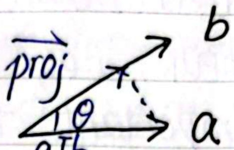
Perceptron

Key idea: Try to learn a hyperplane

设 $\vec{a}$ 为列向量 (by default)，$a^Tb = \sum_{d=1}^{D} a_d b_d$

L_2-norm: $\|\vec{a}\|_2 = \sqrt{a^Ta}$

则 $\vec{a}$ 在 $\vec{b}$ 上的投影为？



Δ: $\vec{a}\cdot\vec{b} = \|b\|\|a\|\cos\theta$

而 $a^Tb = \|b\|\cdot\|proj\|$，$\|proj\| = \frac{a^Tb}{\|b\|_2}$

故：$\vec{proj} = \frac{(a^Tb)}{\|b\|_2^2}\, b$，因为方向为 $\vec{b}$ 方向

在 2D 中，$w_1x_1+w_2x_2+b=0$ 为 line；3D 中为 plane

4+D 中 表示为 hyperplane，$w^Tx+b=0$，值得注意的是：

w 向量指向总为 $w^Tx+b>0$ 的部分！

因此，hyperplane creates 2 halfspaces.

Perceptron 算法：如何更新 w 与 b？

初始化：$w=\vec{0}$，$b=0$

input: $x^{(t)}$ $y^{(t)} = \{\pm1\}$，$\hat{y} = sign(w^Tx+b) = \begin{cases} +1 & \text{if } w^Tx+b \geq 0 \\ -1 & \text{otherwise} \end{cases}$

若 misclassify 一个 +1 例子 ($y^{(t)}=+1, \hat{y}=-1$)：

$w \leftarrow w + x^{(t)}$ $\quad b \leftarrow b+1$

若 misclassify 一个 -1 例子 ($y^{(t)}=-1, \hat{y}=+1$)：

$w \leftarrow w - x^{(t)}$ $\quad b \leftarrow b-1$

为何这样更新？ 若：$b+w^Tx<0$，$y^{(t)}=+1$，欲 $w^Tx\uparrow$

$w \leftarrow w+x^{(t)}$，$w^Tx = (w+x^{(t)})^Tx = wx + \underbrace{x^Tx}_{\geq 0}$，$\uparrow$

[反之亦然] 同时 b 也 $\uparrow$，$b+w^Tx\uparrow$

而更新逻辑可简化为：$\begin{cases} w \leftarrow w + y^{(t)}x^{(t)} \\ b \leftarrow b + y^{(t)} \end{cases}$

可见，w 可视为 $x^{(1)}, \ldots, x^{(n)}$ 的线性组合

同时还有一个trick: $\vec{x} = \begin{bmatrix}1\\ \vec{x}\end{bmatrix}$ $\vec{\theta} = \begin{bmatrix} b \\ w\end{bmatrix}$. 则:

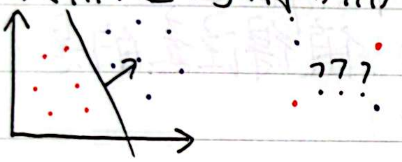
$\hat{y} = \mathrm{sign}(w^T x + b) = \mathrm{sign}(\theta^T x)$

$\theta \leftarrow \theta + y^{(t)} x'^{(t)}$, where $x'(t) = \begin{bmatrix}1\\ x^{(t)}\end{bmatrix}$

那么Perceptron的Inductive Bias是什么? Decision Boundary should be linear, i.e., hyperplane，且 Recent mistakes are more important than older ones.

而当θ.b训到不断产生正确预测时,我们就认为感知机收敛(converge)了。同时, Perceptron也可能受overfitting影响

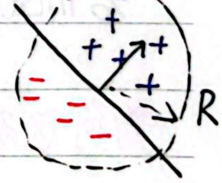
之前介绍到,学的其实是一个超平面;在binary分类中,如果两类点之间的分隔不是"面"而是"space"呢?



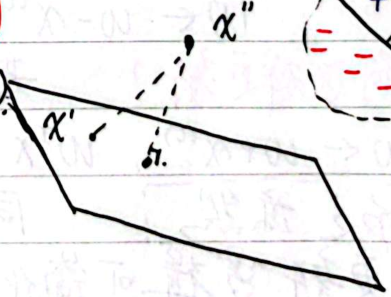
Def: 对二分类问题,样本集S is linearly seperable if there exists a linear boundary that seperate the points

"space"大小又引出下面定义: Def: The margin γ for dataset D is the greatest possible distance between a linear separator and the closest point in D to the linear separator

Theorem: 若Data margin为γ. 所有点都在半径为R的球内,则online perceptron算法 makes $\leq (R/\gamma)^2$ mistakes*



那么margin如何计算? 相当于是$(x''-x')$向量在 $w/\|w\|_2$ 上的投影长度(i.e.,内积)



则 $\left|\dfrac{w^T(x''-x')}{\|w\|_2}\right| = \dfrac{|w^T x'' - w^T x'|}{\|w\|_2}$

$= \dfrac{|w^T x'' + b|}{\|w\|_2}$