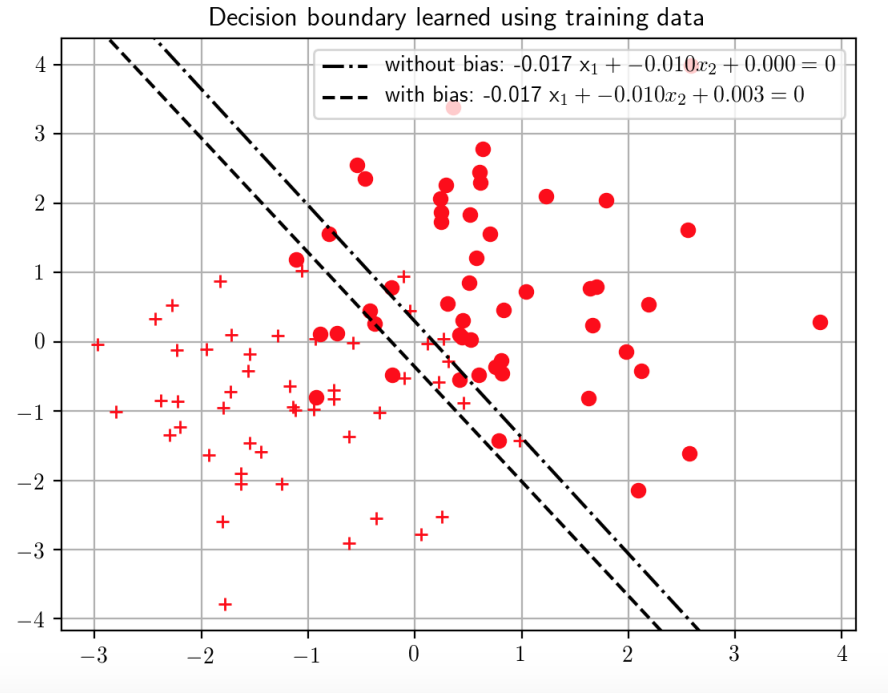
# Sprind 2018: CSCI-UA 473
# Machine Learning
# Assignment 1 (due Feb. 12 , 2018)


## Kexin Huang
## NetID: kh2383


1. It depends on the size of input.
If the input x has a large size, due to the law of large numbers, no matter how skewed is the distribution, the data distribution itself already represents the dataset itself, and the law says that the expected value approximates to the average of each example.
However, if the input size is small, then the input distribution couldn't be approximated by average value and it does matter. For example, we have only 10 examples of input x where only one x equals to 1, 1 x equals to 2, 8 x equals to 3 then the expected value should be 0.1*1+0.1*2+0.8*3=2.7 while the average gives 2.

2. Bias B gives more variation on classifying data, i.e. given only w, the decision boundary has to pass the origin, but with bias, it can translate in the plane. For example,




in the graph, we can clearly see due to the effect of translation , the linear function with bias classify better as it included points where function without bias couldn't classify

3. When $w^T x + b = 0$, it becomes a problem. Because when $w^T x + b = 0$, the

distance function will always be 0 no matter x is classified right or wrong as $D = -(M^*(x) - M(x)) * 0 = 0$. Then if x is classified not coincide with the ground truth value (i.e. when ground truth is positive as sign function defines to be negative if x $\leq$ 0), the distance function should give a distance but it will actually give 0. This does not serve the purpose of distance function.

One possible solution is add a small function on D, i.e.
$D = -(M^*(x) - M(x)) * (w^T X + b) + \mathbb{1}_{w^T X+b=0}(M(X) * \eta)$
where $\mathbb{1}_{w^T X+b=0} = 1$when $w^T X + b = 0$and =0 when otherwise.

The reason is when $w^T X + b = 0$, the problem with this distance function is when it misclassifies, it should give some distance but it gives 0. So we can add on some small number attached to D to give it a small distance. The reason why we give small distance $\eta$ is that near the decision boundary the distance is smaller than away from the contrary side. And the reason why we include M(x) is when the truth label is negative, M(x)=0=W*X+b, then it classifies right, so we don't need the later add function, and this M(x)=0 help automatically make the add function becomes 0. Also notice, this function won't change any other distance measure for any other X values as the $\mathbb{1}_{w^T X+b=0}$ will become 0.

4. $$D(M^*(X), M, X) = -(y^* log(\sigma(w^T X + b))) + (1 - y^*) log(1 - \sigma(w^T X + b)))$$

Using chain rule, we have

$$\frac{dD}{dw} = -y^* log'(\sigma(w^T X + b)) * \sigma'(w^T X + b) * (w^T X + b)' + (1 - y^*) log'(1 -$$

$$\sigma(w^T X + b)) * \sigma'(w^T X + b) * (w^T X + b)'$$

We know following derivative:

$$1. log'(x) = 1/x \text{ given log base is natural log}$$

$$2. \sigma'(x) = -\frac{1}{(1+e^{-x})^2} * e^{-x} * (-1) = \frac{1}{(1+e^{-x})} * \frac{e^{-x}}{(1+e^{-x})} = \frac{1}{(1+e^{-x})} * \frac{1+e^{-x}-1}{(1+e^{-x})} =$$

$$\sigma(x) * (1 - \sigma(x))$$

$$3. (w^T X + b)'(w) = X$$

Therefore, the above equation becomes:

$$\frac{dD}{dw} = -y^* * \frac{1}{\sigma(w^T X+b)} * \sigma(w^T X + b) * (1 - \sigma(w^T X + b))) * X$$

$$+(1 - y^x) \frac{1}{1-\sigma(w^T X+b)} * \sigma(w^T X + b) * (1 - \sigma(w^T X + b)) * X$$

$$= -y^*(1 - \sigma(w^t x + b)) * X + (1 - y^*)(\sigma(w^t x + b)) * X$$

Notice that $\sigma(w^t x + b) = M(x)$, we have

$$\frac{dD}{dw} = -y^*(1 - M(x)) * X + (1 - y^*) M(x) * X$$

$$= X * (M(x) * y^* - y^* + M(x) - M(x) * y^*)$$

$$= X * (M(x) - y^*) = -X * (y^* - M(x))$$

This is the gradient.