1.

(1) F. The $l_2$ regularized logistic regression has only 1 global optimal solution

(2) F with penalty term $\gamma\|w\|_2^2$, the values of the weight should be distributed evenly

$\geq$ / since there exist $w'$ that can totally separate the data and achieve highest likely hood, multiplying a scalar $\alpha$ of arbitrary large will not reduce the likelyhood. Hence $w=\alpha w'$ $\alpha\to\infty$ $w\to\infty$

(4) F As $\lambda$ increase, $l(w, D_{train})$ decrease, hence $l(w, D_{train})$ increase, meaning that the penalty cause the a worse performance in training set, hence a lower likelyhood

(5) F If $\lambda$ is too large, a underfit tend to happen, hence L is lower

---

2. (1)

Let the problem with $\xi_i \geq 0$ constraint be problem A, without the constraint be problem B. The goal is to prove A and B are equivalent

B $\to$ A: suppose $w^*$ is the optimal solution of problem A, so it satisfy:

A $\to$ B

Let the problem with $\xi_i \geq 0$ constraint be problem A, without the constraint is problem B, it is to prove that A and B are equivalent:

$A \Leftrightarrow B$

① A $\to$ B

suppose there exist some case when the optimal solution of B is not that of A. the optimal value of problem B is less than that of problem A:

$Opt(B) < Opt(A)$

$\frac{1}{2}\|w_B\|^2 + \frac{C}{2}\sum \xi_{Bi} < \frac{1}{2}\|w_B\|^2 + \frac{C}{2}\sum \xi_{Bi}$;

for this to happen, there must exist $\xi_i < 0$ for some $j$

$y^i(w_B^T x + b) \geq 1 - \xi_{Bi}$, $i = 1, \dots, m$

suppose let set $S$ be the subset of $\{1, \dots, m\}$ such that $\forall \xi_{Bi} < 0$ $i \in S$

then $y^i(w_B^T x + b) \geq 1 - \xi_{Bi} \geq 1 - \xi_{Bi}'$

where $\forall \xi_{Bj}' = 0$

then

$Opt(C)$

$$Opt(B) = \frac{1}{2}\|w_B\|^2 + \frac{C}{2}\sum \xi_{Bi} = \frac{1}{2}\|w_B\|^2 + \frac{C}{2}\sum_{i\notin S} \xi_{Bi}^2 + \frac{C}{2}\sum_{j\in S} \xi_{Bj}^2$$

$$\geq \frac{1}{2}\|w_B\|^2 + \frac{C}{2}\sum_{i\notin S} \xi_{Bi}^2 + \frac{C}{2}\sum_{j\in S} \xi_{Bj}'^2 = \frac{1}{2}\|w_B\|^2 + \frac{C}{2}\sum_{i\notin S} \xi_{Bi}^2$$

hence this $w_B$ is not the optimal solution which is a contradiction.

hence $Opt(B) \geq Opt(A)$

② B $\to$ A

since B has a less strict constraint, all value objective value in A can be achived in B

hence $Opt(B) \leq Opt(A)$

①② $\Rightarrow Opt(B) = Opt(A)$ Q.E.D.

(2) $\min \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum_{i=1}^{m} \xi_i^2$

s.t. $-y^i(w^T x^i + b) + 1 - \xi_i \le 0$, $i = 1, \dots, m$

Lagrangian:

$L(w, b, \xi, \alpha) = \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum \xi_i^2 + \sum_{i=1}^{m} \alpha_i(-y^i(w^T x^i + b) + 1 - \xi_i)$

(3) $\frac{\partial L}{\partial w_0} = w_0 - \sum_{i=1}^{m} \alpha_i y_i x = 0$

$w = \alpha_i y_i x$ $\quad w = \sum_{i=1}^{m} \alpha_i y_i x_i$

$\frac{\partial L}{\partial b} = \sum_{i=1}^{m} -\alpha_i y_i = 0 \Rightarrow \sum_{i=1}^{m} \alpha_i y_i = 0$

$\frac{\partial L}{\partial \xi_i} = C\xi_i - \alpha_i$

Hence $\frac{\partial L}{\partial \xi} = C\xi - \alpha$

$\frac{\partial L}{\partial \xi} = 0 \Rightarrow \xi = \frac{\alpha}{C}$

(4) $g(u, v) = \min_{s.t.\ u\alpha \ge 0} L(w, b, \xi, \alpha)$

$w = \sum_{i=1}^{m} \alpha_i y_i x_i$ $\quad \sum_{i=1}^{m} \alpha_i y_i = 0$ $\quad \xi_i = \frac{\alpha_i}{C}$

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

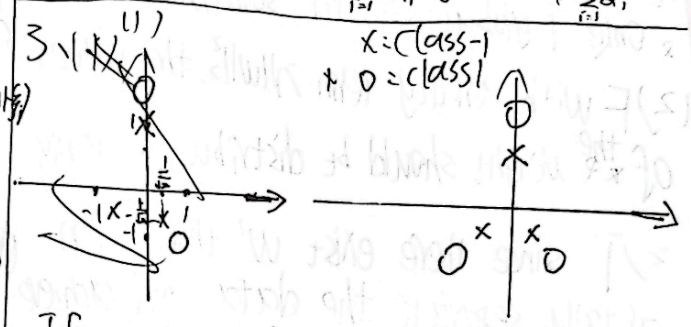
$= \frac{1}{2}\|w\|^2 + \frac{C}{2}\sum \xi_i^2 - \sum_{i=1}^{m} \alpha_i y_i x_i^T w - b\sum_{i=1}^{m} \alpha_i y_i + \sum_{i=1}^{m} \alpha_i + \sum_{i=1}^{m} \xi_i \alpha_i - \sum_{i=1}^{m} \xi_i \alpha_i$

$= \frac{1}{2}\|w\|^2 + \frac{C}{2}\left(\frac{\alpha_i}{C}\right)^2 - w \sum_{i=1}^{m} \alpha_i y_i x_i^T \left(\sum_{i=1}^{m} \alpha_i y_i x_i\right) + \sum_{i=1}^{m} \alpha_i \frac{\alpha_i^2}{C}$

$= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j (y_i x_i^T)(\alpha_j y_j x_j) - \sum_{i=1}^{m}\sum_{j=1}^{m} (\alpha_i y_i x_i^T)(\alpha_j y_j x_j) + \frac{1}{2C}\sum \alpha_i^2 - \frac{1}{C}\sum \alpha_i^2 + m\sum_{i=1}^{m} \alpha_i$

$= -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j x_i^T x_j - \frac{1}{2C}\sum \alpha_i^2 + m\sum_{i=1}^{m} \alpha_i$

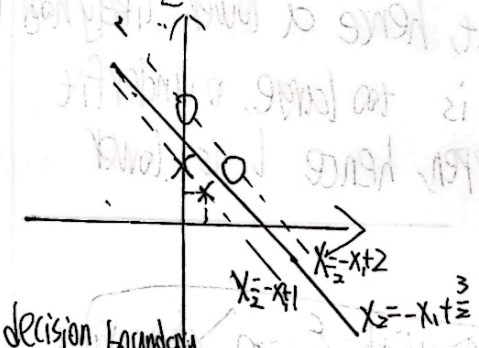
s.t. $\alpha \ge 0$, $\sum_{i=1}^{m} \alpha_i y_i = 0$




If no kernel is applied, SVM classifier cannot split this dataset since it is linearly non-separable

(2)

class -1: $\begin{bmatrix} (0 & 1) \\ (-\frac{1}{2} & \frac{1}{2}) \\ (\frac{1}{2} & \frac{1}{2}) \end{bmatrix}$ class 1: $\begin{bmatrix} (0 & 2) \\ (1 & 1) \\ (1 & 1) \end{bmatrix}$




decision boundary

prediction: $w^T x + b = 0$ $\quad x_2 = -x_1 + \frac{3}{2} \Rightarrow x_1 + x_2 - \frac{3}{2} = 0$

$w^T x + b = 0$ $\quad w_1 x_1 + w_2 x_2 + b = 0 \Rightarrow w_1 = 1, w_2 = 1, b = -\frac{3}{2}$

so $y = \text{sgn}\left(\begin{bmatrix} 1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \frac{3}{2}\right)$

$\text{sgn}\begin{bmatrix} -1 \\ -1 \end{bmatrix}$ $\quad \text{sgn}\left(\begin{bmatrix} 1 & 1 \end{bmatrix}\begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \sqrt{2} \end{bmatrix} - \frac{3}{2}\right)$

$= \text{sgn}\left(\frac{1}{\sqrt{2}} - \frac{3}{2}\right) = -1$

so table is wr

$x = \begin{bmatrix} -\frac{1}{\sqrt{2}} & \sqrt{2} \end{bmatrix} \Rightarrow \phi(x) = x' = \phi(x) = \begin{bmatrix} -\frac{1}{2} & 2 \end{bmatrix}$

$w^T x' + b = \frac{1}{2} + 2 - \frac{3}{2} = 1 \Rightarrow \text{sgn}(w^T x' + b) = 1$

so class 1

4

1) $H(pass) = -\frac{4}{6}\log\frac{4}{6} - \frac{2}{6}\log\frac{2}{6} = 0.92$

(2) ~~$H(Pass|GPA)$~~

$P(T|L) = \frac{1}{2}$ $P(T,L) = \frac{1}{6}$ | $H(pass|GPA)$

$P(F|L) = \frac{1}{2}$ $P(F,L) = \frac{1}{6}$ | ~~$-4\times\frac{1}{6}\log$~~

$P(T|M) = \frac{1}{2}$ $P(T,M) = \frac{1}{6}$ | $= -4\times\frac{1}{6}\log\frac{1}{2} - \frac{1}{3}\log$

$P(F|M) = \frac{1}{2}$ $P(F,M) = \frac{1}{6}$ | $= \frac{2}{3} = 0.67$

$P(T|H) = 1$ $P(T,H) = \frac{1}{6}\ \frac{1}{3}$

$P(F|H) = 0$ $P(F,H) = 0$

~~$H(Pass|GPA) = \frac{1}{2}\times\frac{1}{6} + \frac{1}{2}\times\frac{1}{6} + \frac{1}{2}\times\frac{1}{6} + \frac{1}{2}\times\frac{1}{6}$~~

~~$+ 1\times\frac{1}{3} + 0\times 0 = \frac{2}{3} = 0.67$~~

~~$H(Pass|GPA) = -\frac{1}{2}\log\frac{1}{6}\times 4 - 1\log\frac{1}{3} =$~~

(3)

$P(T|T) = 1$ $P(T,T) = \frac{1}{2}$

$P(F|T) = 0$ $P(F,T) = 0$

$P(T|F) = \frac{1}{3}$ $P(T,F) = \frac{1}{6}$

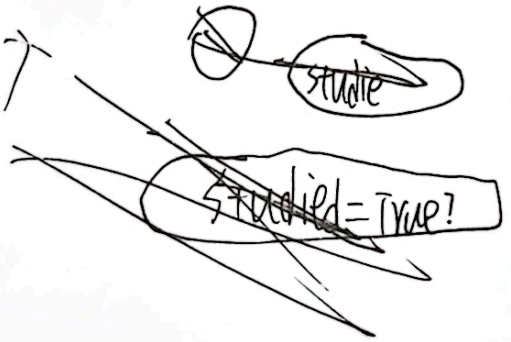
$P(F|F) = \frac{2}{3}$ $P(F,F) = \frac{1}{3}$

~~$H(Pass|Studied) = 1\times\frac{1}{2} + \frac{1}{6}\times 0 + \frac{1}{3}\times\frac{1}{3} + \frac{2}{3}\times\frac{1}{3} = 0.78$~~

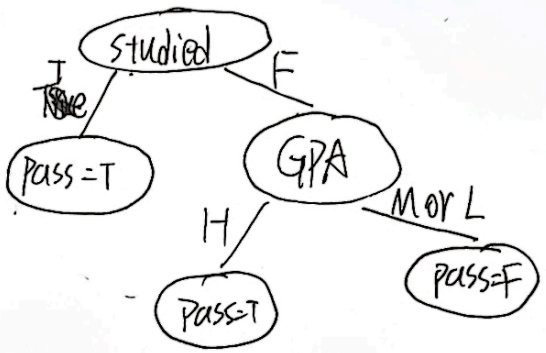
$H(Pass|Studied) = \frac{1}{2}\log 1 + \frac{1}{6}\log\frac{1}{3} + \frac{1}{3}\log\frac{2}{3}$

$= 0.47$

(4)




(4)



# Report-Assignment2

## Overview

In this assignment, I solve the wine analysis problem. Specifically, given its 13 features, a wine should be classified one of the three classes. To conduct such a multi-classes classification task, support verctor machine(SVM) is utilized.

## Support Vector Machine

### Hard Margin SVM

The basic idea of SVM is to maximize the minimum distance between data and the decision boundary(the so-called margin). Hence, the optimization problem can be formulated as:

$$max_{w,b}min_i \frac{y_i(w^T x_i + b)}{||w||}$$

where w, b are the parameter of the linear classifier, and $(x_i, y_i)$ are the supervised training data pair.

Since scaling (w,b) by some factor does not affect the objective function, the solution space of (w,b) can be reduced to a subspace where

$$y_i(w^T \hat{x} + b) = 1$$

where $\hat{x}$ is the data sample that is closest to the decision boundary. Hence the optimization problem can be converted to

$$\begin{aligned} &min_{w,b} \frac{1}{2}||w||^2 \\ s.t.\ \ &y_i(w^T x_i + b) \geq 1, \forall i \end{aligned}$$

## SVM With Slack Variables

For hard margin SVM, it is supposed that the data is linear separable. However, in real case, this is often impossible. Hence, slack variables are needed to relax this constraint. Specifically, each data sample is allowed to cross the margin even the decision boundary, but the number of these samples should not be too high. The problem can be formulated as

$$
\begin{gathered}
min_{w,b} \frac{1}{2}||w||^2 + C\sum_{i}^{m}\xi_i \\
s.t.\ \ y_i(w^T x_i + b) \geq 1 - \xi_i, \forall i \\
\xi \geq 0
\end{gathered}
$$

## Solving SVM With Dual

By using KKT conditions, the dual problem of the original optimization of SVM with slack variable can be converted to

$$
\begin{gathered}
max_{\alpha} \sum_{i}^{m}\alpha_i - \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j x_i^T x_j \\
s.t. \sum_{i}^{m}\alpha_i y_i = 0 \\
0 \leq \alpha_i \leq C
\end{gathered}
$$

w can be obtained by

$$
w = \sum_{i}^{m}\alpha_i y_i x_i
$$

b can be obtained by

$$
b = \frac{1}{|M|}\sum_{j \in M}(y_j - \sum_{i}^{m}\alpha_i y_i x_i^T x_j)
$$

## SVM With Kernel

Since in many cases, the training samples are not linearly separable, it is necessary to project data into a dimension that can be linearly splitted. Hence kernel is needed. the simple dot product $x_i x_j$ in the objective function is replaced by more complicated and flexible kernels $k(x_i, x_j)$. In this task, different kernels are tested.

## Training and Testing Setting

In this assignment, I use 80% of the data as training set, with a total number of 142, and the rest 20% of the data is used for testing, with 36 samples.

## Linear SVM Without Slack Variables

In this task, the kernel of SVM is the simplest linear model, and no slack variables are used. The training error rate is 0, meaning that the model well split the training data without error. However the testing error rate is 0.0278. Some samples are misclassified due to overfitting. For more information such as the support vector, weight and bias, please refer to A2_119010156.ipynb.

## Linear SVM With Slack Variables

| C | Train error | Test error |
|---|---|---|
| 0.1 | 0.0141 | 0.0278 |
| 0.2 | 0.0211 | 0.0278 |
| 0.3 | 0.0211 | 0.0000 |
| 0.3 | 0.0211 | 0.0000 |
| 0.5 | 0.0211 | 0.0000 |
| 0.6 | 0.0211 | 0.0000 |
| 0.7 | 0.0211 | 0.0000 |
| 0.8 | 0.0070 | 0.0000 |

| C | Train error | Test error |
|---|---|---|
| 0.9 | 0.0211 | 0.0000 |
| 1.0 | 0.0211 | 0.0000 |

As is shown in the table, when the constraint put on the slack variables are small, the classifer tend to output a worse test error, meaning an underfit. As the constraint on slack variables grow stronger, the test error decrease and eventually fall to 0, which is suitable. For more information please refer to A2_119010156.ipynb.

## SVM With Kernels

| kernel | Train error | Test error |
|---|---|---|
| 2nd-poly | 0.0000 | 0.0278 |
| 3nd-poly | 0.0000 | 0.0278 |
| RBF | 0.0070 | 0.0000 |
| Sigmoidal | 0.5634 | 0.4444 |

From the table, it can be concluded that RBF is the most suitable kernel, while Sigmoidal fail to map the data space into a linear separable space.