Statistical Machine Learning Homework1

1.1 Optimization

The objective with Lagrange multiplers is

$$\min_{x_1,x_2}\max_{\lambda_1,\lambda_2} L = x_1^2 + x_2^2 - 1 + \lambda_1(x_1 + x_2 - 1) - \lambda_2(x_1 - 2x_2) \tag{1}$$

thus using KKT codnition we get

$$\begin{aligned}
\frac{\partial L}{\partial x_1} = 2x_1 - \lambda_1 - \lambda_2 = 0 && (2)\\
\frac{\partial L}{\partial x_2} = 2x_2 - \lambda_1 + 2\lambda_2 = 0 && (3)\\
\lambda_1, \lambda_2 \geq 0 && (4)\\
\lambda_1(x_1 + x_2 - 1) = 0 && (5)\\
\lambda_2(x_1 - 2x_2) = 0 && (6)\\
x_1 + x_2 - 1 \geq 0 && (7)\\
x_1 \geq 2x_2 && (8)
\end{aligned}$$

from Eq.(2), (3), we have

$$\begin{aligned}
x_1 = \frac{\lambda_1 + \lambda_2}{2} && (9)\\
x_2 = \frac{\lambda_1 - 2\lambda_2}{2} && (10)
\end{aligned}$$

put them into Eq.$(5) - (8)$, we have

$$\begin{aligned}
\lambda_1(\lambda_1 - \frac{\lambda_2}{2} - 1) = 0 && (11)\\
\lambda_1 - \frac{\lambda_2}{2} - 1 \geq 0 && (12)\\
\lambda_2(-\frac{\lambda_1}{2} + \frac{5\lambda_2}{2}) = 0 && (13)\\
\frac{-\lambda_1}{2} + \frac{5\lambda_2}{2} \geq 0 && (14)
\end{aligned}$$

let's discuss about Eq.(11), (13)

1. if $\lambda_1, \lambda_2 = 0$, it contradicts (12)
2. if $\lambda_1 = 0, -\frac{\lambda_1}{2} + \frac{5\lambda_2}{2} = 0$, we get $\lambda_2 = 0$, it contradicts (12)
3. if $\lambda_1 - \frac{\lambda_2}{2} - 1 = 0, \lambda_2 = 0$, we get $\lambda_1 = 1,$, it contradicts (14)
4. if $\lambda_1 - \frac{\lambda_2}{2} - 1 = 0, -\frac{\lambda_1}{2} + \frac{5\lambda_2}{2} = 0$, we get $\lambda_1 = \frac{10}{9}, \lambda_2 = \frac{2}{9}$, which doesn't contradict $(12), (14)$

thus $\lambda_1 = \frac{10}{9}, \lambda_2 = \frac{2}{9}$. Take them into $(9), (10)$, we have

$$\begin{aligned} x_1^\star &= \frac{2}{3} \\ x_2^\star &= \frac{1}{3} \\ L^\star &= -\frac{4}{9} \end{aligned}$$

1.2 Calculus

1 Proof.

$$\begin{aligned} \Gamma(x+1) &= \int_0^\infty u^x e^{-u} \mathrm{d}u && (15) \\ &= -\int_0^\infty u^x \mathrm{d}e^{-u} = -u^x e^{-u}\Big|_0^\infty + \int_0^\infty e^{-u} \mathrm{d}u^x && (16) \\ &= 0 + x\int_0^\infty e^{-u} u^{x-1} \mathrm{d}u = x\Gamma(x-1) && (17) \end{aligned}$$

2 Proof.

$$\begin{aligned} \Gamma(a)\Gamma(b) &= \int_0^\infty e^{-x} x^{a-1} \mathrm{d}x \int_0^\infty e^{-y} y^{b-1} \mathrm{d}y && (18) \\ &= \int_0^\infty x^{a-1} \mathrm{d}x \int_0^\infty e^{-(x+y)} y^{b-1} \mathrm{d}y && (19) \\ &= \int_0^\infty x^{a-1} \mathrm{d}x \int_x^\infty e^{-t} (t-x)^{b-1} \mathrm{d}t \quad (\text{let } t = x+y) && (20) \\ &= \int_0^\infty e^{-t} [\int_0^t (t-x)^{b-1} x^{a-1} \mathrm{d}x] \mathrm{d}t && (21) \\ &= \int_0^\infty e^{-t} [\int_0^1 (t-xt)^{b-1} (xt)^{a-1} \mathrm{d}(xt)] \mathrm{d}t && (22) \\ &= \int_0^\infty e^{-t} t^{a+b-1} \mathrm{d}t \int_0^1 (1-x)^{b-1} x^{a-1} \mathrm{d}x && (23) \\ &= \Gamma(a+b) \int_0^1 (1-x)^{b-1} x^{a-1} \mathrm{d}x && (24) \end{aligned}$$

that is (change x by μ)

$$\int_0^1 \mu^{a-1} (1-\mu)^{b-1} \mathrm{d}\mu = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \qquad (25)$$

1.3 Probability

prior $\lambda \sim \Gamma(\lambda|\alpha, \beta)$, so $p(\lambda) = \frac{e^{-\lambda\beta}\lambda^{\alpha-1}\beta^{\alpha}}{\Gamma(a)}$,

likelihood $x|\lambda \sim Poisson(x|\lambda)$, so $p(x|\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}$

so the joint distribution $p(\lambda, x) = p(x|\lambda)p(\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!}\frac{e^{-\lambda\beta}\lambda^{\alpha-1}\beta^{\alpha}}{\Gamma(a)} = \frac{e^{-\lambda(1+\beta)}\lambda^{(x+\alpha-1)}\beta^{\alpha}}{x!\Gamma(a)}$

so the margin $p(x) = \int_0^\infty p(\lambda, x)\mathrm{d}\lambda = \frac{\int_0^\infty e^{-\lambda(1+\beta)}\lambda^{(x+\alpha-1)}\beta^{\alpha}\mathrm{d}\lambda}{x!\Gamma(a)} = \frac{\Gamma(\alpha+x-1)\beta^{\alpha}}{\Gamma(\alpha)(1+\beta)^{\alpha}x!}$

so the posterior $p(\lambda|x) = \frac{p(\lambda, x)}{p(x)} = \frac{e^{-\lambda(1+\beta)}\lambda^{(x+\alpha-1)}\beta^{\alpha}}{x!\Gamma(a)} / \frac{\Gamma(\alpha+x-1)\beta^{\alpha}}{\Gamma(\alpha)(1+\beta)^{\alpha}x!} = \frac{e^{-\lambda(1+\beta)}\lambda^{(x+\alpha-1)}(1+\beta)^{\alpha}}{\Gamma(a+x)} = \Gamma(\lambda|\alpha+x, \beta+1)$

1.4 Stochastic Process

Noatation:k times of consecutive : T^k. starting state with no toss before: O.

- Let's consider the expected numebr b_k of tosses to get the first HT^k given the last toss is H.

HT^k changes to HT^{k+1} with probability $1/2$, and changes to HT^kH (it equals returning to the starting state) with with probability $1/2$, thus we get a recursive formula

$$b_{k+1} = (b_k + 1)/2 + (b_k + 1 + b_{k+1})/2 \tag{26}$$

the solution is $b_k = 2^{k+1} - 2$

- Then let's consider the expected numebr a_k of tosses to get the first HT^k starting from no toss before.

$\{O\ or\ T\}$ changes to H with probability $1/2$, and changes to T with probability $1/2$. If we get H, the problem becomes the situation mentioned above. Thus we get a recursive formula

$$a_k = (1 + b_k)/2 + (1 + a_k)/2 \tag{27}$$

the solution is $a_k = b_k + 2 = 2^{k+1}$.

So, the expected number of toss to get the first HT^k starting from no toss before, is 2^{k+1}.

2 SVM

The objective with Lagrange multipliers is

$$\begin{aligned}\min_{\boldsymbol{w},\boldsymbol{b},\boldsymbol{\xi},\hat{\boldsymbol{\xi}}}\ \max_{\boldsymbol{\alpha},\hat{\boldsymbol{\alpha}},\boldsymbol{\beta},\hat{\boldsymbol{\beta}}} L = &\frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{N}(\xi_i + \hat{\xi}_i) \\ &- \sum_{i=1}^{N}[\alpha_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i + b + \epsilon + \xi_i - y_i) + \hat{\alpha}_i(-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i - b + \epsilon + \hat{\xi}_i + y_i) + \beta_i\xi_i + \hat{\beta}_i\hat{\xi}_i] \quad (28) \\ \text{s.t.} \quad &\alpha_i, \hat{\alpha}_i, \beta_i, \hat{\beta}_i \geq 0\end{aligned}$$

Since L is convex for $\boldsymbol{w}, \boldsymbol{b}, \boldsymbol{\xi}, \hat{\boldsymbol{\xi}}$ can concave (exactly linear) for $\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}$, we exchange $\max$ and $\min$. For $\min_{\boldsymbol{w},\boldsymbol{b},\boldsymbol{\xi},\hat{\boldsymbol{\xi}}}$, we have

$$\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{N}(\alpha_i\boldsymbol{x}_i + \hat{\alpha}_i\hat{\boldsymbol{x}}_i) = 0 \quad (29)$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{N}(\alpha_i - \hat{\alpha}_i) = 0 \quad (30)$$

$$\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \beta_i = 0 \quad (31)$$

$$\frac{\partial L}{\partial \hat{\xi}_i} = C - \hat{\alpha}_i - \hat{\beta}_i = 0 \quad (32)$$

that is

$$\boldsymbol{w} = \sum_{i=1}^{N}(\alpha_i\boldsymbol{x}_i + \hat{\alpha}_i\hat{\boldsymbol{x}}_i) \quad (33)$$

$$\sum_{i=1}^{N}(\alpha_i - \hat{\alpha}_i) = 0 \quad (34)$$

$$\alpha_i + \beta_i = C \quad (35)$$

$$\hat{\alpha}_i + \hat{\beta}_i = C \quad (36)$$

$$0 \leq \alpha_i \leq C,\ 0 \leq \hat{\alpha}_i \leq C \quad (37)$$

take (33)-(36) into (28), we get

$$\begin{aligned}\max_{\boldsymbol{\alpha},\hat{\boldsymbol{\alpha}}} -\frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i - \hat{\alpha}_i)\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j(\alpha_j - \hat{\alpha}_j) + \sum_{i=1}^{N}(\alpha_i(y_i - \epsilon) - \hat{\alpha}_i(y_i + \epsilon)) \quad (38) \\ \text{s.t.} \quad 0 \leq \alpha_i \leq C,\ 0 \leq \hat{\alpha}_i \leq C \\ \sum_{i=1}^{N}(\alpha_i - \hat{\alpha}_i) = 0\end{aligned}$$

so the dual problem for support vector regression is

$$\begin{gathered}
\min_{\boldsymbol{\alpha},\hat{\boldsymbol{\alpha}}} \frac{1}{2}\sum_{i,j=1}^{N}(\alpha_i-\hat{\alpha}_i)\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j(\alpha_j-\hat{\alpha}_j)+\sum_{i=1}^{N}(\alpha_i(\epsilon-y_i)+\hat{\alpha}_i(y_i+\epsilon)) \\
\text{s.t.} \quad 0\le\alpha_i\le C,\ 0\le\hat{\alpha}_i\le C \\
\sum_{i=1}^{N}(\alpha_i-\hat{\alpha}_i)=0
\end{gathered} \tag{39}$$

3 IRLS for Logistic Regression

Derivaition

Feature $\boldsymbol{x}$, the last dimension is 1; weight $\boldsymbol{w}$, the last dimension is bias b; label $y \in 0, 1$. The prediction probability by logistic regression is:

$$P(y|\boldsymbol{x})=\frac{e^{y\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}}}{1+e^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}}} \tag{40}$$

IRLS with no regularization

the objective is to maximize the log-likelihood

$$\max_{\boldsymbol{w}} L(\boldsymbol{w})=\ln\prod_{i=1}^{N}P(y_i|\boldsymbol{x}_i)=\sum_{i=1}^{N}[y_i\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i-\ln(1+e^{\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i})] \tag{41}$$

then

$$\nabla_{\boldsymbol{w}}L=\sum_{i=1}^{N}(y_i\boldsymbol{x}_i-\boldsymbol{x}_i\sigma(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i))=\sum_{i=1}^{N}\boldsymbol{x}_i(y_i-\mu_i) \tag{42}$$

where $\mu_i=\sigma(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i)$, so the Hessian matrix is

$$H=\nabla_{\boldsymbol{w}}^2L=-\sum_{i=1}^{N}\boldsymbol{x}_i\mu_i(1-\mu_i)\boldsymbol{x}_i^{\mathrm{T}}=-XRX^{\mathrm{T}} \tag{43}$$

where $X=[\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_N]$, $\boldsymbol{x}_i$ is a column, diagonal matrix $R_{ii}=\mu_i(1-\mu_i)$. Thus from Newton's Method (44) we get IRLS (45) for logistic regression without regularization:m

$$\begin{aligned}
\boldsymbol{w}_{t+1}&=\boldsymbol{w}_t-H^{-1}\nabla_{\boldsymbol{w}}L|_{\boldsymbol{w}_t} && (44)\\
&=\boldsymbol{w}_t-(XRX^{\mathrm{T}})^{-1}X(\boldsymbol{\mu}-\boldsymbol{y})\\
&=(XRX^{\mathrm{T}})^{-1}[XRX^{\mathrm{T}}\boldsymbol{w}_t-X(\boldsymbol{\mu}-\boldsymbol{y})]\\
&=(XRX^{\mathrm{T}})^{-1}XR[X^{\mathrm{T}}\boldsymbol{w}_t-R^{-1}(\boldsymbol{\mu}-\boldsymbol{y})] && (45)
\end{aligned}$$

IRLS with l2-norm regularization

the objective is to maximize the log-likelihood

$$\max_{\boldsymbol{w}} \quad L(\boldsymbol{w}) - \frac{\lambda}{2}||\boldsymbol{w}||_2^2 = \sum_{i=1}^{N}[y_i \boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i - \ln(1 + e^{\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i})] - \frac{\lambda}{2}||\boldsymbol{w}||_2^2 \tag{46}$$

then

$$\nabla_{\boldsymbol{w}} L = \sum_{i=1}^{N}(y_i \boldsymbol{x}_i - \boldsymbol{x}_i \sigma(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i)) - \lambda \boldsymbol{w} = \sum_{i=1}^{N} \boldsymbol{x}_i (y_i - \mu_i) - \lambda \boldsymbol{w} \tag{47}$$

where $\mu_i = \sigma(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i)$, so the Hessian matrix is

$$H = \nabla_{\boldsymbol{w}}^2 L = -\sum_{i=1}^{N} \boldsymbol{x}_i \mu_i (1 - \mu_i) \boldsymbol{x}_i^{\mathrm{T}} - \lambda I = -XRX^{\mathrm{T}} - \lambda I \tag{48}$$

where diagonal matrix $R_{ii} = \mu_i(1 - \mu_i)$. Thus from Newton's Method (49) we get IRLS (50) for logistic regression without regularization:

$$\begin{aligned} \boldsymbol{w}_{t+1} &= \boldsymbol{w}_t - H^{-1} \nabla_{\boldsymbol{w}} L|_{\boldsymbol{w}_t} & (49) \\ &= \boldsymbol{w}_t - (XRX^{\mathrm{T}} + \lambda I)^{-1} X(\boldsymbol{\mu} - \boldsymbol{y}) \\ &= (XRX^{\mathrm{T}} + \lambda I)^{-1}[(XRX^{\mathrm{T}} + \lambda I)\boldsymbol{w}_t - X(\boldsymbol{\mu} - \boldsymbol{y})] \\ &= (XRX^{\mathrm{T}} + \lambda I)^{-1}\{XR[X^{\mathrm{T}} \boldsymbol{w}_t - R^{-1}(\boldsymbol{\mu} - \boldsymbol{y})] + \lambda \boldsymbol{w}_t\} & (50) \end{aligned}$$

Experiement

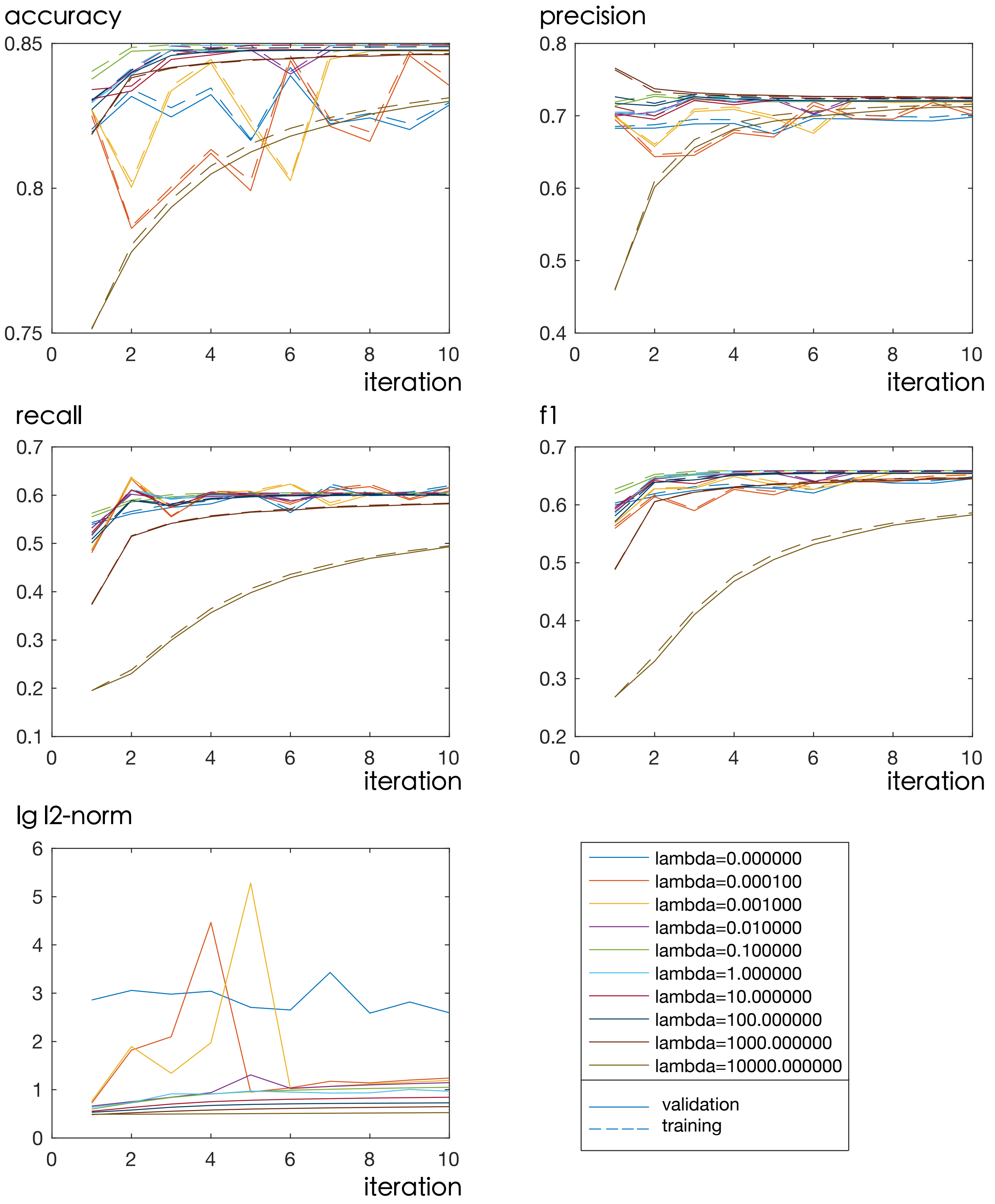
We using IRLS with and without (e.i. $\lambda = 0$) regularization to classify the binary dataset UCI a9a. The input is an 123-dimension binary vector, and the label is a 1-dimension binary value. We use accuracy, precision, recall and f1 score to evaluate the performance, and meanwhile report l2-norm $||\boldsymbol{w}||_2^2$. Since the proportion of positive data is about 0.3 which is unbalanced, we should pay attention to f1 score.

Cross Validation

I apply 20-fold cross validation on the training data of UCI a9a dataset to selection appropriate λ and iteration number.

Coarse Selection

To find the coarse range of λ, I set it to be 0.0001, 0.001, 0.01, 0.1, 1, 10, 100 ,1000, 10000 and 0. The results are shown in the figure below.



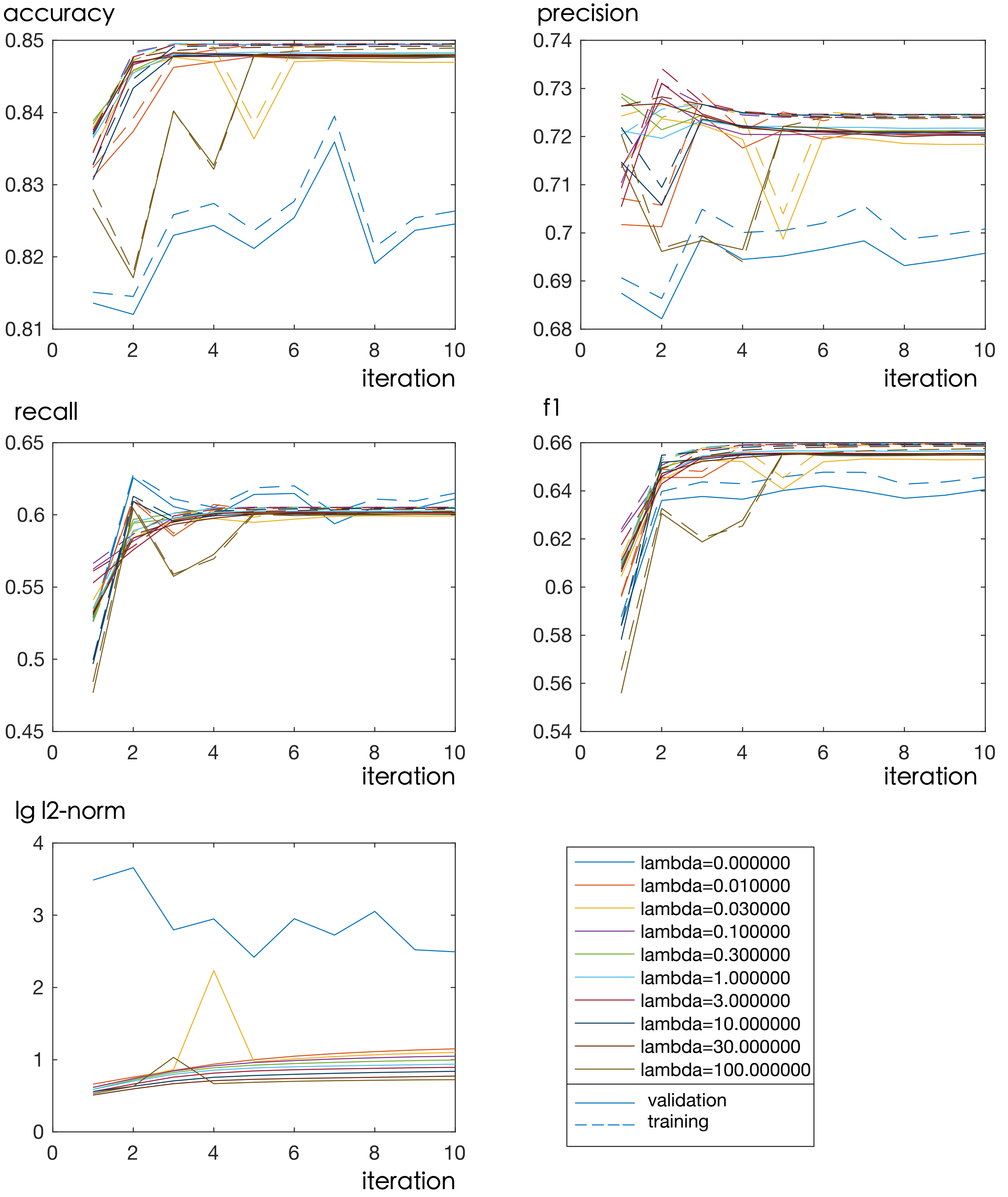
From the figure, we can see:

- $\lambda = 0.01 \sim 100$ perform well on accuracy, recall and f1. They all converges in 10 iteration steps (usually 6 steps are enough).
- IRLS with weaker or no regularization performs unstably during training, that is the evaluation indicators fluctuate with iteration number.
- IRLS with stronger regularization converges slowoy and under-fit the data, which results in low performance.

- The higher λ is, the lower l2-norm is at the same iteration, which is shown in the l2-norm - iteration curve above. No or low regularization occurs extremely large $||\boldsymbol{w}||_2^2$.

Fine Selection

Based on the coarse range $\lambda = 0.01 \sim 100$, we can find finer range. I set it to 0.01, 0.03, 0.1, 0.3, 1, 3, 10, 30, 100 and 0.

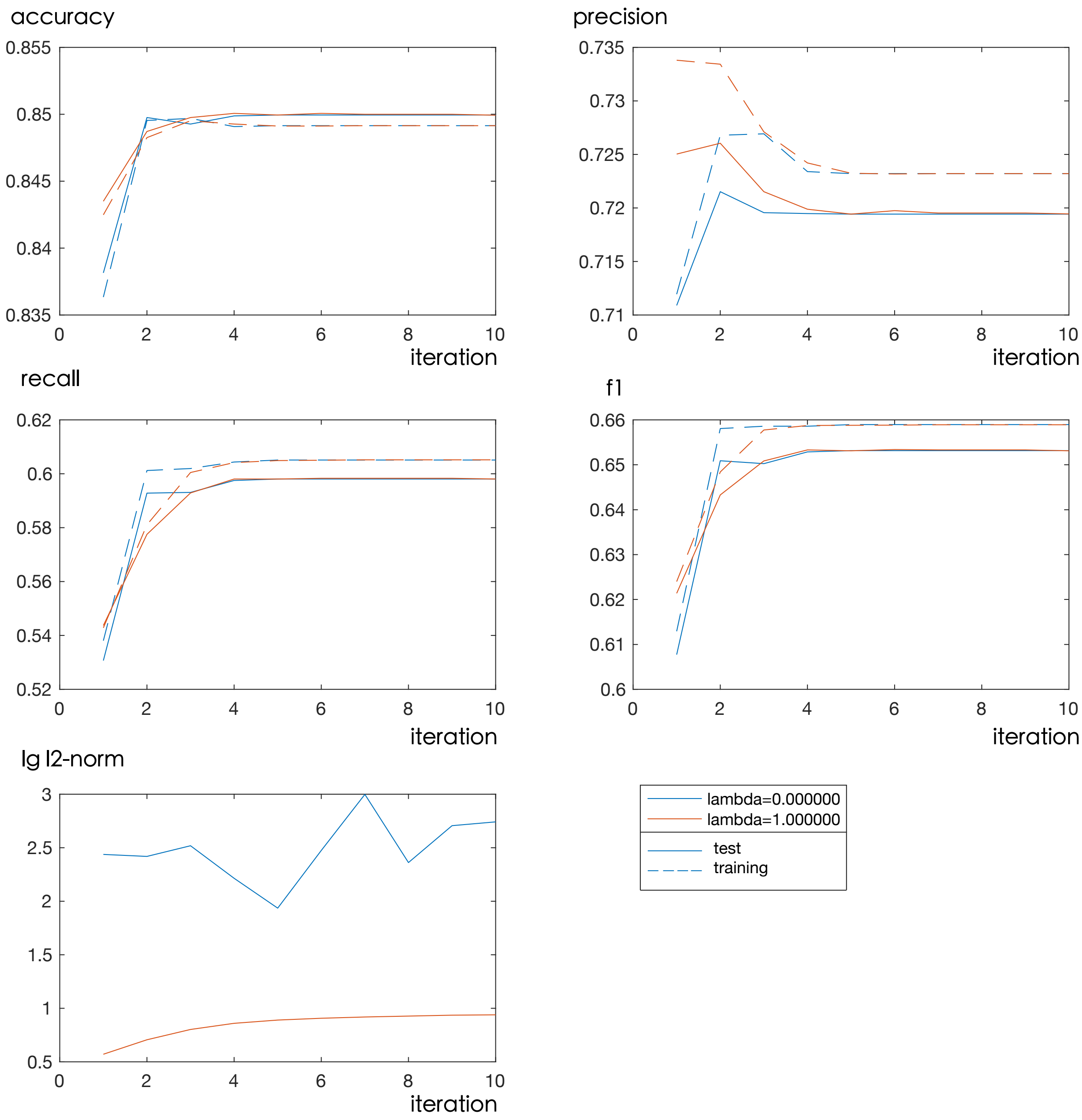


From the figure above, we can see:

- $\lambda = 0.1 \sim 30$ performs well on each evaluation indicator. They all converges in 10 iteration steps (usually 6 steps are enough). The highest is $\lambda = 1$.
- lower regularization and higher regularization performs similar the that mentioned in coarse selection respectively.

Final Test

I adopt $\lambda = 1$ as the best model selected by cross validation to train on the whole training set and test on the testset compare with no regularization. The result is shown below.



The table below is test on the 10th iteration

λ	**data**	**accuracy**	**precision**	**recall**	**f1**	**l2-norm**
0	training	0.849144682	0.723212925	0.605152404	0.658936259	548.1370925
0	testing	0.849947792	0.71942446	0.598023921	0.653130768	-
1	training	0.849947792	0.71942446	0.598023921	0.653130768	8.70818154
1	testing	0.849947792	0.71942446	0.598023921	0.653130768	-

From the figure above and table we can see:

- logistic regression with and without regularization performs similar on accuracy, precision, recall and f1 after enough iterations.
- no obvious overfitting occurs in logistic regression without regularization though the l2-norm is quite large (100~1000). That is because logistic regression is a linear model with very few parameters, which is the dimension of features plus one.
- regularization can effectively suppress the growing of the norm of weights in logistic regression and stabilize the performance as iteration steps increases.

Besides these, we notice that in accuracy -- iteration curve, accuracy on test data is higher than that on training data. That's due to the difference of class size ratio between training data and test data.