Zheng Wen 7112-8072-12

1-a, for $\forall \underline{x_1}, \underline{x_2} \in H$, $g(\underline{x_1}) - g(\underline{x_2}) = (w_0 + \underline{w}^T\underline{x_1}) - (w_0 + \underline{w}^T\underline{x_2}) = \underline{w}^T(\underline{x_1} - \underline{x_2}) = 0$

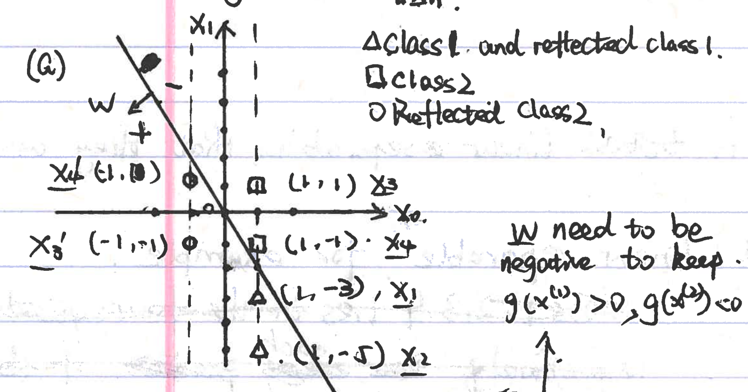
so $\underline{w} \perp (\underline{x_1} - \underline{x_2})$ for $\underline{x_1}, \underline{x_2}$ are vectors start from origin, $\underline{x_1} - \underline{x_2}$ is an arbitrary vector on H, so $\underline{w} \perp (\underline{x_1} - \underline{x_2}) \Rightarrow \underline{w} \perp H$

1-b. for $\underline{x} = \underline{x_1} + \underline{w}$, where $\underline{x_1}$ is a point on H. the distance between $\underline{x}$ and H $= \frac{w_0 + \underline{w}^T\underline{x}}{\|\underline{w}\|} = \frac{w_0 + \underline{w}^T\underline{x_1} + \underline{w}^T\underline{w}}{\|\underline{w}\|} = \|\underline{w}\| > 0$, so, after adding $\underline{w}$, a point on H now have a positive distance between H and $\underline{x}$, so $\underline{w}$ point to the positive side of H

1-C. In augmented space, for $\forall \underline{x}$ on H, $\underline{w}^T\underline{x} = 0$, so $\underline{w} \perp \underline{x}$, ie $\underline{w} \perp H$.
So for any point x' in the space, its distance between it and H could be calculated by projection of $\underline{x}$ on $\underline{w}$, So $dist(\underline{x}', H) = \frac{\underline{w}^T\underline{x}'}{\|\underline{w}\|}$.

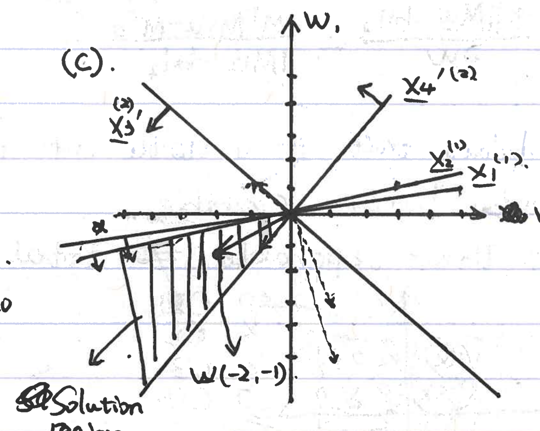
1-d for $\forall \underline{w}$ on hyperplane $g(\underline{x}) = \underline{w}^T\underline{x} = 0$, because $\underline{w}^T\underline{x} = 0$, so $\underline{x} \perp \underline{w}$, same as (c), $dist(\underline{w}, g(\underline{x})) = \frac{\underline{w}^T\underline{x}}{\|\underline{x}\|}$.

2. (a)



Feature Space. $H: -2x_0 - x_1 = 0$

(c).



Weight Space.

(b) for points from different classes x_i, $g(x_i) > 0$, ~~...~~, so the decision boundary classifies them correctly.

(d). $w(-2, -1)$ in the solution region

3.(a). For $\forall i$, $\frac{\partial f[p(\underline{x})]}{\partial x_i} = \frac{d f[p(\underline{x})]}{d p(\underline{x})} \cdot \frac{\partial p(\underline{x})}{\partial x_i}$.

So $\nabla_{\underline{x}} f[p(\underline{x})] = \begin{bmatrix} \frac{\partial f[p(\underline{x})]}{\partial x_1} \\ \vdots \\ \frac{\partial f[p(\underline{x})]}{\partial x_n} \end{bmatrix} = \frac{d f[p(\underline{x})]}{d p(\underline{x})} \begin{bmatrix} \frac{\partial p(\underline{x})}{\partial x_1} \\ \vdots \\ \frac{\partial p(\underline{x})}{\partial x_n} \end{bmatrix} = \frac{d f(p)}{dp} \nabla_{\underline{x}} p(\underline{x})$.

(b). For $\frac{\partial}{\partial \underline{x}}[\underline{x}^T M \underline{x}] = [M + M^T]\underline{x}$, $\frac{\partial}{\partial \underline{x}}(\underline{x}^T M \underline{x}) \overset{M=I}{=} \frac{\partial}{\partial \underline{x}}(\underline{x}^T\underline{x}) = 2\underline{x}$

(c) For $\forall i$, $\frac{\partial}{\partial x_i}[\underline{x}^T\underline{x}] = \frac{\partial}{\partial x_i}\left[(x_1 \cdots x_n)\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}\right] = \frac{\partial}{\partial x_i}\sum_{i=1}^{n} x_i^2 = 2x_i$,

So $\frac{\partial}{\partial \underline{x}}[\underline{x}^T\underline{x}] = \nabla_{\underline{x}}[\underline{x}^T\underline{x}] = \begin{bmatrix} 2x_1 \\ \vdots \\ 2x_n \end{bmatrix} = 2\underline{x}$

~~(d). $\nabla_{\underline{x}}[(\underline{x}^T\underline{x})^3] = \nabla_{\underline{x}}[(\underline{x}^T\underline{x})(\underline{x}^T\underline{x})(\underline{x}^T\underline{x})]$ for $(\underline{x}^T\underline{x})^T = \underline{x}^T\underline{x}$, suppose the second $\underline{x}^T\underline{x} = M$, so using (b), $\nabla_{\underline{x}}[(\underline{x}^T\underline{x})(\underline{x}^T\underline{x})(\underline{x}^T\underline{x})] = [(\underline{x}^T\underline{x}) + (\underline{x}^T\underline{x})^T](\underline{x}^T\underline{x}) = 2(\underline{x}^T\underline{x})(\underline{x}^T\underline{x})$~~

(d) $\nabla_x[(x^Tx)^3] = \frac{\partial[(x^Tx)^3]}{\partial x} = \frac{d[(x^Tx)^3]}{d x^Tx}\cdot\frac{\partial x^Tx}{\partial x} = 3(x^Tx)^2\cdot 2x = 6(x^Tx)x$

4. (a) $\frac{\partial\|w\|_2}{\partial w} = \frac{\partial\sqrt{w^Tw}}{\partial w} = \frac{d\sqrt{w^Tw}}{d\, w^Tw}\,\frac{\partial w^Tw}{\partial w} = \frac{1}{2\sqrt{w^Tw}}\cdot 2w = \frac{w}{\|w\|_2}$

(b) $\frac{\partial\|Mw-b\|_2}{\partial w} = \frac{\partial\sqrt{(Mw-b)^T(Mw-b)}}{\partial w} = \frac{d\sqrt{(Mw-b)^T(Mw-b)}}{d(Mw-b)^T(Mw-b)}\cdot\frac{\partial(Mw-b)^T(Mw-b)}{\partial w}$

$\frac{d\sqrt{(Mw-b)^T(Mw-b)}}{d(Mw-b)^T(Mw-b)} = \frac{1}{2\sqrt{\cdots}} = \frac{1}{2\|Mw-b\|_2}$

$(Mw-b)^T(Mw-b) = [(Mw)^T - b^T](Mw-b) = (w^TM^T - b^T)(Mw-b)$

$= w^TM^TMw - w^TM^Tb - b^TMw + b^Tb$.

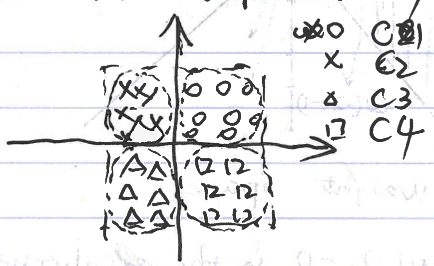
So $\frac{\partial(Mw-b)^T(Mw-b)}{\partial w} = \frac{\partial w^TM^TMw}{\partial w} - \frac{\partial w^TM^Tb}{\partial w} - \frac{\partial b^TMw}{\partial w} + \frac{\partial b^Tb}{\partial w}$

$= 2M^TMw - M^Tb - M^Tb + 0 = 2M^TMw - 2M^Tb$.

So $\frac{\partial\|Mw-b\|_2}{\partial w} = \frac{M^TMw - M^Tb}{\|Mw-b\|_2}$

5. It's obvious that if a ~~data~~ set is ~~total~~ linear separable, that they are pair-wise linear separable.

but linear separable $\not\Rightarrow$ total linear separable. for example:



~~class 1, 2, 3, 4 lies in the four quadrants~~ respectively, ~~the~~, ~~each pair of them are linear separable~~, ~~but they are not total linear separable, for we can't find a line to separate any one of them apart from other three classes~~.

class 1, 2, 3, 4 lies in four quardrents respectively, using nearest neighboorhood method, ~~i.e. arg max~~, i.e. ~~arg~~ $\arg\max_x -\|x-\bar{x}\|$, they could be correctly classified. and ~~they~~ the decision boundary between them are ~~x~~ x-axis and y-axis. So using MVM, they could be classified correctly, but obviously, they are not ~~linear~~ totally linear separable, because we cannot find a line to separate one class from the others.