Homework 3

Yunian Pan

April 11, 2019

1 Optimization

I did normalization on the loss function to avoid numerical problem:

$$J = -\frac{1}{N}\sum_{i=1}^{N}\log(\sigma(y_i w_i^\top k_i)) + \lambda w^\top w$$

The gradient then becomes:

$$\nabla_w J = -\frac{1}{N} column[(1-\sigma(y_i w_i^\top k_i))y_i k_i^\top] + 2\lambda w$$

1.1 GD

After some exploration, I set the parameters are as below and get the results:

Table 1: GD

λ	1e-3
step size	2e-3
ϵ	1e-5
step length	3000
accuracy	90.2%

The learning curve is in 1.1

1.2 SGD

For SGD, first I tried the window size $p = 1$

The learning curve is in 1.2

For the setting of $p = 100$, as 3 shows,

The learning curve is in 1.3

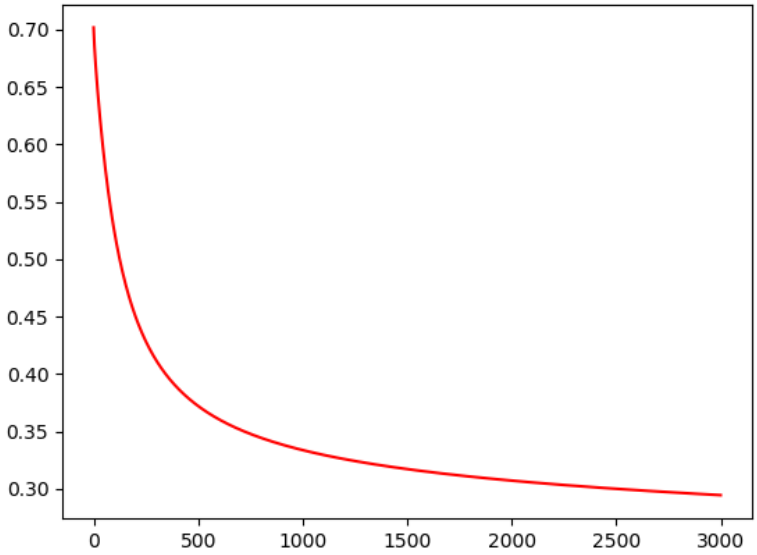

Figure 1.1: GD

Table 2: SGD1

p	1
λ	1e-3
step size	3e-4
ϵ	1e-5
step length	3000
accuracy	88.9%

Table 3: SGD2

p	100
λ	1e-3
step size	3e-4
ϵ	1e-5
step length	3000
accuracy	91.2%

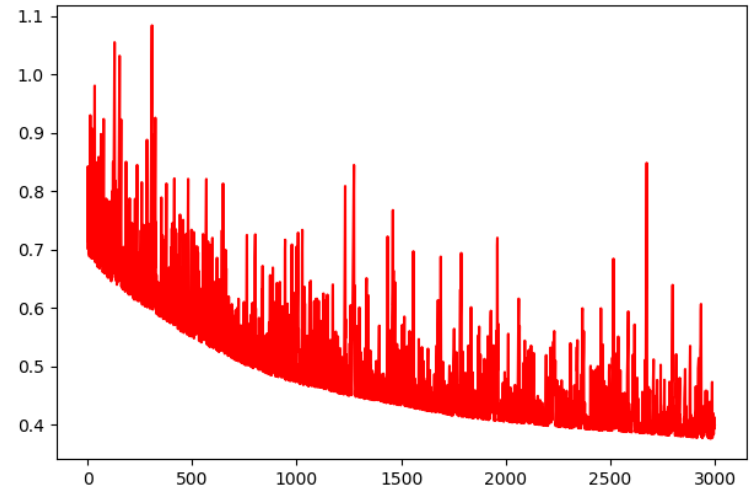

Figure 1.2: p = 1

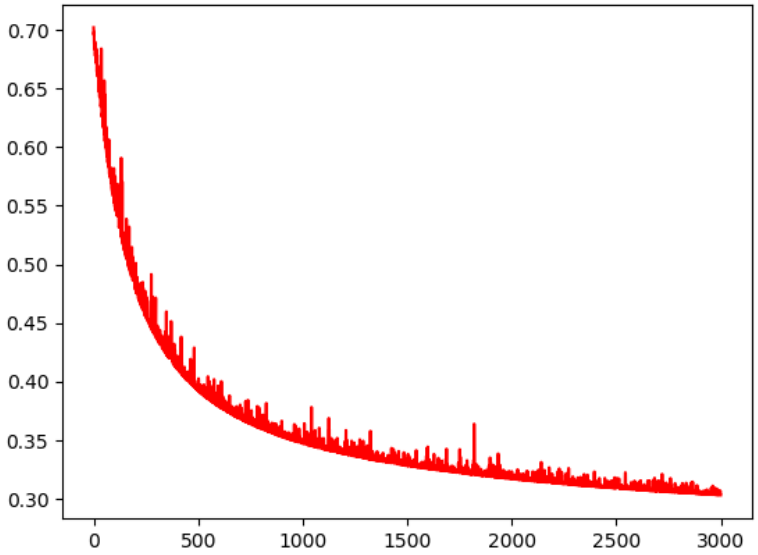

Figure 1.3: p = 100

1.3 BFGS

Table 4: BFGS

λ	1e-3
ϵ	1e-5
step length	100
accuracy	92.9%

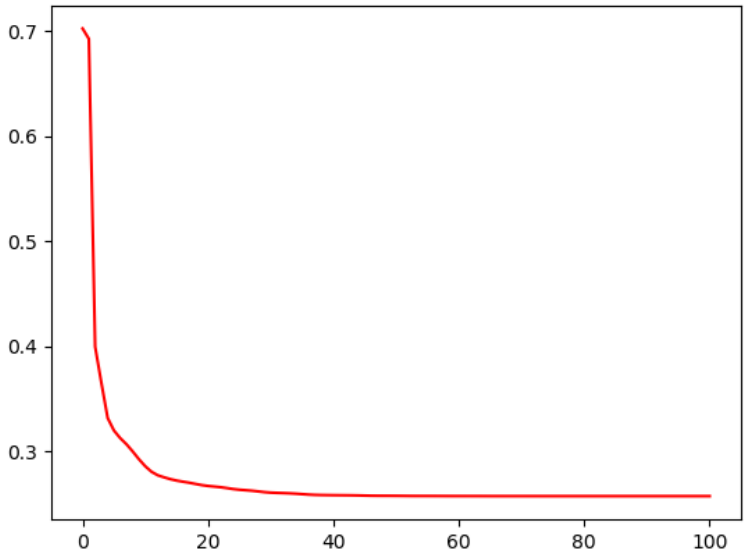

Figure 1.4

1.4 LBFGS

Remark that the convergence is not stable

Table 5: LBFGS

λ	1e-3
ϵ	1e-5
step length	50
accuracy	90.7%

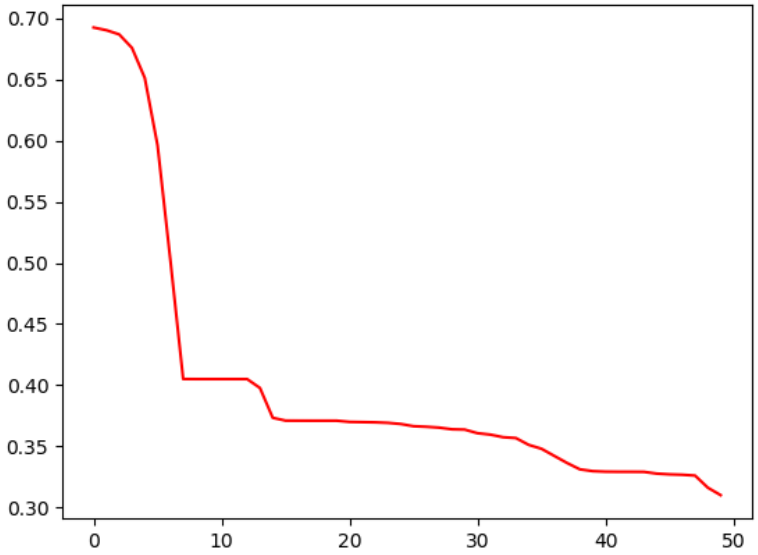

Figure 1.5

2 EM

2.1 E step:

$$\begin{aligned} Q_i(z_i) &= p(z_i|x_i;\theta) \\ &= \frac{p(z_i,x_i|\theta)}{\sum_{z_i} p(z_i,x_i|\theta)} \\ &= \frac{\pi_{z_i}\prod_{j=1}^{M}\mu_{z_i}(j)^{x_i(j)}}{\sum_{z_i}^{k}\pi_{z_i}\prod_{j=1}^{M}\mu_{z_i}(j)^{x_i(j)}} \end{aligned}$$

2.2 M step:

$$\begin{aligned} \mathcal{L}(\theta) &= \sum_{i=1}^{N}\sum_{z_i} Q_i(z_i)\log\frac{p(x_i,z_i;\theta)}{Q_i(z_i)} \\ &= \sum_{i=1}^{N}\sum_{z_i} Q_i(z_i)\log p(x_i,z_i;\theta) - \sum_{i=1}^{N}\sum_{z_i} Q_i(z_i)\log Q_i(z_i) \\ &= Q(\theta) - const \end{aligned}$$

$$\begin{aligned} \mathcal{Q}(\theta) &= \sum_{i=1}^{N}\sum_{z_i} Q_i(z_i)\log p(x_i,z_i;\theta) \\ &= \sum_{i=1}^{N}\sum_{z_i} \tau_{i,z_i}(\log\pi_{z_i} + \log\prod_{j=1}^{M}\mu_{z_i}(j)^{x_i(j)}) \\ &= \sum_{i=1}^{N}\sum_{z_i} \tau_{i,z_i}(\log\pi_{z_i} + x_i(q)\log\mu_{z_i}(q)) \qquad (x_i(q)=1) \end{aligned}$$

Where we write $Q_i(z_i)$ as τ_{i,z_i}.

There are 2 constraints: $\sum_j^M \mu_k(j) = 1$, $\sum_i^K \pi_i = 1$, with an observation $\sum_j^M x_i(j) = 1\ \forall x_i \in \{x_1,\ldots,x_N\}$.

Use lagrange multiplier to find the optimum value of π_{z_i} and μ_{z_i} as below:

$$L(\mu, \pi, \lambda_1, \lambda_2) = \mathcal{Q}(\pi, \mu) - \lambda_1(\sum_{j}^{M} \mu_k(j) - 1) - \lambda_2(\sum_{i}^{K} \pi_i - 1)$$

$$\frac{\partial L}{\partial \pi_k} = \sum_{i=1}^{N} \tau_{i,k} \frac{1}{\pi_k} - \lambda_2 = 0$$

$$sum \;\Rightarrow\; 1 = \sum_{k=1}^{K} \pi_k = \frac{1}{\lambda_2} \sum_{k=1}^{K} \sum_{i=1}^{N} \tau_{i,k}$$

$$\Rightarrow \lambda_2 = \sum_{k=1}^{K} \sum_{i=1}^{N} \tau_{i,k}$$

$$plug\ in \Rightarrow \pi_k = \frac{\sum_{i=1}^{N} \tau_{i,k}}{\sum_{k=1}^{K} \sum_{i=1}^{N} \tau_{i,k}}$$

$$\frac{\partial L}{\partial \mu_k(q)} = \sum_{i=1}^{N} \tau_{i,k} x_i(q) \frac{1}{\mu_k(q)} - \lambda_1 = 0$$

$$\Rightarrow \mu_k(q) = \sum_{i=1}^{N} \tau_{i,k} x_i(q) \frac{1}{\lambda_1}$$

$$sum \;\Rightarrow\; 1 = \sum_{q=1}^{M} \mu_k(q) = \sum_{i=1}^{N} \tau_{i,k} \sum_{q=1}^{M} x_i(q) \frac{1}{\lambda_1}$$

$$\Rightarrow \lambda_1 = \sum_{i=1}^{N} \tau_{i,k}$$

$$plug\ in \Rightarrow \mu_k(q) = \sum_{i=1}^{N} \frac{\tau_{i,k} x_i(q)}{\sum_{i=1}^{N} \tau_{i,k}}$$

To conclude, the probability for class $z = k$ is $\pi_k = \frac{\sum_{i=1}^{N} \tau_{i,k}}{\sum_{k=1}^{K} \sum_{i=1}^{N} \tau_{i,k}}$, the probability of x_i which belongs to the class $z = k$ taking on the q^{th} value is $\mu_k(q) = \frac{\sum_{i=1}^{N} \tau_{i,k} x_i(q)}{\sum_{i=1}^{N} \tau_{i,k}}$.

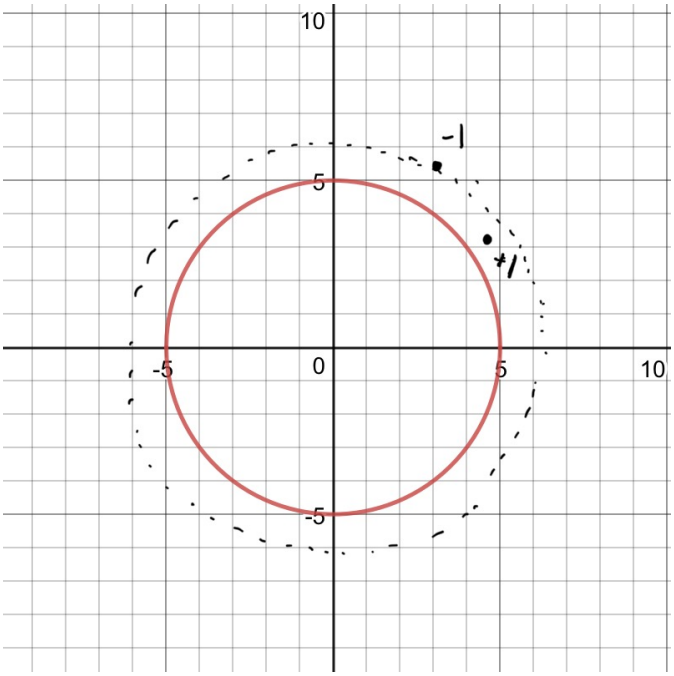

Figure 3.1

3 VC-dimension

3.1 a

The VC-dimension $h = 1$,

Justification: the case for 1 point is trivial, I can adjust the radius to label it as positive or negative no matter what the position is; For case of 2 points, once the distance from the 2 points to the original point is different, then there's no way to classify them if the outer one is negative and the inner one is positive as 3.1 shows.

3.2 b

The VC-dimension $h = 2$,

Justification: Since I can reverse the classier by changing the sign of a, b, the case for 2 points is trivial as the inner one can be labeled as positive no matter how the another is labeled; For the case of 3 points, as 3.2 illustrates, no matter where the third point is, its label will put some restriction on the labeling of other points, for instance if the third point is in region (c) and be labeled negative, then either the intermedian point will be "sandwiched" and forced to be negative or the

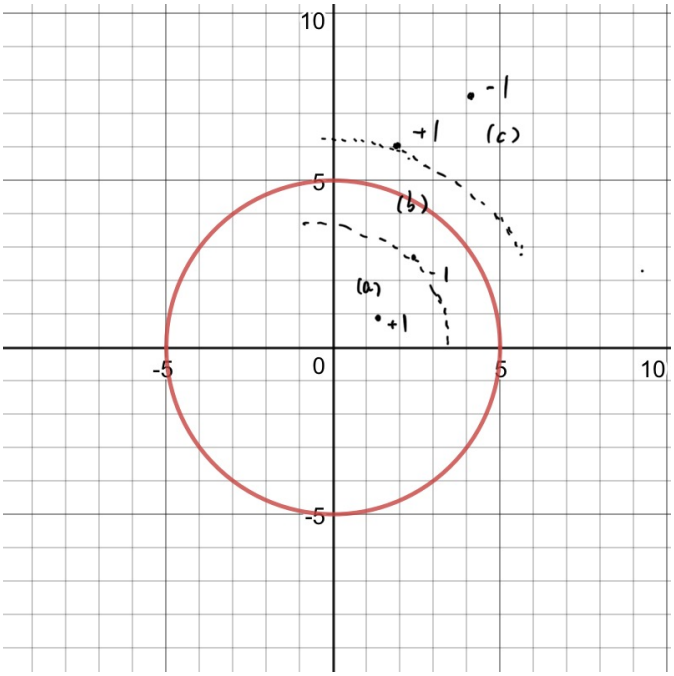

Figure 3.2

inner one will be forced to be positive, and the situation is the same with region (a) and (b).