Machine Learning HW5

B09902005 資工三盧冠綸

December 10, 2022

For this homework, I discussed with B09502019 and B09901133.

programming part

problem 1: 請說明你是如何 normalize discrete 跟 continous 的 feature (1%)

I only normalized continuous data. Below is the pseudo code of how I do this:

```
def normalize(X):
    for label in (continuous_labels):
        X[label] = pd.DataFrame((X[label]-X[label].mean())/X[label].std())
    return X
```

I've tried to normalize discrete features, and I've also tried to give different weights to different features, or use log on those continuous features, but the performance doesn't seem to be better.

problem 2: 使用 DNN 做 feature transformation，將 output dimension 設爲 4 跟 1024，並丟進 linear SVM 訓練，比較 leaderboard 上的結果，並說明造成這樣結果的原因 (hint: linear SVM 本身是 linear classifier，資料必須是 linearly separable 的資料) (1%)

Below is the table of scores on leaderboard of the model trained with different output dimensions: (I set $lr = 0.01$, $C = 1$, and $epoch = 15$ for all of the output dimensions)

output dimension	public score	private score
4	0.85540	0.85235
32	0.85184	0.84731
1024	0.84508	0.84375
8192	0.83488	0.82704

From the table, we can easily notice that the score gets lower when the output dimension of DNN gets larger.

The reason is that when the dimension gets higher, it will become easier to set the margin to separate the two different types of data. (For the most extreme case, if the dimension is larger than the number of training data, and all data are linearly independent, then the data is linearly separable, thus training accuracy will be 100%.) However, this is not a good thing for predicting since there are too many possible ways that lead to the highest training accuracy or the lowest training loss. However, these minima of loss or maxima of accuracy do not imply a good testing loss or accuracy because there are too many possibilities that is beyond what we can see from the value of the training loss or accuracy. So, a larger dimension for linear SVM may not lead to better performance of the models.

mathematics part

problem 3: Support Vector Regression (2%)

1. Here, we use $f(w, b, \xi)$ to denote the function we want to minimize. That is,

$$f(w, b, \xi) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i$$

Then, we use $g_{1,i}(x)$ and $g_{2,i}(x)$ and $g_{3,i}(x)$ to deal with (2) and (3) and (4).

$$g_{1,i}(w, b, \xi) = y_i - w^T x_i - b - \epsilon - \xi_i$$

$$g_{2,i}(w, b, \xi) = w^T x_i + b - y_i - \epsilon - \xi_i$$

$$g_{3,i}(w, b, \xi) = -\xi_i$$

This way, we have $g_{1,i}(x) \leq 0$, $g_{2,i}(x) \leq 0$, and $g_{3,i}(x) \leq 0$ for all $1 \leq i \leq m$.

This way, the Lagrangian for the optimization problem is

$$L(w, b, \xi, \alpha, \alpha^\star, \beta) = f(w, b, \xi) + \sum_{i=1}^{m}(\alpha_i g_{1,i}(w, b, \xi) + \alpha_i^\star g_{2,i}(w, b, \xi) + \beta_i g_{3,i}(w, b, \xi))$$

2. The dual optimization problem is

$$\text{maximize } \theta(\alpha, \alpha^\star, \beta) = \inf_{w \in R^{n+1}, b \in R, \xi \in R^m} \{L(w, b, \xi, \alpha, \alpha^\star, \beta)\}$$

$$\text{subject to } \alpha_i \geq 0, \alpha_i^\star \geq 0, \beta_i \geq 0, \forall 1 \leq i \leq m, \text{ where } \alpha_i, \alpha_i^\star, \beta_i \in R$$

And in fact,

$$\nabla_w L = w + \sum_{i=1}^{m}(-\alpha_i + \alpha_i^\star)x_i$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{m}(-\alpha_i + \alpha_i^\star)$$

$$\frac{\partial L}{\partial \xi_i} = C - (\alpha_i + \alpha_i^\star + \beta_i)$$

Here, we have to let $\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \xi_i} = 0$ for all i, otherwise $\theta(\alpha, \alpha^\star, \beta)$ will be $-\infty$ because the value of L may go to $-\infty$ if the values of b and ξ_i are adjusted to ∞ or $-\infty$.

Then, since L is convex with respect to w, thus
$\theta(\alpha, \alpha^\star, \beta) = L(w, b, \xi, \alpha, \alpha^\star, \beta)$ if and only if $\nabla_w L = 0$. Which is,
$w = \sum_{i=1}^{m}(\alpha_i - \alpha_i^\star)x_i$.
This way, given $\frac{\partial L}{\partial b} = \frac{\partial L}{\partial \xi_i} = 0$, then

$$\theta(\alpha, \alpha^\star, \beta) = f(w, b, \xi) + \sum_{i=1}^{m}(\alpha_i g_{1,i}(w, b, \xi) + \alpha_i^\star g_{2,i}(w, b, \xi) + \beta_i g_{3,i}(w, b, \xi))$$

$$= \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}((\alpha_i - \alpha_i^\star)(y_i - w^T x_i - b) + (\alpha_i + \alpha_i^\star)\epsilon - (\alpha_i + \alpha_i^\star + \beta_i)\xi_i)$$

$$= \frac{1}{2}||w||^2 + \sum_{i=1}^{m}((\alpha_i - \alpha_i^\star)(y_i - w^T x_i - b) + (\alpha_i + \alpha_i^\star)\epsilon)$$

$$= \frac{1}{2}||w||^2 + \sum_{i=1}^{m}(\alpha_i - \alpha_i^\star)y_i - \sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i - \alpha_i^\star)((\alpha_j - \alpha_j^\star)x_j)^T x_i + \sum_{i=1}^{m}(\alpha_i + \alpha_i^\star)\epsilon$$

Since

$$\begin{aligned}\frac{1}{2}||w||^2 &= \frac{1}{2}(\sum_{i=1}^{m}(\alpha_i-\alpha_i^\star)x_i)^T(\sum_{i=1}^{m}(\alpha_i-\alpha_i^\star)x_i)\\ &= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}((\alpha_i-\alpha_i^\star)x_i)^T(\alpha_j-\alpha_j^\star)x_j\\ &= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i-\alpha_i^\star)((\alpha_j-\alpha_j^\star)x_j)^Tx_i\end{aligned}$$

So,

$$\theta(\alpha,\alpha^\star,\beta)=\sum_{i=1}^{m}(\alpha_i-\alpha_i^\star)y_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i-\alpha_i^\star)((\alpha_j-\alpha_j^\star)x_j)^Tx_i+\sum_{i=1}^{m}(\alpha_i+\alpha_i^\star)\epsilon$$

So, the dual optimization problem is

$$\text{maximize } \theta(\alpha,\alpha^\star,\beta)=\sum_{i=1}^{m}(\alpha_i-\alpha_i^\star)y_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i-\alpha_i^\star)(\alpha_j-\alpha_j^\star)x_j^Tx_i+\sum_{i=1}^{m}(\alpha_i+\alpha_i^\star)\epsilon$$

$$\text{subject to } \alpha_i\geq 0,\alpha_i^\star\geq 0,\alpha_i+\alpha_i^\star\leq C,\forall i,\text{ and }\sum_{i=1}^{m}(-\alpha_i+\alpha_i^\star)=0$$

3. (a) Note that the original convex optimization problem can be modified as minimizing $\frac{1}{2}||w||^2+C\sum_{i=1}^{m}\max(y_i-w^Tx_i-b-\epsilon,w^Tx_i+b-y_i-\epsilon,0)$. That is:

$$\text{maximize } \frac{1}{2}||w||^2+C\sum_{i=1}^{m}\max(|y_i-w^Tx_i-b|-\epsilon,0)$$

$$\text{variables } w\in R^{n+1},b\in R,\xi\in R^m$$

From the modification, we can know that given optimal weighting vector $\bar{w}$, then the optimal bias $\bar{b}$ is

$$\bar{b}=\arg\min_{b\in\mathbb{R}}C\sum_{i=1}^{m}\max(|y_i-(\bar{w}^Tx_i+b)|-\epsilon,0)$$

(b) Now, we have the following three stationary conditions:
(S1) $\nabla_w L = 0$, (S2) $\frac{\partial L}{\partial b} = 0$, and (S3) $\frac{\partial L}{\partial \xi_i} = 0 \forall 1 \leq i \leq m$

And, the primal and dual feasibility conditions are
(P1) $g_{1,i}(w, b, \xi) \leq 0$, (P2) $g_{2,i}(w, b, \xi) \leq 0$, (P3) $g_{3,i}(w, b, \xi) \leq 0$,
(D1) $\alpha_i \geq 0$, (D2) $\alpha_i^\star \geq 0$, and (D3) $\beta_i \geq 0$ for all $1 \leq i \leq m$.

The complementary slackness condition is
(C1) $\alpha_i g_{1,i}(w, b, \xi) = 0$, (C2) $\alpha_i^\star g_{2,i}(w, b, \xi) = 0$, and (C3)
$\beta_i g_{3,i}(w, b, \xi) = 0$.
Consider $e = y_i - (\bar{w}^T x_i + \bar{b})$.

If $|e| < \epsilon$, then $g_{1,i}(w, b, \xi) < 0$, and $g_{2,i}(w, b, \xi) < 0$. From (C1) and (C2), we can know $\bar{\alpha_i} = \bar{\alpha_i^\star} = 0$. From (S3), we know $\bar{\beta_i} = C > 0$, so (C3) tells us that $g_{3,i}(w, b, \xi) = -\xi_i = 0$, thus $\bar{\xi_i} = 0$.

If $e = \epsilon$, then $g_{2,i}(w, b, \xi) < 0$. Thus from (C2), we can know $\bar{\alpha_i^\star} = 0$. Then, $\bar{\xi_i} = 0$ (Otherwise, from (C1) and (C3), $\bar{\alpha_i} = 0$, and $\bar{\beta_i} = 0$, but (S3) will be violated). Thus, from (S3), (D1), and (D3), we can conclude that $0 \leq \bar{\alpha_i} \leq C$.

If $e = -\epsilon$, then $g_{1,i}(w, b, \xi) < 0$. Thus from (C1), we can know $\bar{\alpha_i} = 0$. Then, $\bar{\xi_i} = 0$. Thus, from (S3), (D2), and (D3), we can conclude that $0 \leq \bar{\alpha_i^\star} \leq C$.

If $e > \epsilon$, then from (P1), we can know $\bar{\xi_i} > 0$. From (C2) and (C3), we can know $\bar{\alpha_i^\star} = 0$ and $\bar{\beta_i} = 0$. From (S3), we can see $\bar{\alpha_i} = C$. From (C1), we can conclude that $\bar{\xi_i} = e - \epsilon$.

If $e < -\epsilon$, then from (P2), we can know $\bar{\xi_i} > 0$. From (C1) and (C3), we can know $\bar{\alpha_i} = 0$ and $\bar{\beta_i} = 0$. From (S3), we can see $\bar{\alpha_i^\star} = C$. From (C2), we can conclude that $\bar{\xi_i} = -(e + \epsilon)$.