by z5244467 Chen Wu

1.(a) $\hat{\theta}^{\oplus} = (\frac{5}{23}, \frac{3}{23}, \frac{9}{23}, \frac{6}{23})$

$\hat{\theta}^{\ominus} = (\frac{11}{16}, \frac{3}{16}, \frac{0}{16}, \frac{2}{16})$

vector $x = (1, 0, 0, 2)$

Therefore, $P(x|\oplus) = 3! \frac{(\frac{5}{23})^1 \cdot (\frac{3}{23})^0 \cdot (\frac{9}{23})^0 \cdot (\frac{6}{23})^2}{1! \cdot 0! \, 0! \, 2!} = 0.04446$

$P(x|\ominus) = 3! \frac{\frac{11}{16} \cdot (\frac{3}{16})^0 \cdot 0^0 \cdot (\frac{2}{16})^2}{1! \, 0! \, 0! \, 2!} = 0.0322$

$$P(+|x_*) = \frac{P(x_*|+) \cdot P(+)}{P(x_*)} = \frac{3! \frac{5}{23} (\frac{6}{23})^2}{1!2!} \cdot \frac{1}{2} \Big/ \frac{1}{2} P(x|+) + \frac{1}{2} P(x|-)$$

$$= 0.5796.$$

(b). apply added 1 smoothing.

$\hat{\theta}^{\oplus} = (\frac{6}{27}, \frac{4}{27}, \frac{10}{27}, \frac{7}{27})$

$\hat{\theta}^{\ominus} = (\frac{12}{20}, \frac{4}{20}, \frac{1}{20}, \frac{3}{20})$

$P(x|+) = 3! \frac{(\frac{6}{27} \cdot (\frac{4}{27})^0 \cdot (\frac{10}{27})^0 \cdot (\frac{7}{27})^2)}{1! \, 0! \, 0! \, 2!} = 0.0648$

$P(x|-) = 3! \frac{(\frac{12}{20} \cdot (\frac{4}{20})^0 \cdot (\frac{1}{20})^0 \cdot (\frac{3}{20})^2}{1! \, 0! \, 0! \, 2!} = 0.0405$

$P(-|x_*) = \frac{P(x_*|-) P(-)}{P(x_*)} = \frac{0.0405 \times \frac{1}{2}}{\frac{1}{2}(0.0648 + 0.0405)} = 0.4375$

c. convert every word to 1, which can be obtained.

A	B	C	D	
1	0	1	1	+
0	1	1	0	+
1	0	0	1	+
0	0	1	0	+
0	0	0	1	−
1	0	0	0	−
1	1	0	0	−
1	0	0	1	−

$\theta^{\wedge}+ = (\frac{2}{4}, \frac{1}{4}, \frac{3}{4}, \frac{2}{4})$

$\theta^{\wedge}- = (\frac{3}{4}, \frac{1}{4}, \frac{0}{4}, \frac{2}{4})$

vector $x = (1, 0, 0, 1)$

$P(x|+) = \frac{2}{4} \times (1-\frac{1}{4}) \times (1-\frac{3}{4}) \times \frac{2}{4} = 0.0469$

$P(x|-) = \frac{3}{4} \times (1-\frac{1}{6}) \times 1 \times \frac{2}{4} = \frac{3}{4} \times \frac{3}{4} \times \frac{2}{4} = 0.283$

Therefore, $P(+|x_*) = \frac{P(x_*|+) \cdot P(+)}{P(x_*)} = \frac{1}{7} = 0.1429$

d. apply add 1 smoothing.

A	B	C	D	
1	1	1	1	+
0	0	0	0	+
1	0	1	1	+
0	1	1	0	+
1	0	0	1	+
0	0	1	0	+
0	0	0	1	−
1	0	0	0	−
1	1	0	0	−
1	0	0	1	−
1	1	1	1	−
0	0	~	~	−

$\theta^{\wedge}\oplus = (\frac{3}{6}, \frac{2}{6}, \frac{4}{6}, \frac{3}{6})$

$\theta^{\wedge}\ominus = (\frac{4}{6}, \frac{2}{6}, \frac{1}{6}, \frac{3}{6})$

vector $x = (1, 0, 0, 1)$

$P(x|+) = \frac{3}{6} \times (1-\frac{2}{6}) \times (1-\frac{4}{6}) \times \frac{3}{6} = \frac{1}{18}$

$P(x|-) = \frac{4}{6} \times (1-\frac{2}{6}) \times (1-\frac{1}{6}) \times \frac{3}{6} = \frac{5}{27}$

$P(-|x_*) = \frac{P(x_*|-) \cdot P(-)}{P(x_*)} = \frac{10}{13} = 0.7692$

2. (a) $\frac{\partial L_c(y_i, \hat{y}_i)}{\partial w_0} = \frac{\partial\left(\frac{1}{c^2}(y_i - (w_0 + w_1 x_i))^2 + 1\right)}{2\sqrt{\frac{1}{c^2}(y_i - (w_0 + w_1 x_i))^2 + 1}}$

$= \frac{\frac{2}{c^2}(y_i - (w_0 + w_1 x_i))(-1)}{2\sqrt{\frac{1}{c^2}(y_i - (w_0 + w_1 x_i))^2 + 1}}$

$= \frac{w_0 + w_1 x_i - y_i}{c\sqrt{(y_i - (w_0 + w_1 x_i))^2 + c^2}}$

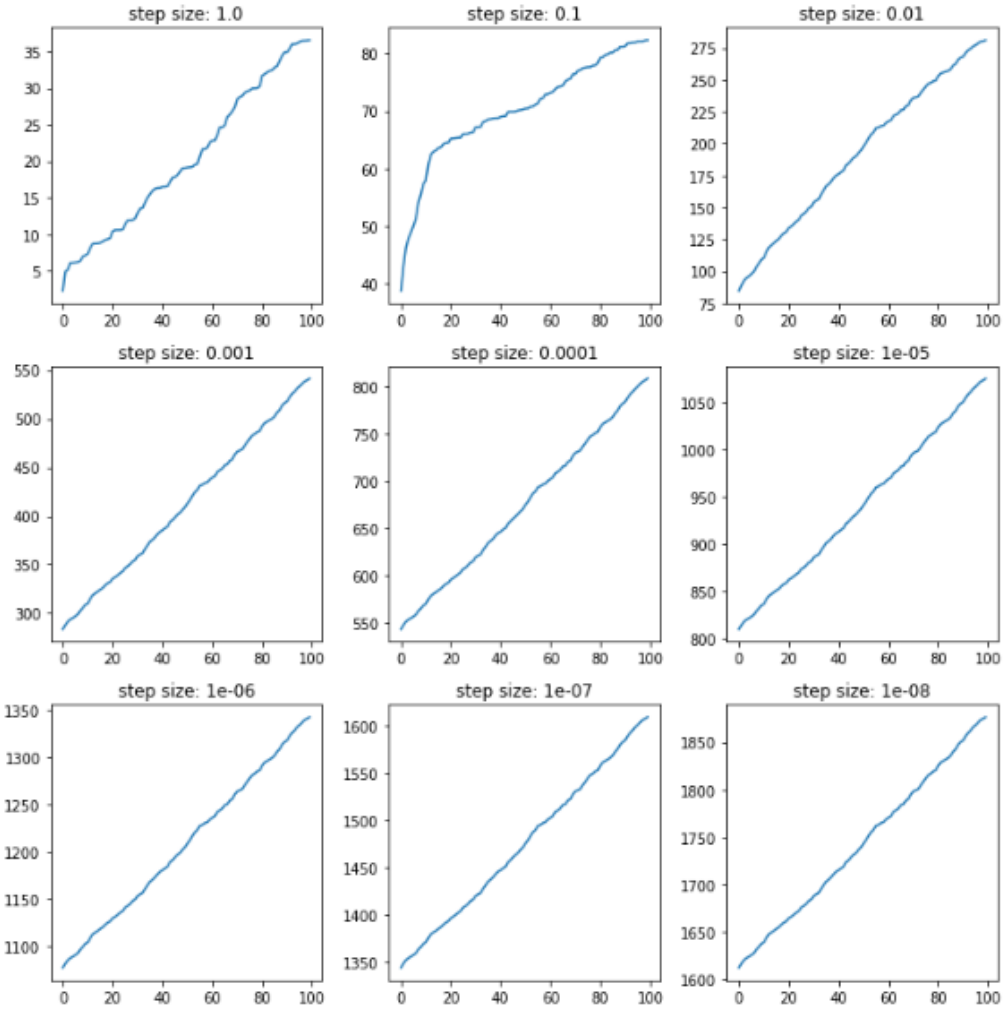
$\frac{\partial L_c(y_i, \hat{y}_i)}{\partial w_1} = \frac{\partial\left(\frac{1}{c^2}(y_i - (w_0 + w_1 x_i))^2 + 1\right)}{2\sqrt{\frac{1}{c^2}(y_i - (w_0 + w_1 x_i))^2 + 1}}$

$= \frac{-y_i x_i + w_0 x_i + x_i^2 w_1}{c\sqrt{(y_i - (w_0 + w_1 x_i))^2 + c^2}}$

(b). $w_0^{(t+1)} = w_0^{(t)} - 2\frac{-y_i + w_0 + w_1 x_i}{\sqrt{(y_i - w^T x_i)^2 + c^2}}$

$w_1^{(t+1)} = w_1^{(t)} - 2\frac{-y_i x_i + w_0 x_i + x_i^2 w_1}{\sqrt{(y_i - w^T x_i)^2 + c^2}}$

2.(c)

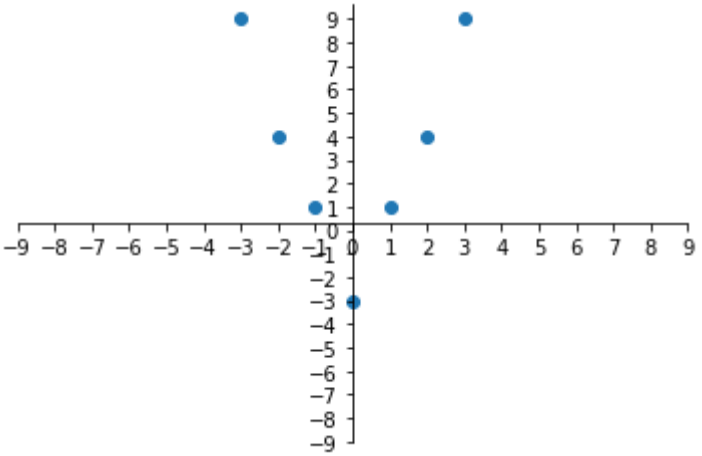


(d). From these graphs, it is clear that in the first three figures, In the first three pictures, with the change of step size, the line changes a lot. However, then even if the step size continues to shrink, the lines are basically unchanged. It means that no matter how to update step size, the optimal solution of the model has been reached at this time.

(f) Since different values of c can induce different optimal step-size varies, it is necessary to choose various values of c to run the codes. For example, c= [2 ,4, 8, 16…100], and plot graphs which can show the clear result.

4.(a)

Yes, it is linearly separable.



(b). According to data analysis, a straight line that can divide (-1, 1), (2, 4), (1, 1) meets the condition, so

$$X=\begin{pmatrix}-1 & 1\\ 1 & 1\\ 2 & 4\end{pmatrix} \qquad y=\begin{pmatrix}-1\\ 1\\ -1\end{pmatrix}$$

the X' incorporates the class labels, then calculate X' and X^T,

$$X'=\begin{pmatrix}1 & -1\\ 1 & 1\\ -2 & -4\end{pmatrix} \qquad X^T=\begin{pmatrix}1 & 1 & -2\\ -1 & 1 & -4\end{pmatrix}$$

The Gram matrix is

$$X'X'^T=\begin{pmatrix}2 & 0 & 2\\ 0 & 2 & -6\\ 2 & -6 & 20\end{pmatrix}$$

Hence, the dual optimisation problem is thus

$$argmax_{a1,a2,a3}-\frac{1}{2}(2a_1^2+2a_1a_3+2a_1a_3+2a_2^2-6a_2a_3-6a_3a_2+20a_3^2)+a_1+a_2+a_3$$

$$= argmax_{a1,a2,a3}-\tfrac{1}{2}(2a_1^2+4a_1a_2+2a_2^2-12a_2a_3+20a_3^2)+a_1+a_2+a_3$$

Subject to $a_1 \geq 0\,, a_2 \geq 0, a_3 \geq 0$ and $-a_1+a_2-a_3=0.\ then\ a_2=a_1+a_3$

From this we obtain the next expression, and its following simplification

$$argmax_{a1,a2,a3}-\frac{1}{2}(4a_1^2-4a_1a_3+10a_3^2)+2a_1+2a_3$$

$$-4a_1+2a_3+2=0\ and\ -10a_3+2a_1+2=0$$

The result is $a_1=\frac{2}{3}\ , a_3=\frac{1}{3},$ then $a_2=1$.

(c). Therefore, we can obtain the solution

$$w = \frac{2}{3} * x1 + 1 * x2 + \frac{1}{3} * x3$$

$$= \frac{2}{3}(1 \quad -1) + 1(1 \quad 1) + \frac{1}{3}(-2 \quad -4)$$

$$= (1 \quad -1)$$

Thus, the margin $\frac{1}{||w||} = \frac{1}{\sqrt{(1^2+(-1)^2)}} = \frac{\sqrt{2}}{2}$

t can be obtained from any support vector, say x_3, Since $y_3(w * x_3 - t) = 1$,So $t = -1$ $w = (1 \quad -1)$ margin is $\frac{\sqrt{2}}{2}$

(d). Linearity and non-linearity are based on model parameters and input characteristics; for example, if input x, model y = ax + ax ^ 2, then it is a non-linear model. If the input is x and X ^ 2, the model is linear. The linear classifier has good interpretability and low computational complexity. However, it does not have the fitting effect of the nonlinear function. If the linear classifier is used for representation, tend to be underfitting.

(e).

In short, the kernel is a shortcut that can help us perform certain calculations faster, otherwise it will involve calculations in high-dimensional spaces. It allows us to operate in the original feature space without calculating data coordinates in the high-dimensional space. Via the kernel, there are unlimited things can be done. For instance, f(x) can be a mapping from n dimensions to infinite dimensions, and we may hardly understand how to deal with it. Then the kernel provides us with a great shortcut.

5.(a)