Gamma(\alpha)}}=\frac{(\beta+1)^{\alpha+x}}{\Gamma(\alpha+x)}\lambda^{x+a-1}e^{-\lambda(\beta+1)} \tag{1.34}$$

可以看出 λ 的后验分布满足 $\lambda|x\sim Ga(\alpha+x,\beta+1)$
即泊松分布式伽马分布的共轭先验分布

□

1.4 Stochastic Process

Proof. 设定两个事件，即 A_k代表出现序列 $H,\underbrace{T,T,\cdots,T}_{k}$的平均投掷次数,而 B_k 代表在出现H 后出现序列$H,\underbrace{T,T,\cdots,T}_{k}$的平均投掷次数

显然，考虑第一次投掷有两种情况：

- 结果为H，那么接下来即为B_n
- 结果为T，那么接下来即为A_n

有

$$A_n = 1 + \frac{1}{2}A_n + \frac{1}{2}B_n \tag{1.35}$$

推得

$$A_n = B_n + 2 \tag{1.36}$$

在序列 B_{n-1}后再次进行一次投掷，有两种结果：

- 如果是T，则结束为 $B_{n-1}+1$
- 如果是H，则重新开始为 $B_{n-1}+1+B_n$

接下来得到

$$B_n = \frac{1}{2}(B_{n-1}+1) + \frac{1}{2}(B_{n-1}+B_n+1) \tag{1.37}$$

得到递推公式

$$B_n = 2B_{n-1} + 2 \tag{1.38}$$

设 $f(n) = B_n + 2$则由$f(0) = B_0 + 2 = 2$有

$$f(n) = 2^{n+1} \tag{1.39}$$

从而

$$B_n = 2^{n+1} - 2 \tag{1.40}$$

从而得到

$$A_k = B_k + 2 = 2^{k+1} \tag{1.41}$$

其观察到这样的模式的所需要的实验的平均次数为 2^{k+1} □

2 SVM

2.1 From Primal to Dual

Consider the regression problem with training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$ $(\boldsymbol{x}_i \in \mathbb{R}^d, y_i \in \mathbb{R})$. $\epsilon > 0$ debotes a fixed small value. Derive the dual problem of the following primal problem of linear SVM (Please use $\alpha_i, \hat{\alpha}_i, \beta_i$, and $\hat{\beta}_i$ as the Lagrange multipliers):

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi},\hat{\xi}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^N \left(\xi_i + \hat{\xi}_i\right) \tag{2.1}$$

s.t.

$$y_i \leq \boldsymbol{w}^\top \boldsymbol{x}_i + b + \epsilon + \xi_i, i = 1, \ldots, N \tag{2.2}$$

$$y_i \geq \boldsymbol{w}^\top \boldsymbol{x}_i + b - \epsilon - \hat{\xi}_i, i = 1, \ldots, N \tag{2.3}$$

$$\xi_i \geq 0 \quad \forall i = 1, \ldots, N \tag{2.4}$$

$$\hat{\xi}_i \geq 0 \quad \forall i = 1, \ldots, N \tag{2.5}$$

Proof. 首先将问题变为标准形式：

$$\min \quad {}_{\boldsymbol{w}, b, \boldsymbol{\xi}, \hat{\xi}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\left(\xi_i + \hat{\xi}_i\right) \tag{2.6}$$

s.t.

$$\boldsymbol{w}^\top \boldsymbol{x}_i + b + \epsilon + \xi_i - y_i \geq 0, \quad i = 1, \ldots, N \tag{2.7}$$

$$y_i - \boldsymbol{w}^\top \boldsymbol{x}_i - b + \epsilon + \hat{\xi}_i \geq 0, \quad i = 1, \ldots, N \tag{2.8}$$

$$\xi_i \geq 0 \quad \forall i = 1, \ldots, N \tag{2.9}$$

$$\hat{\xi}_i \geq 0 \quad \forall i = 1, \ldots, N \tag{2.10}$$

问题的对偶问题为

$$\max \quad \theta(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}, \boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) \tag{2.11}$$

s.t.

$$\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}, \boldsymbol{\beta}, \hat{\boldsymbol{\beta}} \geq \boldsymbol{0} \tag{2.12}$$

其中

$$\theta = \inf\{\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\left(\xi_i + \hat{\xi}_i\right) - \sum_{i}^{N}\alpha_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b + \epsilon + \xi_i - y_i) - \tag{2.13}$$

$$\sum_{i}^{N}\hat{\alpha}_i(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i - b + \epsilon + \hat{\xi}_i) - \beta_i\xi_i - \hat{\beta}_i\hat{\xi}_i\} = \inf\{F\} \tag{2.14}$$

上式中F应该对$\boldsymbol{w}, b, \xi, \hat{\xi}$求最小值，即偏导数分别为0：

$$\frac{\partial F}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i}^{N}\alpha_i x_i + \sum_{i}^{N}\hat{\alpha}_i x_i = 0 \tag{2.15}$$

$$\frac{\partial F}{\partial b} = -\sum_{i}^{N}\alpha_i + \sum_{i}^{N}\hat{\alpha}_i = 0 \tag{2.16}$$

$$\frac{\partial F}{\partial \xi_i} = C - \alpha_i - \beta_i = 0 \tag{2.17}$$

$$\frac{\partial F}{\partial \hat{\xi}_i} = C - \hat{\alpha}_i - \hat{\beta}_i = 0 \tag{2.18}$$

则有

$$\boldsymbol{w} = \sum_i^N \alpha_i x_i - \sum_i^N \hat{\alpha}_i x_i \tag{2.19}$$

$$\sum_i^N \alpha_i = \sum_i^N \hat{\alpha}_i \tag{2.20}$$

$$C = \alpha_i + \beta_i \tag{2.21}$$

$$C = \hat{\alpha}_i + \hat{\beta}_i \tag{2.22}$$

代入F的表达式得到：

$$\theta = \frac{1}{2}\boldsymbol{x}^T(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}})(\boldsymbol{\alpha} - \hat{\boldsymbol{\alpha}})^T\boldsymbol{x} - 2\epsilon\boldsymbol{\alpha_i}^T\mathbf{1} \tag{2.23}$$

即为所求的对偶问题 □

2.2 Finding Support Vectors (Optional)

State the corresponding set of KKT conditions. Now please argue from the KKT conditions why the following hold:

$$\alpha_i = 0 \Rightarrow \boldsymbol{w}^\top \boldsymbol{x}_i + b + \epsilon - y_i \geq 0 \tag{2.24}$$

$$0 < \alpha_i < C \Rightarrow \boldsymbol{w}^\top \boldsymbol{x}_i + b + \epsilon - y_i = 0 \tag{2.25}$$

$$\alpha_i = C \Rightarrow \boldsymbol{w}^\top \boldsymbol{x}_i + b + \epsilon - y_i \leq 0 \tag{2.26}$$

$$\hat{\alpha}_i = 0 \Rightarrow \boldsymbol{w}^\top \boldsymbol{x}_i + b - \epsilon - y_i \leq 0 \tag{2.27}$$

$$0 < \hat{\alpha}_i < C \Rightarrow \boldsymbol{w}^\top \boldsymbol{x}_i + b - \epsilon - y_i = 0 \tag{2.28}$$

$$\hat{\alpha}_i = C \Rightarrow \boldsymbol{w}^\top \boldsymbol{x}_i + b - \epsilon - y_i \geq 0 \tag{2.29}$$

Please show that either α_i or $\hat{\alpha}_i$(or both) must be zero. What are the support vectors?

Proof. 对于 KKT 条件中的导数条件和广义Lagrange函数的导数条件相同即：

$$\boldsymbol{w} - \sum_i^N \alpha_i x_i + \sum_i^N \hat{\alpha}_i x_i = 0 \tag{2.30}$$

$$-\sum_i^N \alpha_i + \sum_i^N \hat{\alpha}_i = 0 \tag{2.31}$$

$$C - \alpha_i - \beta_i = 0 \tag{2.32}$$

$$C - \hat{\alpha}_i - \hat{\beta}_i = 0 \tag{2.33}$$

下面给出 KKT 条件中的互补松弛条件：

$$\alpha_i\left(\boldsymbol{w}^\top \boldsymbol{x}_i + b + \epsilon + \xi_i - y_i\right) = 0, \quad \forall i = 1, \ldots, N \tag{2.34}$$

$$\hat{\alpha}_i\left(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i - b + \epsilon + \hat{\xi}_i\right) = 0, \quad \forall i = 1, \ldots, N \tag{2.35}$$

$$\beta_i \xi_i = 0 \quad \forall i = 1, \ldots, N \tag{2.36}$$

$$\hat{\beta}_i \hat{\xi}_i = 0 \quad \forall i = 1, \ldots, N \tag{2.37}$$

$$\alpha_i \geq 0 \quad \forall i = 1, \ldots, N \tag{2.38}$$

$$\hat{\alpha}_i \geq 0 \quad \forall i = 1, \ldots, N \tag{2.39}$$

$$\beta_i \geq 0 \quad \forall i = 1, \ldots, N \tag{2.40}$$

$$\hat{\beta}_i \geq 0 \quad \forall i = 1, \ldots, N \tag{2.41}$$

如果 $\alpha_i = 0$，由互补松弛条件有

$$\boldsymbol{w}^\top \boldsymbol{x}_i + b + \epsilon + \xi_i - y_i \geq 0 \tag{2.42}$$

而上面的 Lagrange 函数的导数为零可以得到

$$C = \alpha_i + \beta_i \tag{2.43}$$

则有

$$\beta_i = C > 0 \tag{2.44}$$

由于互补松弛条件有

$$\xi_i = 0 \tag{2.45}$$

推得

$$\boldsymbol{w}^\top \boldsymbol{x}_i + b + \epsilon - y_i \geq 0 \tag{2.46}$$

如果 $0 < \alpha_i < C$，由互补松弛条件有

$$\boldsymbol{w}^\top \boldsymbol{x}_i + b + \epsilon + \xi_i - y_i = 0 \tag{2.47}$$

另外同样由 $C = \alpha_i + \beta_i$可以得到

$$\beta_i > 0 \tag{2.48}$$

则

$$\xi_i = 0 \tag{2.49}$$

推得

$$\boldsymbol{w}^\top \boldsymbol{x}_i + b + \epsilon - y_i = 0 \tag{2.50}$$

如果 $\alpha_i = C$，由互补松弛条件有

$$\boldsymbol{w}^\top \boldsymbol{x}_i + b + \epsilon + \xi_i - y_i = 0 \tag{2.51}$$

即

$$\boldsymbol{v}^\top \boldsymbol{x}_i + b + \epsilon - y_i = -\xi_i \tag{2.52}$$

再由 $\alpha_i+\beta_i=C$ 可以得到

$$\beta_i=0 \tag{2.53}$$

有

$$\xi_i \geq 0 \tag{2.54}$$

推得

$$\boldsymbol{v}^\top \boldsymbol{x}_i+b+\epsilon-y_i=-\xi_i \leq 0 \tag{2.55}$$

对于 $\hat{\alpha}$ 的情况可以完全类似的讨论

如果 α_i 和 $\hat{\alpha}_i$中没有一个等于0，分下面四种情况讨论

- $0<\alpha_i<C$ 和 $0<\hat{\alpha}_i<C$
  此时有

$$\boldsymbol{w}^\top \boldsymbol{x}_i+b+\epsilon-y_i=0 \tag{2.56}$$

$$\boldsymbol{w}^\top \boldsymbol{x}_i+b-\epsilon-y_i=0 \tag{2.57}$$

  两式相减得到

$$\epsilon=0 \tag{2.58}$$

  和

$$\epsilon>0 \tag{2.59}$$

  矛盾

- $\alpha_i=C$ 和 $\hat{\alpha}_i=C$
  有

$$\boldsymbol{w}^\top \boldsymbol{x}_i+b+\epsilon-y_i \leq 0 \tag{2.60}$$

$$\boldsymbol{w}^\top \boldsymbol{x}_i+b-\epsilon-y_i \geq 0 \tag{2.61}$$

  两个不等式相减得到

$$\epsilon \leq 0 \tag{2.62}$$

  和

$$\epsilon>0 \tag{2.63}$$

  矛盾

- $\alpha_i=C$ 和 $0<\hat{\alpha}_i<C$
  有

$$\boldsymbol{w}^\top \boldsymbol{x}_i+b+\epsilon-y_i \leq 0 \tag{2.64}$$

$$\boldsymbol{w}^\top \boldsymbol{x}_i+b-\epsilon-y_i=0 \tag{2.65}$$

推得

$$\epsilon \leq 0 \tag{2.66}$$

和

$$\epsilon > 0 \tag{2.67}$$

矛盾

- $0 < \alpha_i < C$ 和 $\hat{\alpha}_i = C$
  有

$$\boldsymbol{w}^\top \boldsymbol{x}_i + b + \epsilon - y_i = 0 \tag{2.68}$$

$$\boldsymbol{w}^\top \boldsymbol{x}_i + b - \epsilon - y_i \geq 0 \tag{2.69}$$

有

$$\epsilon \leq 0 \tag{2.70}$$

和

$$\epsilon > 0 \tag{2.71}$$

矛盾

综上， α_i 和 $\hat{\alpha}_i$ 中至少有一个为0
满足方程

$$\boldsymbol{w}^\top \boldsymbol{x}_i + b + \epsilon + \xi_i - y_i = 0 \tag{2.72}$$

$$y_i - \boldsymbol{w}^\top \boldsymbol{x}_i - b + \epsilon + \hat{\xi}_i = 0 \tag{2.73}$$

的点称为支持向量。 □

3 IRLS for Logistic Regression

3.1 推导

Proof. 该迭代方法基于 Newton-Raphson迭代，其对权重的更新形式为：

$$\mathbf{w}^{(\text{new})} = \mathbf{w}^{(\text{old})} - \mathbf{H}^{-1} \nabla E(\mathbf{w}) \tag{3.1}$$

对于对数似然函数：

$$\mathscr{L}(\boldsymbol{w}) = \sum_{i=1}^{N} \left(y_i \boldsymbol{w}^\top \boldsymbol{x}_i - \log\left(1 + \exp\left(\boldsymbol{w}^\top \boldsymbol{x}_i\right)\right)\right) \tag{3.2}$$

对于 logictic 函数的导数：

$$\frac{\partial sig(\Phi_n)}{\partial w} = sig(\Phi_n)(1 - \Phi_n)\Phi_n \tag{3.3}$$

梯度为

$$\nabla E(\mathrm{w}) = \sum_{n=1}^{N} \left(\mathrm{y}_n - t_n\right) \phi_n = \phi^T (\mathrm{Y} - \mathrm{T}) \tag{3.4}$$

其中

$$Y = sig(\Phi W) \tag{3.5}$$

从而

$$H = \nabla\nabla E(\mathrm{w}) = \nabla \phi^T (Y - T) = \phi^T \nabla Y = \phi^T \sum_{n=1}^{N} y_n \left(1 - y_n\right) \phi_n = \phi^T \sum_{n=1}^{N} y_n \left(1 - y_n\right) \phi \tag{3.6}$$

其中 Φ 是 $M \times N$ 维的数据矩阵，也就是样本的个数为 M，维度为 N， T 则是$M \times 1$的特征的标签矩阵。

用 $R_{nn} = y_n(1 - y_n)$ 表示 $\sum_{n=1}^{N} \mathrm{y}_n \left(1 - \mathrm{y}_n\right)$，有

$$\begin{aligned} w^{new} &= w^{old} - H^{-1} \nabla E(w) && (3.7) \\ &= w^{old} - \left(\phi^T R_{nn} \phi\right)^{-1} \phi^T (\mathrm{Y} - \mathrm{T}) && (3.8) \\ &= \left(\phi^T R_{nn} \phi\right)^{-1} \phi^T R_{nn} \left[\phi w^{old} - R_{nn}^{-1} (\mathrm{Y} - \mathrm{T})\right] && (3.9) \end{aligned}$$

如果加上正则项：

$$L_2 = -\frac{\lambda}{2} \| w \|_2^2 \tag{3.10}$$

由

$$\nabla L_2 = -\lambda w \tag{3.11}$$

$$H_{L_2} = \nabla\nabla L_2 = -\lambda \tag{3.12}$$

得到其更新公式为

$$\begin{aligned} w^{new} &= w^{old} - H^{-1} \nabla E(w) && (3.13) \\ &= w^{old} - \left(\phi^T R_{nn} \phi - \lambda I_n\right)^{-1} \left[\phi^T (\mathrm{Y} - \mathrm{T}) - \lambda w_{old} I\right] && (3.14) \\ &= \left(\phi^T R_{nn} \phi - \lambda I_n\right)^{-1} \left[\left(\phi^T R_{nn} \phi - \lambda I\right) w_{old} - \phi^T (\mathrm{Y} - \mathrm{T}) + \lambda w_{old} I\right] && (3.15) \\ &= \left(\phi^T R_{mn} \phi + \lambda I_n\right)^{-1} \phi^T R_{mn} \left[\phi w^{old} - R_{nn}^{-1} (\mathrm{Y} - \mathrm{T})\right] && (3.16) \end{aligned}$$

□

3.2 程序编写和结果

(1)

该问题的程序见 **LR-IRLS.py**. 其结果为在训练集上的精度为0.84816,在测试集上的预测精度为0.84817，其迭代一次即收敛。

其

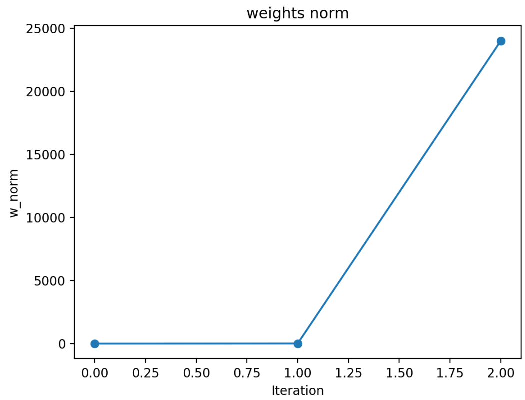

Figure 3.1: Weights norm in iteration

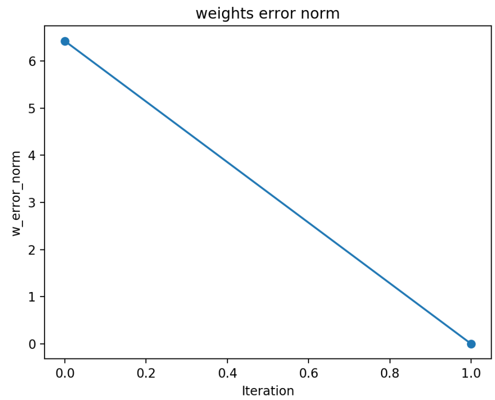

Figure 3.2: Weights error norm in iteration

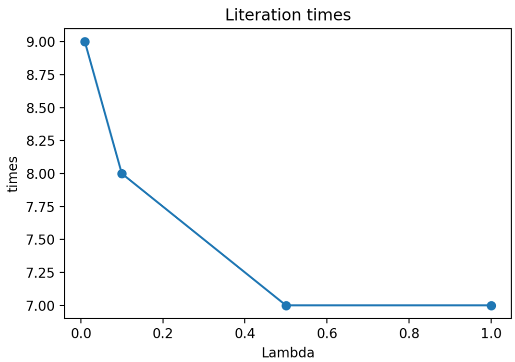

Figure 3.3: Iteration times

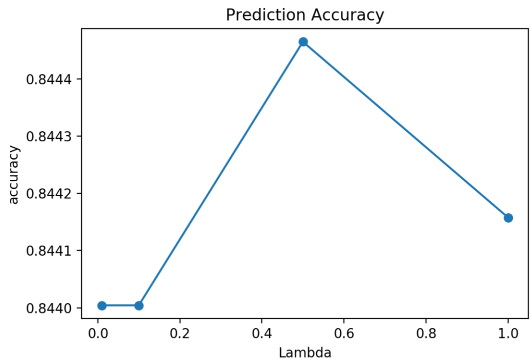

Figure 3.4: Prediction error in different λ

(2)

该问题的程序见 **LR-IRLS-L2.py**. 通过比较得出的 λ 的较优值为 0.5,各个不同的 λ 的取值分别为[0.01,0.1,0.5,1]，其在不同参数下的迭代次数和预测精度见图3.2 可以看出较好的 λ 的取值为 0.5,其在训练集上的精度为0.8445，在测试集上的精度为0.8501
对于 $\lambda = 0.5$的情况，其迭代过程中参数 w 的范数和更新的 w 和之前的 w 的差的范数见图3.2

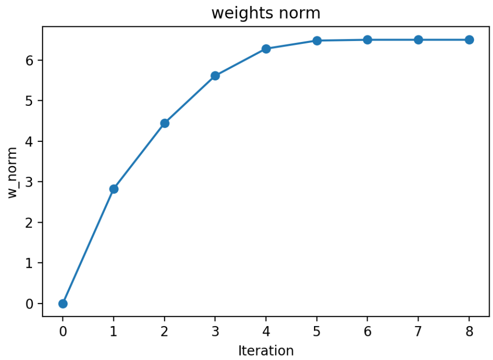

Figure 3.5: weights norm

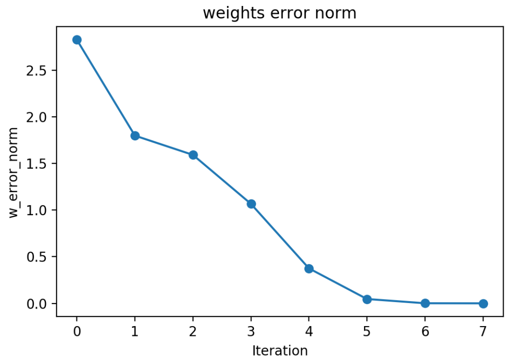

Figure 3.6: weights error norm

3.3 (3)

最后利用 **SKlearn** 的结果和前面的精度进行一个参照，源代码见**LR-SKlearn.py**. 在训练集上的精度为 0.8492，在测试集上的精度为 0.8498.

3.4 代码和文档

代码和文档见 IRLS-Logistic-Regression