Homework 1

Chen Lu

April 2, 2018

1 Mathematics Basics

1.1 Optimization

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

Answer:
Assume $f(x) = x_1^2 + x_2^2 - 1$, $g(x) = x_1 - 2x_2$ and $h(x) = x_1 + x_2 - 1$.

$$\nabla f(x) = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} \qquad \nabla g(x) = \begin{bmatrix} 1 \\ -2 \end{bmatrix} \qquad \nabla h(x) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Because f is a convex function, g is a concave function, h is a linear function, it's a convex programming. We only need to satisfy the KKT condition. We can get the

$$\begin{aligned} \nabla L_x(\mathbf{x}, \mathbf{w}, \mathbf{v}) &= \nabla f(x) - w\nabla g(x) - v\nabla h(x) \\ &= \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} - w \begin{bmatrix} 1 \\ -2 \end{bmatrix} - v \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 0 \\ g(x) &= x_1 - 2x_2 \geq 0 \\ h(x) &= x_1 + x_2 = 0 \\ wg(x) &= w(x_1 - 2x_2) = 0 \\ w &\geq 0 \end{aligned}$$

Solve the equation group above. Get $x_1 = \frac{2}{3}$ and $x_2 = \frac{1}{3}$, that is, the optimal solution. And the optimal object is $f(x) = x_1^2 + x_2^2 - 1 = -\frac{4}{9}$

1.2 Calculus

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

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

(2) Show that

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

Proof:

$$\begin{aligned}
\Gamma(a)\Gamma(b) &= \int_0^\infty \mu^{a-1}e^{-\mu}d\mu \int_0^\infty e^{-v}v^{b-1}dv \\
&= \int_{v=0}^\infty \int_{\mu=0}^\infty e^{-\mu-v}\mu^{a-1}v^{b-1}d\mu dv
\end{aligned}$$

Changing variables by $\mu = f(z,t) = zt$ and $v = g(z,t) = z(1-t)$ shows that this is

$$\begin{aligned}
\Gamma(a)\Gamma(b) &= \int_{z=0}^\infty \int_{t=0}^1 e^{-z}(zt)^{a-1}(z(1-t))^{b-1}|J(z,t)|dtdz \\
&= \int_{z=0}^\infty \int_{t=0}^1 e^{-z}(zt)^{a-1}(z(1-t))^{b-1}zdtdz \\
&= \int_{z=0}^\infty e^{-z}z^{a+b-1}dz \cdot \int_{t=0}^1 t^{a-1}(1-t)^{b-1}dt \\
&= \Gamma(a+b)\int_{t=0}^1 t^{a-1}(1-t)^{b-1}dt
\end{aligned}$$

that is,

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

where $|J(z,t)|$ is the absolute value of the Jacobian determinant of $u = f(z,t)$ and $g = (z,t)$.

1.3 Probability

1.4 Stochastic Process

We toss a fair coin for a number of times and use H(head) and T(tail) to denote the two sides of the coin. Please compute the expected number of tosses we

need to observe a first time occurrence of the following consecutive pattern

$$H, \underbrace{T, T, \cdots, T}_{k}$$

Answer:

Let's first calculate for k consecutive Tails the expected number of tosses needed. Let's denote $E_k(Y)$ for k consecutive Tails. Now if we get one more Tail after $E_{k-1}(Y)$, then we have k consecutive Tails or if it is a Head, then again we have to repeat the procedure. So $E_k(Y) = \frac{1}{2}(E_{k-1}(Y) + 1) + \frac{1}{2}(E_{k-1}(Y) + E_k(Y) + 1)$.

Lets calculate it for n consecutive tosses the expected number of tosses needed.
Let's denote E_n for n consecutive heads. Now if we get one more head after E_{n-1}, then we have n consecutive heads or if it is a tail then again we have to repeat the procedure.
So for the two scenarios:

1. $E_{n_1} + 1$
2. $E_{n+1}(1 for a tail)$

So,

$$E_n = \frac{1}{2}(E_{n-1} + 1) + \frac{1}{2}(E_{n-1} + E_n + 1) \tag{1}$$
$$E_n = 2E_{n-1} + 2 \tag{2}$$
$$\tag{3}$$

We have the general recurrence relation. Define $f(n) = E_n + 2$ with f(0)=2. So

$$f(n) = 2f(n-1) \tag{4}$$
$$f(n) = 2^{n+1} \tag{5}$$

Therefore,$E_n = 2^{n+1} - 2 = 2(2^n - 1)$. that is,

$$E_k(Y) = 2^{k+1} - 2$$

So the expected number of tosses is

$$e = 2^{k+1}$$

2 SVM

original problem:

$$\begin{aligned} \underset{w,b,\xi,\hat{\xi}}{min} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \hat{\xi}_i) \\ s.t. \quad & y_i \le w^T x_i + b + \epsilon + \xi_i, i = 1, ..., N \\ & y_i \ge w^T x_i + b - \epsilon - \xi_i, i = 1, ..., N \\ & \xi_i \ge 0 \quad \forall i = 1, ..., N \\ & \hat{\xi}_i \ge 0 \quad \forall i = 1, ..., N \end{aligned}$$

the langrange function of original problem is

$$\begin{aligned} L(w, b, \xi, \hat{\xi}, \alpha, \beta, \gamma, \lambda) =& \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}(\xi_i + \hat{\xi}_i) \\ & - \sum_{i=1}^{N}\alpha_i(-y_i + w^T x_i + b + \epsilon + \xi_i) - \sum_{i=1}^{N}\beta_i(y_i - w^T x_i - b + \epsilon + \hat{\xi}_i) \\ & - \sum_{i=1}^{N}\gamma_i\xi_i - \sum_{i=1}^{N}\lambda_i\hat{\xi}_i \end{aligned} \tag{6}$$

where $\alpha_i \ge 0$, $\beta_i \ge 0$ The dual problem of original problem is max min problem of Lagrange function.
First, we should derive the minimum of $L(w, b, \xi, \hat{\xi}, \alpha, \beta, \gamma, \lambda)$ as to $w, b, \xi, \hat{\xi}$.

$$\begin{aligned} \nabla_w L(w, b, \xi, \hat{\xi}, \alpha, \beta, \gamma, \lambda) &= w - \sum_{i=1}^{N}\alpha_i x_i + \sum_{i=1}^{N}\beta_i x_i = 0 \\ \nabla_b L(w, b, \xi, \hat{\xi}, \alpha, \beta, \gamma, \lambda) &= -\sum_{i=1}^{N}\alpha_i + \sum_{i=1}^{N}\beta_i = 0 \\ \nabla_{\xi_i} L(w, b, \xi, \hat{\xi}, \alpha, \beta, \gamma, \lambda) &= C - \alpha_i - \gamma_i = 0 \\ \nabla_{\hat{xi}_i} L(w, b, \xi, \hat{\xi}, \alpha, \beta, \gamma, \lambda) &= C - \beta_i - \lambda_i = 0 \end{aligned}$$

that is,

$$w = \sum_{i=1}^{N}(\alpha_i - \beta_i)x_i \tag{7}$$

$$0 = \sum_{i=1}^{N}(\alpha_i - \beta_i) \tag{8}$$

$$\alpha_i = C - \gamma_i \le 0 \tag{9}$$

$$\beta_i = C - \lambda_i \le 0 \tag{10}$$

bring the equations into initial Lagrange function, then we can get

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

and then derive the maximum of $\min_{w,b,\xi,\hat{\xi}} \quad L(w,b,\xi,\hat{\xi},\alpha,\beta,\gamma,\lambda)$ as to α,β. According to the Lagrange dual property, the dual problem of original problem is max min problem:

$$\max_{\alpha,\beta} \min_{w,b,\xi,\hat{\xi}} L(w,b,\xi,\hat{\xi},\alpha,\beta,\gamma,\lambda)$$

Taking the equations into equation (6), we can get

$$\min_{w,b,\xi,\hat{\xi}} \quad L(w,b,\xi,\hat{\xi}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{i=1}^{N}(\alpha_i-\beta_i)(\alpha_j-\beta_j)x_i x_j - \sum_{i=1}^{N} b(\alpha_i-\beta_i) - \sum_{i=1}^{N}\epsilon(\alpha_i+\beta_i)$$

(2) Get the maximum of $\min_{w,b,\xi,\hat{\xi}} \quad L$ as to α, that is

$$\max_{\alpha,\beta} \quad -\frac{1}{2}\sum_{i=1}^{N}\sum_{i=1}^{N}(\alpha_i-\beta_i)(\alpha_j-\beta_j)x_i x_j + \sum_{i=1}^{N} y_i(\alpha_i-\beta_i) - \sum_{i=1}^{N}\epsilon(\alpha_i+\beta_i)$$

$$\begin{aligned} s.t. \quad & \sum_{i=1}^{N}(\alpha_i-\beta_i) = 0 \\ & 0 \le \alpha_i \le C \qquad \forall i = 1,...,N \\ & 0 \le \beta_i \le C \qquad \forall i = 1,...,N \end{aligned}$$

Finally, get the dual problem:

$$\min_{\alpha,\beta} \quad \frac{1}{2}\sum_{i=1}^{N}\sum_{i=1}^{N}(\alpha_i-\beta_i)(\alpha_j-\beta_j)x_i x_j - \sum_{i=1}^{N} y_i(\alpha_i-\beta_i) + \sum_{i=1}^{N}\epsilon(\alpha_i+\beta_i)$$

$$\begin{aligned} s.t. \quad & \sum_{i=1}^{N}(\alpha_i-\beta_i) = 0 \\ & 0 \le \alpha_i \le C \qquad \forall i = 1,...,N \\ & 0 \le \beta_i \le C \qquad \forall i = 1,...,N \end{aligned}$$

3 IRLS for Logistic Regression

3.1 Logistic regression

$$\max_{w} \quad L(w)$$

Using Newton's method to optimize the parameter w:

$$\nabla_w L(w)|_w = X(y-\mu)$$

where $\mu = sigmoid(Xw)$ and the Hessian matrix of L(w) is

$$H = \nabla_w^2 L(w) = -XRX^T$$

where R is a diagonal matrix and $Rii = \mu_i(1 - \mu_i)$
The update equation is is

$$w_{t+1} \leftarrow w_t - H^{-1}\nabla_w L(w) \tag{11}$$
$$\leftarrow w_t - (XRX^T)^{-1}X(\mu - y) \tag{12}$$

3.2 L2-norm regularized logistic regression

$$\max_w \quad -\frac{\lambda}{2}\|\omega\|_2^2 + L(\omega)$$

where λ is the positive regularization constant.
The gradient of L is

$$\nabla_w L = X(y - \mu) - \lambda w$$
$$H(w) = -XRX^T - \lambda I$$

The update equation is

$$w_{t+1} \leftarrow w_t - H^{-1}\nabla_w L(w) \tag{13}$$
$$\leftarrow w_t + (XRX^T + \lambda I)^{-1}[X(y - \mu) - \lambda w_t] \tag{14}$$

3.3 Experiment

3.3.1 Initial

1. $w_0 = 10^{-13} * I$
2. max iter = 50
3. loss threshold = 0.001
4. train = 27676
5. dev = 4885
6. test = 16281

3.3.2 Process

First, given a fixed λ, get the weight of last iteration.
Second, in dev set, get accuracy score from the weight in different λ. Choose the weight of the best score as the weight of test set.
Last, use the optimal weight we get from dev set to evaluate test set and get the result of prediction.

3.4 Result

3.4.1 Training set

I set a group of *lambda* for training set, and get the last iteration's of weight, stop iteration and accuracy. Show the result in Table 1.

Table 1: Accuracy Table

λ	Stop Iteration	Accuracy
0	14	0.848569
0.01	10	0.848569
0.1	8	0.848569
1	8	0.848641
10	8	0.847594
100	7	0.844775

Table 2: Prediction Accuracy

set	Accuracy
train	0.848569
dev	0.847697
test	0.849764

3.4.2 Accuracy

I draw a picture which shows the accuracy in every iteration in training set and compare the result of different λ. From Figure 1, it's obvious that the speed of converge is almost the same. After 2 iterations, it's nearly achieve the optimal accuracy. And in different λ, the accuracy is very close.

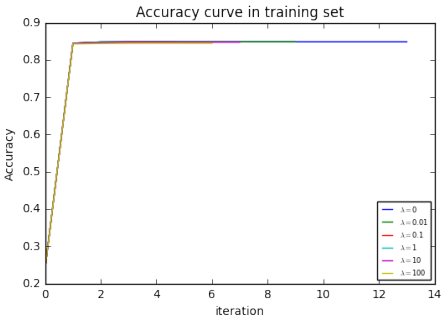

Figure 1: Accuracy

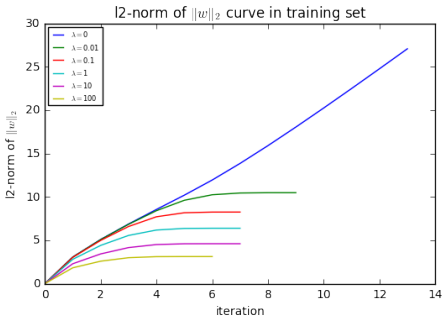

Figure 2: l2-norm $\|w\|_2$

3.4.3 L2-norm of $\|w\|_2$

In every training iteration, I recorded the L2-norm of $\|w\|_2$.
From Figure 2, easily we can draw the conclusion that if IRLS algorithm without regularization, l2-norm of $\|w\|_2$ is approximately a linear line. With l2-norm regularization, $\|w\|_2$ intend to be a constant after certain iteration.
It's obvious that, with l2-norm regularization, larger λ, smaller $\|w\|_2$.

3.4.4 Loss curve in training set

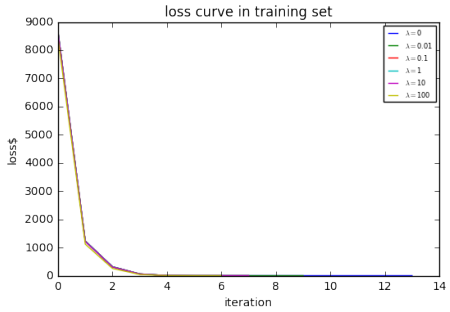

Figure 3: Loss in every iteration

From Figure 3, the speed of loss decent is almost the same in different λ.

4 Reference

Thanks for the help of Li Siyuan, Zheng shun.
Here is some references from the websites and book.

1. Wikipedia Beta function
2. expected-number-of-coin-tosses-to-get-five-consecutive-heads
3. Optimization Method