1.1 Problem: Overfitting and Regularized Logisitc Regression (5 pts.)

1) (2 pts.) Plot the sigmoid function $1/(1+\exp^{-wX})$ for increasing weights $w \in \{1, 5, 100\}$ with $X \in \mathbb{R}$. A qualitative sketch will suffice. Utilize these plots to explain why large weights can lead to overfitting in logistic regression.
2) (3 pts.) To mitigate overfitting, it is preferable to have smaller weights. To accomplish this, rather than utilizing maximum conditional likelihood estimation M(C)LE for logistic regression:

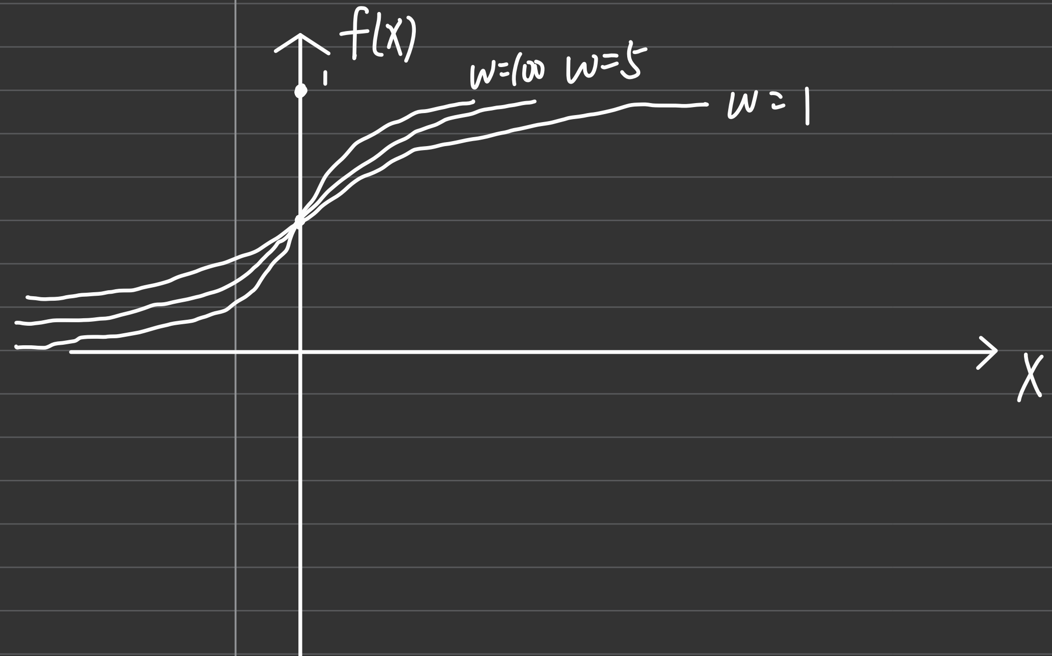
$$\max_{w_0,\cdots,w_d} \prod_{i=1}^{n} P(Y_i|X_i, w_0, \cdots, w_d), \tag{1}$$

we can consider maximum conditional a posterior M(C)AP estimation:

$$\max_{w_0,\cdots,w_d} \prod_{i=1}^{n} P(Y_i|X_i, w_0, \cdots, w_d)P(w_0, \cdots, w_d), \tag{2}$$

where $P(w_0, \cdots, w_d)$ is a prior on the weights.
Given a standard Gaussian prior $\mathcal{N}(0, \mathbf{I})$ for the weight vector w, please derive the **gradient ascent** update rules for the weights and explain why M(C)AP can address overfitting issue.

(1).



Larger weights makes f(x) larger with same X, so the penalty is larger when a sample is misclassified. As a result, more samples are forced to be classified to the "right" class even though they are very close to another class. So it leads to overfitting.

(2). $\max_{w_0,\cdots,w_d} \prod_{i=1}^{n} P(Y_i|X_i, w_0, \cdots, w_d)\, P(w_0, \cdots, w_d)$

$\Leftrightarrow \max_{w_0,\cdots,w_d} \sum_{i=1}^{n} \log(P(Y_i|X_i, \cdots, w_d)) + n \log P(w_0, \cdots, w_d)$

$\Leftrightarrow \max_{w_0\cdots w_d} \sum_{i=1}^{n} y_i \log f_w(x_i) + (1-y_i)\log(1-f_w(x_i)) - \frac{n}{2} w^T w$

∇J denotes gradient of $J(w) = y_i \log f_w(x_i) + (1-y_i)\log(1-f_w(x_i)) - \frac{\eta}{2} w^T w$

$W_{update} = w + \alpha \Delta J$, where α is learning rate.

As you can see, in $J(w)$ we get $-\frac{\eta}{2} w^T w$, since we need to maximize $J(w)$, so we would make w small and thus to mitigate overfitting.

2、(1). $P(y_c^i=1|x^i;w)$

$$= \frac{\exp(w_c^T x^i)}{\sum_{c'} \exp(w_{c'}^T x^i)}$$

$$= \frac{\exp((w_c - w_k)^T x^i)}{\sum_{c'} \exp((w_{c'} - w_k)^T x^i)}$$ (divide $\exp(w_k x^i)$)

$$= \frac{\exp((w_{c'} - w_k)^T x^i)}{1+\sum_{c'=1}^{k-1} \exp((w_{c'} - w_k)^T x^i)}$$

$$= \begin{cases} \dfrac{\exp(w_c^T x^i)}{1+\sum_{c'=1}^{k-1} \exp(w_{c'}^T x^i)}, & \text{if } c<k. \\ \dfrac{1}{1+\sum_{c'=1}^{k-1} \exp(w_{c'}^T x^i)}, & \text{if } c=k \end{cases}$$ (set $w_k^T=0$)

(2). $$L(w) = \ln \prod_{j=1}^{n} p(y_c^j|x^j, w)$$

$$= \ln \prod_{j=1}^{n} \sum_{c'} \frac{\exp(w_c^T x^j)}{\sum_{c'} \exp(w_{c'}^T x^j)} \cdot y_c$$ (Only need the right class probability, so multiply y_c)

$$= \sum_{j=1}^{n} \sum_{c=1}^{k} [y_c^j (w_c^T x^j) - y_c^j \ln(\sum_{c'} \exp(w_{c'}^T x^j))]$$

($y_c=1$ when $j=c$, else 0)

(3). $$\nabla_{w_c} \sum_{c=1}^{k} y_c^j \ln(\sum_{c'} \exp(w_{c'}^T x^j))$$

$$= \left(\frac{\nabla_{w_c}(\exp(w_{c'}^T x^j))}{\sum_{c'} \exp(w_{c'}^T x^j)} \right)$$

$$= x^j \frac{\exp(w_{c'}^T x^j)}{\sum_{c'} \exp(w_{c'}^T x^j)}$$ (consider fixed j, most values 0)

(4). $\nabla_{w_c} L(w) = \sum_{j=1}^{n} y_c^j x^j + \sum_{j=1}^{n} x^j \frac{\exp(w_c^T x^j)}{\sum_{c'} \exp(w_{c'}^T x^j)}$

$= \sum_{j=1}^{n} x^j (y_c^j - p(y_c^j = 1 \mid x^j; w)$

(5).

$W_{update} = W + \eta \cdot \nabla_{w_c} L(w)$

$= W + \eta \cdot \sum_{j=1}^{n} x^j (y_c^j - p(y_c^j = 1 \mid x^j; w))$

1.3 Problem: Support Vector Machine 1 (6 pts.)

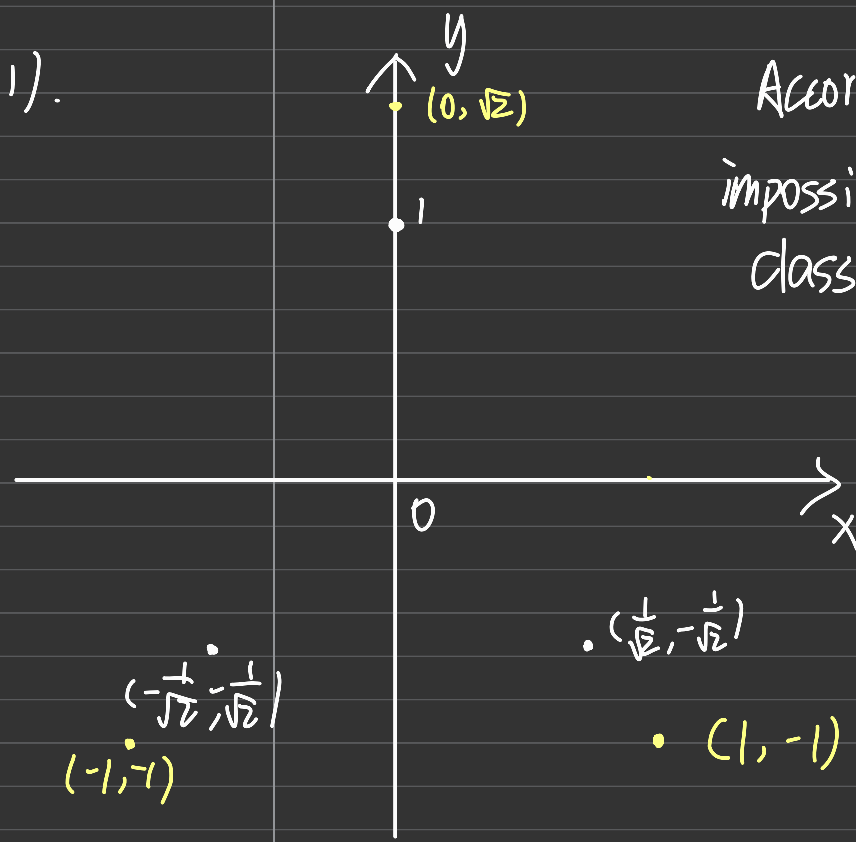
Given a binary data set:

Class -1: $\{(0,1),(\frac{1}{\sqrt{2}},-\frac{1}{\sqrt{2}}),(-\frac{1}{\sqrt{2}},-\frac{1}{\sqrt{2}})\}$. Class +1: $\{(0,\sqrt{2}),(1,-1),(-1,-1)\}$.

1) (3 pts.) Could you find an SVM classifier (without slack variable) for this dataset? Please explain your reasoning, possibly with the help of a plot sketch.

2) (3 pts.) Use SVM by expanding the original feature vector $x=[x_1,x_2]$ to $x=[x_1^2,x_2^2]$, find the svm of this given data set and predict the label of $(-\frac{1}{2},\sqrt{2})$.

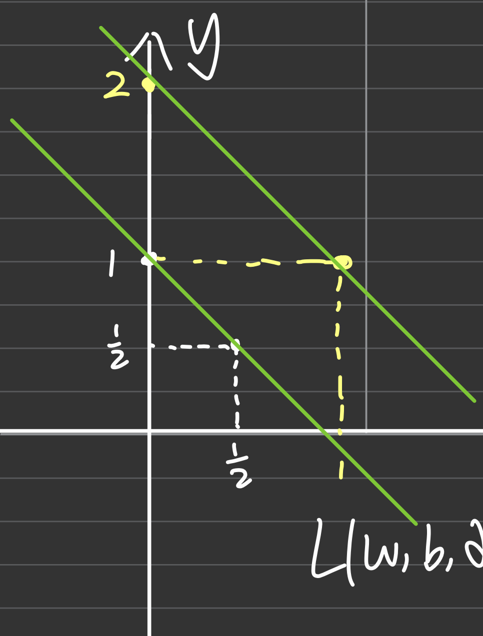
(1).



According to the graph, its impossible to find an SVM classifier.

(2). Class −1: $\{(0,1), (\frac{1}{2},\frac{1}{2})\ (\frac{1}{2},\frac{1}{2})\}$

Class 1: $\{(0,2), (1,1), (1,1)\}$



The opetimization problem is

$$\min \frac{1}{2}\|W\|^2$$

s.t.

$$(w_1+b)\cdot(-1)\geq 1$$
$$(\tfrac{1}{2}w_0+\tfrac{1}{2}w_1+b)\cdot(-1)\geq 1$$
$$(2w_1+b)\cdot 1\geq 1$$
$$(w_0+w_1+b)\cdot 1\geq 1$$

$$L(w,b,\alpha)=\frac{1}{2}(w_1^2+w_0^2)+\alpha_1(1+w_1+b)+\alpha_2(1+\tfrac{1}{2}w_0+\tfrac{1}{2}w_1+b)$$
$$+\alpha_3(1-2w_1-b)+\alpha_4(1-w_0-w_1-b)$$

$$\frac{\partial L}{\partial w} = \begin{pmatrix} w_0 + \frac{1}{2}\alpha_2 - \alpha_4 \\ w_1 + \alpha_1 + \frac{1}{2}\alpha_2 - 2\alpha_3 - \alpha_4 \end{pmatrix} = 0$$

$$\frac{\partial L}{\partial b} = \alpha_1 + \alpha_2 - \alpha_3 - \alpha_4 = 0$$

$\alpha_i \geq 0$, for $i = 1, 2, 3, 4$

$$\begin{cases} 1 + w_1 + b \leq 0 \\ \frac{1}{2}w_0 + \frac{1}{2}w_1 + b + 1 \leq 0 \\ 1 - 2w_1 - b \geq 0 \\ 1 - w_0 - w_1 - b \geq 0 \end{cases}$$

$$\alpha_1 (1 + w_1 + b) = 0$$

$$\alpha_2 (\tfrac{1}{2}w_0 + \tfrac{1}{2}w_1 + b + 1) = 0$$

$$\alpha_3 (1 - 2w_1 - b) = 0$$

$$\alpha_4 (1 - w_0 - w_1 - b) = 0$$

KKT condition

$$\Rightarrow \begin{cases} w_0 = 2 \\ w_1 = 2 \\ b = -3 \end{cases}$$

$\frac{1}{2}$ tl4.

The lable of $(-\frac{1}{2}, \sqrt{2})$ should be +1.

1.4 (1).

$$\min_{W,s} \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{n}(s_i + t_i)$$

s.t.

$$y_i - f(x_i) - \varepsilon \le s_i \quad , i = 1, 2, 3, \cdots, n$$

$$f(x_i) - y_i - \varepsilon \le t_i \quad , i = 1, 2, 3, \cdots, n$$

$$s_i \geq 0, \ t_i \geq 0, \quad i = 1, 2, \cdots, n$$

Where $f(x) = Wx$

(2) $L(W, t, s, \alpha, \beta, r, \theta)$

$$= \frac{1}{2}\|W\|^2 + C\sum_{i=1}^{n} s_i + \sum_{i=1}^{n}\left((y_i - f(x_i) - \varepsilon - s_i)\alpha_i + \beta_i(-y_i + f(x_i) - \varepsilon - s_i)\right)$$

$$- \sum_{i=1}^{n} s_i r_i - \sum_{i=1}^{n}\theta_i t_i$$

(3).

$$g(\alpha, \beta, r, \theta) = \min_{x \in R^n} L(w, b, s, \alpha, \beta, r, \theta)$$

$$\frac{\partial g}{\partial w} = 0 \Rightarrow \quad w = \sum_{i=1}^{n}(\alpha_i - \beta_i)x_i$$

and

$$\alpha_i(\varepsilon + s_i - y_i + f(x_i)) = 0$$

$$\beta_i(\varepsilon + t_i + y_i - f(x_i)) = 0$$

$$0 \le \alpha_i \quad , i = 1, 2, \cdots, n$$

$$0 \le \beta_i \quad , i = 1, 2, \cdots, n$$

so dual form is

$$\max g(\alpha, \beta, r, \theta)$$

$$\Updownarrow$$

$$\min \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \beta_i)(\alpha_j - \beta_j)x_j^T x_i + \sum_{i=1}^{n}\left((\varepsilon - y_i)\cdot\alpha_i + (\varepsilon + y_i)\cdot\beta_i\right)$$

subject to $\alpha_i \geq 0, \ \beta_i \geq 0$

$i = 1, 2, \cdots, n$

(4). Yes.

(5). The support vectors are points lie on the margin or in the ε boundary around the trained decision boundary.

(6). $g(x) = \sum_{i=1}^{n} (\alpha_i - \beta_i) x_i x$ (7). Yes. substitue x with $k(x)$.

(8). The decision area will get larger, which means the error that can be toleranted gets larger if ε is smaller, and the model can be more complex and easier to be overfitting.

If ε is larger, the model gets less complex, more values can be fitted in the area without large penalty. However, it may leads to loss of precision.

(9). If C is smaller, it may result in a larger margin and a simpler model, and the model may be more generalized.

If C is larger, it may result in a smaller margin and a complex model, which means it provides higher precision but may lead to overfitting.

1.5 First iteration:

$$H(D|A) = \sum_{i=1}^{n} \frac{|D_i|}{D} H(D_i)$$

$$= \sum_{i=1}^{n} \frac{|D_i|}{|D|} \sum_{k=1}^{K} \frac{|D_{ik}|}{|D_i|} \log_2 \frac{|D_{ik}|}{|D_i|}$$

$$H(D) = -\frac{6}{15}\log_2\frac{6}{15} - \frac{9}{15}\log_2\frac{9}{15}$$

$$g_R(D,A) = \frac{H(D) - H(D|A)}{H_A(D)}$$

Age: $H(D|A) = \frac{1}{3}\cdot\frac{3}{5}\log_2\frac{3}{5} - \frac{1}{3}\cdot\frac{2}{5}\log_2\frac{2}{5} - \frac{1}{3}\log_2\frac{3}{5} - \frac{1}{3}\log_2\frac{2}{5} - \frac{1}{3}\cdot\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{3}\cdot\frac{1}{5}\log_2\frac{1}{5}$

$$= -\frac{2}{5}\log_2\frac{3}{5} - \frac{4}{15}\log_2\frac{2}{5} - \frac{4}{15}\log_2\frac{4}{5} - \frac{1}{15}\log_2\frac{1}{5}$$

$$H_A(D) = -\frac{3}{5}\log_2\frac{1}{5}$$

$$g_R(D,Age) = \frac{H(D) - H(D|A)}{H_A(D)} = 0.05$$

$g_R(D, work) = 0.35$

$g_R(D, house) = 0.43$

$g_R(D, credit) = 0.21$

$A_g =$ house

$D_1 = \{1, 2, 5, 6, 7, 13, 14, 15\}$

$D_2 = \{4, 8, 9, 10, 11, 12\}$ $A = \{$Age, Work, Credit$\}$

For D_1. $g_R(D_1, Age) = 0.16$

$g_R(D, work) = 1$

so work is chosen

$D_1 \rightarrow \begin{cases} D_3 \{1, 2, 5, 6, 7, 15\} \\ D_4 \{3, 13, 14\} \end{cases}$

For D_2: belongs to Yes.

D_3: belongs to No

D_4: belongs to Yes. Done.