Homework 1 of Statistical Machine Learning

Wang Yikai, 2017310740

March 30, 2018

1 Mathematics Basics

1.1 Optimization

Use the Lagrange multiplier method to solve the following problem:

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

The Lagrange function is:

$$L(x_1,x_2,\alpha,\beta)=x_1^2+x_2^2+\alpha(2x_2-x_1)+\beta(x_1+x_2-1)$$

where α and β are Lagrange multipliers.

The dual problem is:

$$\begin{aligned}&\max_{\alpha,\beta}\min_{x_1,x_2}\ L(x_1,x_2,\alpha,\beta)\\&s.t.\ \alpha\geq 0\end{aligned}$$

First find the derivatives of $L(x_1,x_2,\alpha,\beta)$ w.r.t. x_1,x_2, and let the derivatives equal to zero:

$$\begin{aligned}\frac{\partial L(x_1,x_2,\alpha,\beta)}{\partial x_1}&=2x_1-\alpha+\beta=0\Rightarrow x_1=\frac{\alpha-\beta}{2}\\\frac{\partial L(x_1,x_2,\alpha,\beta)}{\partial x_2}&=2x_2+2\alpha+\beta=0\Rightarrow x_2=-\alpha-\frac{\beta}{2}\end{aligned}$$

Therefore the dual problem turns to:

$$\begin{aligned}&\max_{\alpha,\beta}\ g(\alpha,\beta)=-\frac{5}{4}\alpha^2-\frac{1}{2}\beta^2-\frac{1}{2}\alpha\beta-\beta\\&s.t.\ \alpha\geq 0\end{aligned}$$

Compute the derivatives of $g(\alpha, \beta)$ w.r.t. α, β, there are:

$$\begin{aligned}
\frac{\partial g(\alpha,\beta)}{\partial \alpha} &= -\frac{5}{2}\alpha - \frac{1}{2}\beta = 0 \\
\frac{\partial g(\alpha,\beta)}{\partial \beta} &= -\beta - \frac{1}{2}\alpha - 1 = 0 \\
\Rightarrow \ \alpha = \frac{2}{9}&,\ \beta = -\frac{10}{9}
\end{aligned}$$

Therefore the optimal value of the dual problem is:

$$\max_{\alpha,\beta} \min_{x_1,x_2} \ L(x_1, x_2, \alpha, \beta) = -\frac{4}{9}$$

The optimal solution is:

$$x_1 = \frac{2}{3},\ x_2 = \frac{1}{3}$$

1.2 Calculus

(1) Prove the recursion formula of Gamma function:

$$\begin{aligned}
\Gamma(x+1) &= \int_0^\infty u^{x-1} e^{-u} du \\
&= -\int_0^\infty u^{x-1} de^{-u} \\
&= -u^x e^{-u}\Big|_0^\infty + \int_0^\infty x u^{x-1} e^{-u} du \\
&= \lim_{u\to\infty} -u^x e^{-u} + x\int_0^\infty u^{x-1} e^{-u} du \\
&= x\int_0^\infty u^{x-1} e^{-u} du \\
&= x\Gamma(x)
\end{aligned}$$

(2) Prove the relationship between gamma function and beta function:

$$\begin{aligned}
\Gamma(a)\Gamma(b) &= \int_{u=0}^\infty e^{-u} u^{a-1} du \int_{v=0}^\infty e^{-v} v^{b-1} dv \\
&= \int_{v=0}^\infty \int_{u=0}^\infty e^{-u-v} u^{a-1} v^{b-1} du dv
\end{aligned}$$

Changing variables by $u = f(z, \mu) = z\mu$ and $v = g(z, \mu) = z(1-\mu)$, and let $|J(z, \mu)|$ be the absolute value of the Jacobian determinant of $u = f(z, \mu)$ and $v = g(z, \mu)$:

$$\begin{aligned}\Gamma(a)\Gamma(b) &= \int_{z=0}^{\infty}\int_{\mu=0}^{1} e^{-z}(z\mu)^{a-1}[z(1-\mu)]^{b-1}|J(z,\mu)|d\mu dz \\ &= \int_{z=0}^{\infty}\int_{\mu=0}^{1} e^{-z}(z\mu)^{a-1}[z(1-\mu)]^{b-1}zd\mu dz \\ &= \int_{z=0}^{\infty} e^{-z}(z\mu)^{a-1}z^{a+b-1}dz \int_{\mu=0}^{1} \mu^{a-1}(1-\mu)^{b-1}d\mu \\ &= \Gamma(a+b)\int_{\mu=0}^{1} \mu^{a-1}(1-\mu)^{b-1}d\mu\end{aligned}$$

Therefore there is:

$$\mathrm{Beta}(a,b) = \int_0^1 \mu^{a-1}(1-\mu)^{b-1}d\mu = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

This result implies that the Beta distribution is normalized.

1.3 Probability

The prior distribution $p(\lambda; \alpha, \beta)$ follows Gamma distribution:

$$p(\lambda; \alpha, \beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}$$

The likelihood function $p(x|\lambda)$ follows Poisson distribution:

$$p(x|\lambda) = \frac{\lambda^x}{x!}e^{-\lambda}$$

Then the posterior distribution is:

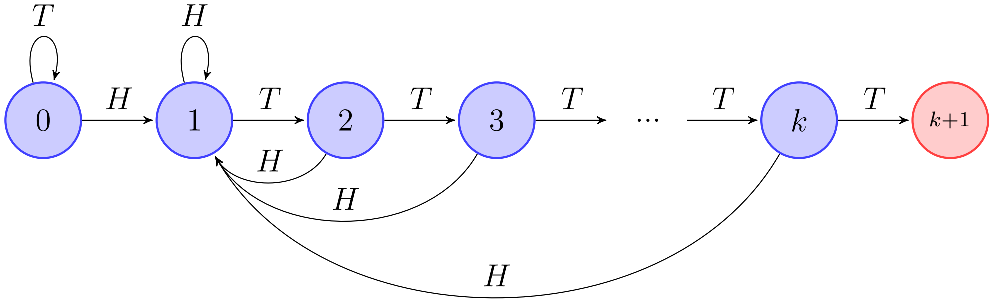
$$\begin{aligned}p(\lambda|x) &= \frac{p(x|\lambda)\,p(\lambda;\alpha,\beta)}{p(x)} \\ &= \frac{\beta^{\alpha}\lambda^{\alpha+x-1}e^{-(\beta+1)\lambda}}{\Gamma(\alpha)\,x!\,p(x)} \\ &\propto \lambda^{\alpha+x-1}e^{-(\beta+1)\lambda}\end{aligned}$$

Therefore the posterior distribution $p(\lambda|x)$ follows Gamma distribution $\Gamma(\lambda|\alpha + x, \beta + 1)$, which implies that the Gamma distribution can serve as a conjugate prior to the Poisson distribution.

1.4 Stochastic Process

We need to observe a consecutive pattern which has the length of $k + 1$, one H and k T.

Consider establishing a Markov process: the initial state means currently the coins are all T; there are $k + 1$ middle states, and the ith middle state means currently we have i coins that match the pattern; at last we have one final state, means that we have observed the given pattern. Obviously in this Markov process, the initial state and $k+1$ middle states are **non − recurrent states**, and the final state is an **absorption state**. In non-recurrent states, if we get another coin that matches the next term of the pattern, the state will change to the next state, otherwise the state will return to the initial state, and the two situations have equal probability. The diagram of this Markov process is shown as below, where all the probability of transitions in the diagram are 0.5:



The question turns to calculate the expectation time from a non-recurrent state to the absorption state. The transition matrix with $(k+2)\times(k+2)$ dimensions is:

$$\boldsymbol{P} = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0.5 & 0.5 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 & \cdots & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0.5 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0.5 & 0 & 0 & 0 & \cdots & 0.5 & 0 \\ 0 & 0.5 & 0 & 0 & 0 & \cdots & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix}_{(k+2)\times(k+2)}$$

Suppose $\boldsymbol{D}$ equals to the first $(k+1)\times(k+1)$ dimensions of $\boldsymbol{P}$:

$$\boldsymbol{D} = \begin{bmatrix} 0.5 & 0.5 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0.5 & 0.5 & 0 & 0 & \cdots & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 & \cdots & 0 \\ 0 & 0.5 & 0 & 0 & 0.5 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0.5 & 0 & 0 & 0 & \cdots & 0.5 \\ 0 & 0.5 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix}_{(k+1)\times(k+1)}$$

Let $\boldsymbol{g} = [1, 1, ..., 1]^T$, which has $(k+1)$ dimensions, and $\boldsymbol{I}$ is a unit matrix with $(k+1)\times(k+1)$ dimensions. The expected number of total tosses equals to the first number of the vector

$(\boldsymbol{I}-\boldsymbol{D})^{-1}\boldsymbol{g}$. Let $\boldsymbol{Q}=\boldsymbol{I}-\boldsymbol{D}$, there is:

$$\boldsymbol{Q}=\begin{bmatrix} 0.5 & -0.5 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0.5 & -0.5 & 0 & 0 & \cdots & 0 \\ 0 & -0.5 & 1 & -0.5 & 0 & \cdots & 0 \\ 0 & -0.5 & 0 & 1 & -0.5 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & -0.5 & 0 & 0 & 0 & \cdots & -0.5 \\ 0 & -0.5 & 0 & 0 & 0 & \cdots & 1 \end{bmatrix}_{(k+1)\times(k+1)}$$

Therefore:

$$(\boldsymbol{I}-\boldsymbol{D})^{-1}\boldsymbol{g}=\boldsymbol{Q}^{-1}\boldsymbol{g}=\frac{\boldsymbol{Q}^{*}}{\det(\boldsymbol{Q})}\boldsymbol{g}=2^{k+1}\boldsymbol{Q}^{*}\boldsymbol{g}$$

in this equation $\boldsymbol{Q}^{*}=(\boldsymbol{Q}_{ij})_{(k+1)\times(k+1)}$, and $\boldsymbol{Q}_{ij}=(-1)^{i+j}\boldsymbol{M}_{ij}$, where $\boldsymbol{M}_{ij}$ is the determinant of $\boldsymbol{Q}$ when deleting the ith row and the jth column. Then the first number of $2^{k+1}\boldsymbol{Q}^{*}\boldsymbol{g}$ is the expectation time, which is found to be 2^{k+1}.

Or there is another easier way to get the result. Suppose $t(i)$ is the expected time from the state i to the absorption state $k+1$. Suppose p_{ij} is the transition probability from state i to j, and the specific value can be obtained by the matrix P shown before. There is:

$$t(i)=1+\sum_{j\in\{j:i\to j\}}p_{i,j}t(j)$$

Given that $t(k+1)=0$, we can get $t(0)=2^{k+1}$.

2 SVM

Consider the regression problem with training data $\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{N}$ where $\boldsymbol{x}_i\in\mathbb{R}^d, y_i\in\mathbb{R}$. The primal problem of linear SVM is:

$$\begin{aligned} &\min_{\boldsymbol{\omega},b,\boldsymbol{\xi},\hat{\boldsymbol{\xi}}}\frac{1}{2}\|\boldsymbol{\omega}\|^2+C\sum_{i=1}^{N}\left(\xi_i+\hat{\xi}_i\right) \\ &s.t.\ y_i\le\boldsymbol{\omega}^T\boldsymbol{x}_i+b+\epsilon+\xi_i,\ 1\le i\le N \\ &\qquad y_i\ge\boldsymbol{\omega}^T\boldsymbol{x}_i+b-\epsilon-\hat{\xi}_i,\ 1\le i\le N \\ &\qquad \xi_i\ge 0,\ 1\le i\le N \\ &\qquad \hat{\xi}_i\ge 0,\ 1\le i\le N \end{aligned}$$

where C is a constant and $\epsilon>0$ denotes a fixed small value.

The corresponding Lagrangian of SVM is:

$$\begin{aligned}
L(\boldsymbol{\omega}, b, \boldsymbol{\xi}, \hat{\boldsymbol{\xi}}, \boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}, \boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) &= \frac{1}{2}\|\boldsymbol{\omega}\|^2 + C\sum_{i=1}^{N}\left(\xi_i + \hat{\xi}_i\right) + \sum_{i=1}^{N}\alpha_i\left[y_i - \left(\boldsymbol{\omega}^T\boldsymbol{x}_i + b + \epsilon + \xi_i\right)\right] \\
&\quad - \sum_{i=1}^{N}\hat{\alpha}_i\left[y_i - \left(\boldsymbol{\omega}^T\boldsymbol{x}_i + b - \epsilon - \hat{\xi}_i\right)\right] - \sum_{i=1}^{N}\beta_i\xi_i - \sum_{i=1}^{N}\hat{\beta}_i\hat{\xi}_i \\
&= \frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\omega} + C\left(\boldsymbol{\xi} + \hat{\boldsymbol{\xi}}\right)^T\mathbf{1} + \sum_{i=1}^{N}\alpha_i\left[y_i - \left(\boldsymbol{\omega}^T\boldsymbol{x}_i + b + \epsilon + \xi_i\right)\right] \\
&\quad - \sum_{i=1}^{N}\hat{\alpha}_i\left[y_i - \left(\boldsymbol{\omega}^T\boldsymbol{x}_i + b - \epsilon - \hat{\xi}_i\right)\right] - \boldsymbol{\xi}^T\boldsymbol{\beta} - \hat{\boldsymbol{\xi}}^T\hat{\boldsymbol{\beta}}
\end{aligned}$$

where $\boldsymbol{\alpha} \geq \mathbf{0}$, $\hat{\boldsymbol{\alpha}} \geq \mathbf{0}$, $\boldsymbol{\beta} \geq \mathbf{0}$, $\hat{\boldsymbol{\beta}} \geq \mathbf{0}$ are Lagrange multipliers.

The dual Lagrangian problem of SVM is:

$$\max_{\boldsymbol{\alpha},\hat{\boldsymbol{\alpha}},\boldsymbol{\beta},\hat{\boldsymbol{\beta}}}\ \min_{\boldsymbol{\omega},b,\boldsymbol{\xi},\hat{\boldsymbol{\xi}}} L(\boldsymbol{\omega}, b, \boldsymbol{\xi}, \hat{\boldsymbol{\xi}}, \boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}, \boldsymbol{\beta}, \hat{\boldsymbol{\beta}})$$

To solve the dual Lagrangian problem, the dual function of SVM needs to be derived first:

$$g(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}, \boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) = \min_{\boldsymbol{\omega},b,\boldsymbol{\xi},\hat{\boldsymbol{\xi}}} L(\boldsymbol{\omega}, b, \boldsymbol{\xi}, \hat{\boldsymbol{\xi}}, \boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}, \boldsymbol{\beta}, \hat{\boldsymbol{\beta}})$$

Let the derivatives of L w.r.t. $\boldsymbol{\omega}$, $b, \boldsymbol{\xi}, \hat{\boldsymbol{\xi}}$ equal to zero:

$$\begin{aligned}
\left.\frac{\partial L}{\partial \boldsymbol{\omega}}\right|_{\boldsymbol{\omega}^*} &= \boldsymbol{\omega}^* - \sum_{i=1}^{N}\left(\alpha_i - \hat{\alpha}_i\right)\boldsymbol{x}_i = 0 \ \Rightarrow\ \boldsymbol{\omega}^* = \sum_{i=1}^{N}\left(\alpha_i - \hat{\alpha}_i\right)\boldsymbol{x}_i \\
\left.\frac{\partial L}{\partial b}\right|_{b^*} &= 0 \ \Rightarrow\ \sum_{i=1}^{N}\left(\alpha_i - \hat{\alpha}_i\right) = 0 \\
\left.\frac{\partial L}{\partial \boldsymbol{\xi}}\right|_{\boldsymbol{\xi}^*} &= C\mathbf{1} - (\boldsymbol{\alpha} + \boldsymbol{\beta}) = 0 \ \Rightarrow\ \begin{cases} \boldsymbol{\beta} = C\mathbf{1} - \boldsymbol{\alpha} \\ 0 \leq \boldsymbol{\alpha} \leq C\mathbf{1} \end{cases} \\
\left.\frac{\partial L}{\partial \hat{\boldsymbol{\xi}}}\right|_{\hat{\boldsymbol{\xi}}^*} &= C\mathbf{1} - (\hat{\boldsymbol{\alpha}} + \hat{\boldsymbol{\beta}}) = 0 \ \Rightarrow\ \begin{cases} \hat{\boldsymbol{\beta}} = C\mathbf{1} - \hat{\boldsymbol{\alpha}} \\ 0 \leq \hat{\boldsymbol{\alpha}} \leq C\mathbf{1} \end{cases}
\end{aligned}$$

So the dual function is:

$$\begin{aligned}
g(\boldsymbol{\alpha},\hat{\boldsymbol{\alpha}},\boldsymbol{\beta},\hat{\boldsymbol{\beta}}) &= L(\boldsymbol{\omega}^*, b^*, \boldsymbol{\xi}^*, \hat{\boldsymbol{\xi}}^*, \boldsymbol{\alpha},\hat{\boldsymbol{\alpha}},\boldsymbol{\beta},\hat{\boldsymbol{\beta}}) \\
&= \frac{1}{2}\|\boldsymbol{\omega}^*\|^2 + \sum_{i=1}^{N}\left(\alpha_i-\hat{\alpha}_i\right)\left[y_i - \left(\boldsymbol{\omega}^{*T}\boldsymbol{x}_i + b^*\right)\right] + \epsilon\left(\boldsymbol{\alpha}+\hat{\boldsymbol{\alpha}}\right)^T\mathbf{1} \\
&= \frac{1}{2}\|\boldsymbol{\omega}^*\|^2 + \sum_{i=1}^{N}\left(\alpha_i-\hat{\alpha}_i\right)\left(y_i - \boldsymbol{\omega}^{*T}\boldsymbol{x}_i\right) + \epsilon\left(\boldsymbol{\alpha}+\hat{\boldsymbol{\alpha}}\right)^T\mathbf{1} \\
&= \frac{1}{2}\left\|\sum_{i=1}^{N}\left(\alpha_i-\hat{\alpha}_i\right)\boldsymbol{x}_i\right\|^2 + \sum_{i=1}^{N}\left(\alpha_i-\hat{\alpha}_i\right)\left\{y_i - \left[\sum_{j=1}^{N}\left(\alpha_j-\hat{\alpha}_j\right)\boldsymbol{x}_j\right]^T\boldsymbol{x}_j\right\} + \epsilon\left(\boldsymbol{\alpha}+\hat{\boldsymbol{\alpha}}\right)^T\mathbf{1} \\
&= \epsilon\left(\boldsymbol{\alpha}+\hat{\boldsymbol{\alpha}}\right)^T\mathbf{1} + \left(\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}}\right)^T\boldsymbol{y} - \frac{1}{2}\left\|\sum_{i=1}^{N}\left(\alpha_i-\hat{\alpha}_i\right)\boldsymbol{x}_i\right\|^2 \\
&= \epsilon\left(\boldsymbol{\alpha}+\hat{\boldsymbol{\alpha}}\right)^T\mathbf{1} + \left(\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}}\right)^T\boldsymbol{y} - \frac{1}{2}\left(\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}}\right)^T\boldsymbol{G}\left(\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}}\right)
\end{aligned}$$

where $\boldsymbol{G}\in\mathbb{R}^{N\times N}$, $G_{ij}=\boldsymbol{x}_i^T\boldsymbol{x}_j$.

Therefore the dual problem turns to:

$$\begin{aligned}
&\max_{\boldsymbol{\alpha},\hat{\boldsymbol{\alpha}}}\ \epsilon\left(\boldsymbol{\alpha}+\hat{\boldsymbol{\alpha}}\right)^T\mathbf{1} + \left(\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}}\right)^T\boldsymbol{y} - \frac{1}{2}\left(\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}}\right)^T\boldsymbol{G}\left(\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}}\right) \\
&s.t.\ 0\le\boldsymbol{\alpha}\le C\mathbf{1},\ 0\le\hat{\boldsymbol{\alpha}}\le C\mathbf{1},\ \left(\boldsymbol{\alpha}-\hat{\boldsymbol{\alpha}}\right)^T\mathbf{1}=0
\end{aligned}$$

where the definitions and constrains of $\boldsymbol{G}$ is shown before.

3 IRLS for Logistic Regression

For a binary classification problem $\{(x_i,y_i)\}_{i=1}^N$ ($x_i\in\mathbb{R}^d$, $y_i\in\{0,1\}$), the probabilistic decision rule according to logistic regression is:

$$P_{\boldsymbol{\omega}}\left(y|\boldsymbol{x}\right) = \frac{\exp\left(y\boldsymbol{\omega}^T\boldsymbol{x}\right)}{1+\exp\left(\boldsymbol{\omega}^T\boldsymbol{x}\right)}$$

And hence the log-likelihood is:

$$\begin{aligned}
\mathcal{L}\left(\boldsymbol{\omega}\right) &= \log\prod_{i=1}^{N}P_{\boldsymbol{\omega}}\left(y_i|\boldsymbol{x}_i\right) \\
&= \sum_{i=1}^{N}\left\{y_i\boldsymbol{\omega}^T\boldsymbol{x}_i - \log\left[1+\exp\left(\boldsymbol{\omega}^T\boldsymbol{x}_i\right)\right]\right\}
\end{aligned}$$

The gradient of $\mathcal{L}(\boldsymbol{\omega})$ w.r.t. $\boldsymbol{\omega}$ is:

$$\begin{aligned}\boldsymbol{g} &= \frac{\partial \mathcal{L}(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}} \\ &= \sum_{i=1}^{N}\left[y_i \boldsymbol{x}_i - \frac{\exp\left(\boldsymbol{\omega}^T \boldsymbol{x}_i\right)\boldsymbol{x}_i}{1+\exp\left(\boldsymbol{\omega}^T \boldsymbol{x}_i\right)}\right] \\ &= \sum_{i=1}^{N}\left[y_i - \sigma\left(\boldsymbol{\omega}^T \boldsymbol{x}_i\right)\right]\boldsymbol{x}_i \\ &= \boldsymbol{X}(\boldsymbol{y}-\boldsymbol{\mu})\end{aligned}$$

where $\sigma(a) = \frac{\exp(a)}{1+\exp(a)}$ is the sigmoid function and there is $\frac{d\sigma}{da} = \sigma(1-\sigma)$; $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N]_{d\times N}$, $\mu_i = \sigma\left(\boldsymbol{\omega}^T \boldsymbol{x}_i\right)$.

The Hessian matrix of $\mathcal{L}(\boldsymbol{\omega})$ is:

$$\begin{aligned}\boldsymbol{H} &= \frac{\partial^2 \mathcal{L}(\boldsymbol{\omega})}{\partial \boldsymbol{\omega}\partial \boldsymbol{\omega}^T} \\ &= \frac{\partial\left[\boldsymbol{X}(\boldsymbol{y}-\boldsymbol{\mu})\right]}{\partial \boldsymbol{\omega}^T} \\ &= -\boldsymbol{X}\mathrm{diag}\left\{\mu_i(1-\mu_i)\boldsymbol{x}_i^T\right\} \\ &= -\boldsymbol{X}\boldsymbol{R}\boldsymbol{X}^T\end{aligned}$$

where $R_{ii} = \mu_i(1-\mu_i) = \sigma\left(\boldsymbol{\omega}^T \boldsymbol{x}_i\right)\left[1-\sigma\left(\boldsymbol{\omega}^T \boldsymbol{x}_i\right)\right]$.

In least square estimate of linear regression, we have:

$$\boldsymbol{\omega} = \left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{y}$$

In IRLS, we have:

$$\begin{aligned}\boldsymbol{\omega}_{t+1} &= \boldsymbol{\omega}_t - \boldsymbol{H}^{-1}\boldsymbol{g} \\ &= \boldsymbol{\omega}_t + \left(\boldsymbol{X}\boldsymbol{R}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}(\boldsymbol{y}-\boldsymbol{\mu}) \\ &= \left(\boldsymbol{X}\boldsymbol{R}\boldsymbol{X}^T\right)^{-1}\left[\boldsymbol{X}\boldsymbol{R}\boldsymbol{X}^T\boldsymbol{\omega}_t + \boldsymbol{X}(\boldsymbol{y}-\boldsymbol{\mu})\right] \\ &= \left(\boldsymbol{X}\boldsymbol{R}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{R}\boldsymbol{z}\end{aligned}$$

where $\boldsymbol{z} = \boldsymbol{X}^T\boldsymbol{\omega}_t + \boldsymbol{R}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})$.

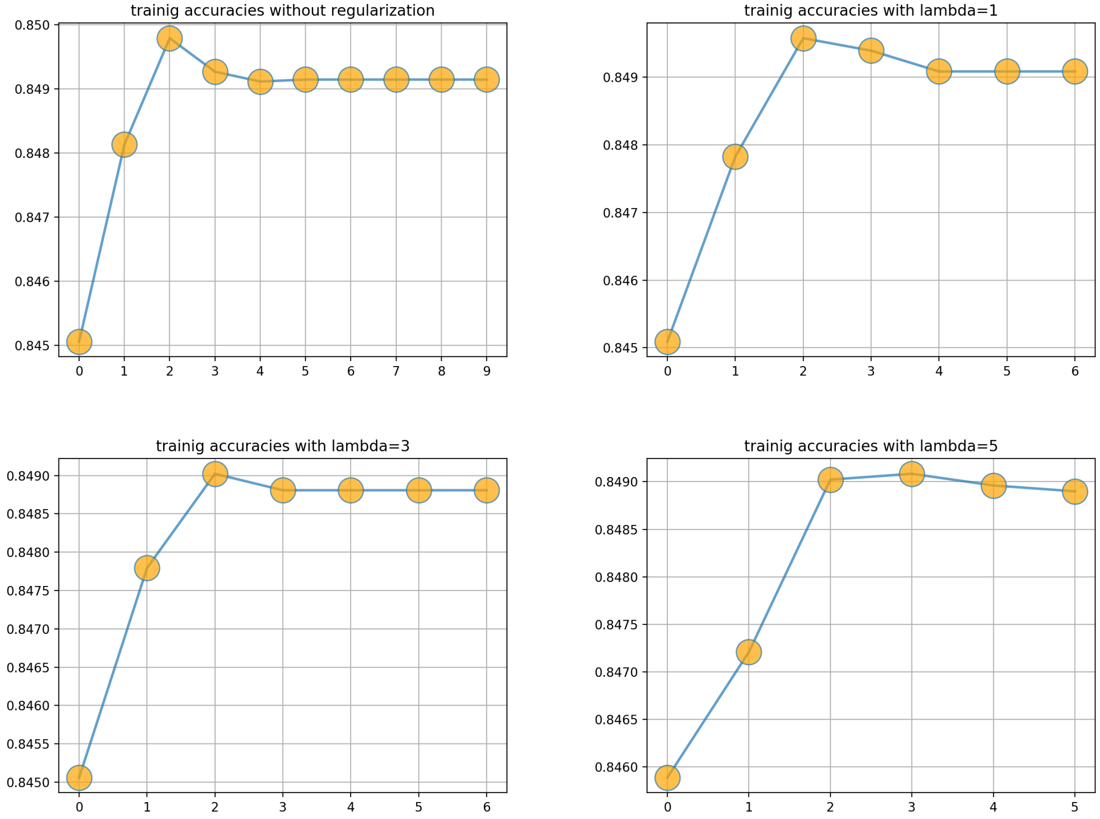
For the L2-norm regularized logistic regression $-\frac{\lambda}{2}\|\boldsymbol{\omega}\|_2^2 + \mathcal{L}(\boldsymbol{\omega})$, the gradients and Hessian matrix are:

$$\begin{aligned}\hat{\boldsymbol{g}} &= -\lambda\boldsymbol{\omega} + \boldsymbol{X}(\boldsymbol{y}-\boldsymbol{\mu}) \\ \hat{\boldsymbol{H}} &= -\lambda\boldsymbol{I} - \boldsymbol{X}\boldsymbol{R}\boldsymbol{X}^T\end{aligned}$$

The IRLS procedure turns to:

$$\begin{aligned}
\hat{\boldsymbol{\omega}}_{t+1} &= \hat{\boldsymbol{\omega}}_t - \hat{\boldsymbol{H}}^{-1}\hat{\boldsymbol{g}} \\
&= \hat{\boldsymbol{\omega}}_t + \left(\lambda \boldsymbol{I} + \boldsymbol{X}\boldsymbol{R}\boldsymbol{X}^T\right)^{-1}\left[-\lambda\hat{\boldsymbol{\omega}}_t + \boldsymbol{X}\left(\boldsymbol{y}-\boldsymbol{\mu}\right)\right] \\
&= \left(\lambda \boldsymbol{I} + \boldsymbol{X}\boldsymbol{R}\boldsymbol{X}^T\right)^{-1}\left[\left(\lambda \boldsymbol{I} + \boldsymbol{X}\boldsymbol{R}\boldsymbol{X}^T\right)\hat{\boldsymbol{\omega}}_t - \lambda\hat{\boldsymbol{\omega}}_t + \boldsymbol{X}\left(\boldsymbol{y}-\boldsymbol{\mu}\right)\right] \\
&= \left(\lambda \boldsymbol{I} + \boldsymbol{X}\boldsymbol{R}\boldsymbol{X}^T\right)^{-1}\left[\boldsymbol{X}\boldsymbol{R}\boldsymbol{X}^T\hat{\boldsymbol{\omega}}_t + \boldsymbol{X}\left(\boldsymbol{y}-\boldsymbol{\mu}\right)\right] \\
&= \left(\lambda \boldsymbol{I} + \boldsymbol{X}\boldsymbol{R}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{R}\boldsymbol{z}
\end{aligned}$$

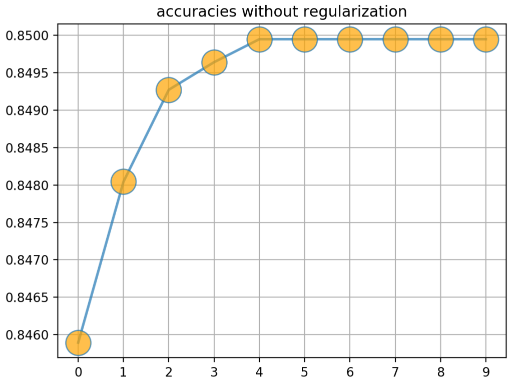
where $\boldsymbol{z} = \boldsymbol{X}^T\boldsymbol{\omega}_t + \boldsymbol{R}^{-1}\left(\boldsymbol{y}-\boldsymbol{\mu}\right)$. The model can reach 85% on the test data of UCI a9a dataset. When there is no regularization, $\|\omega\|^2 = 19.108$, and with $\lambda = 1, 3, 5, 10, 20, 30$, $\|\omega\|^2 = 6.222, 5.412, 5.061, 4.652, 4.270, 4.035$ respectively. The training accuracies with different values of lambda are shown below:

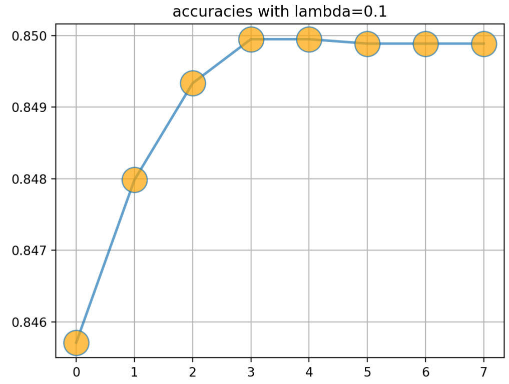


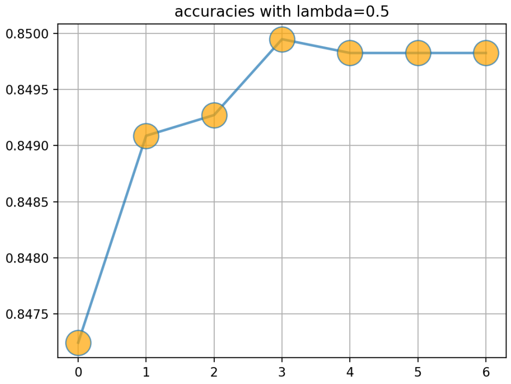
As we can find, the update iteration of ω gets smaller when λ gets larger.

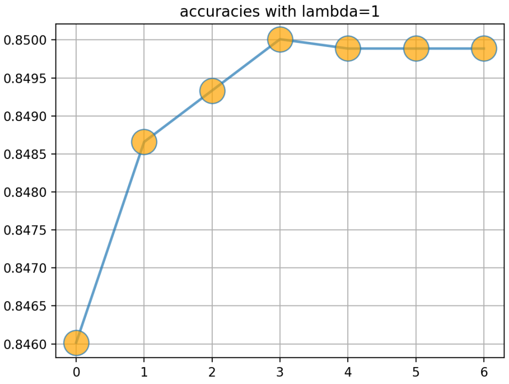
Then we test the weights on the test data, and adjust the value of λ to see the optimal accuracy, as shown in the next page.

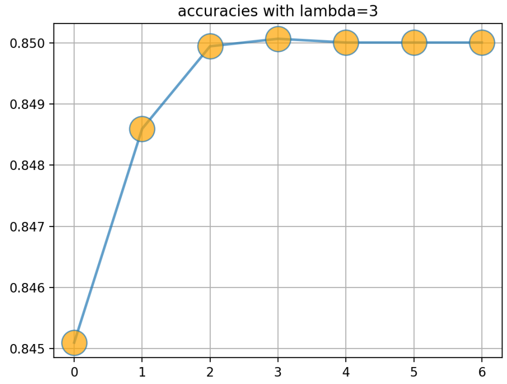
The testing accuracies with different values of lambda are shown below:

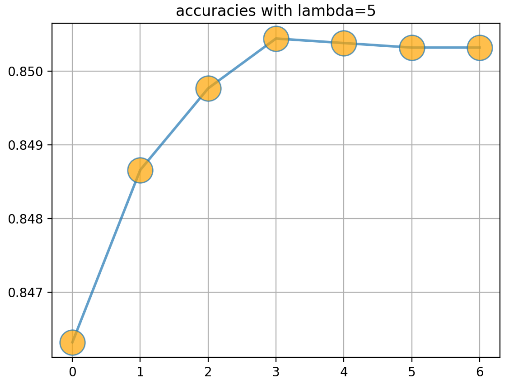




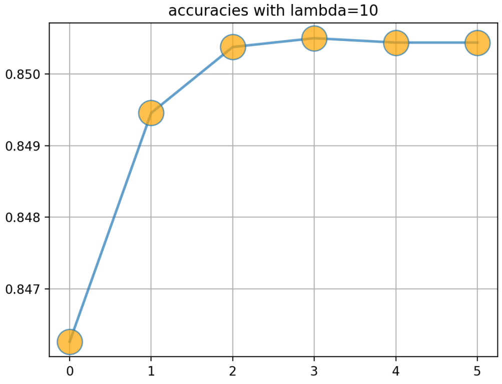


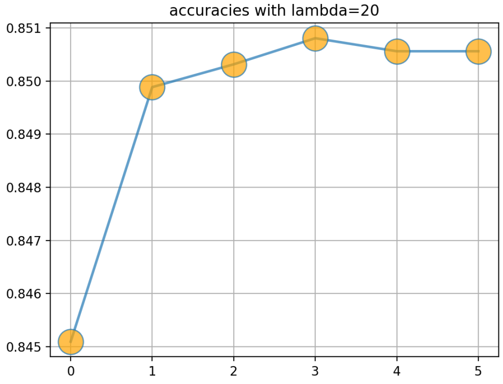


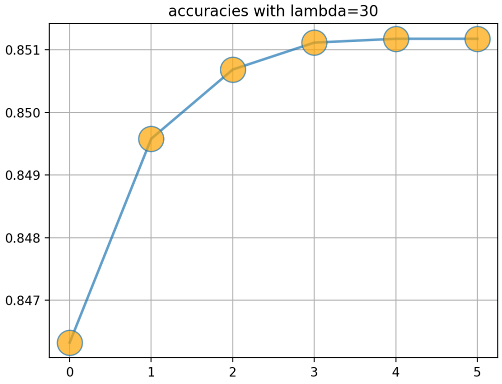


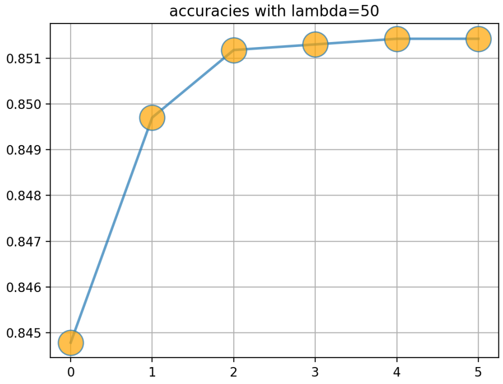


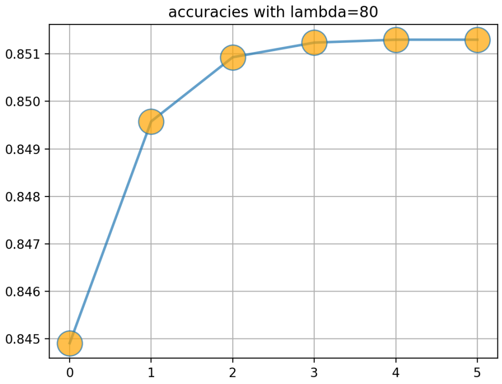
Therefore when $\lambda \in [30, 80]$, the testing accuracy seems to reach the optimal value 85.1%.

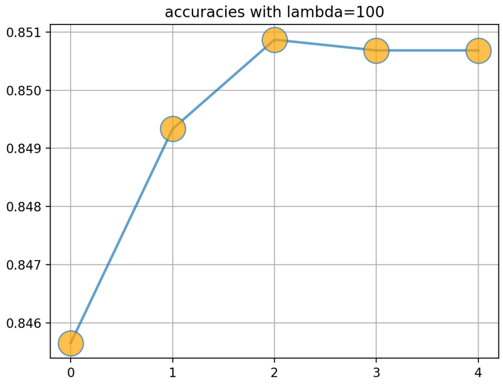

accuracies with lambda=10


accuracies with lambda=20


accuracies with lambda=30


accuracies with lambda=50


accuracies with lambda=80


accuracies with lambda=100