Binary classification

kernel : Data space $\times$ Data space $\to \mathbb{C}$

$\mathcal{X}$ $\quad\quad$ $\mathcal{X}$

(Hilbert space or $\mathbb{R}^n$)

- measures similarity in data

kernel methods : nonlinear generalizations of algorithms that have dot product forms

$\longrightarrow$ Bayes decision rule / classifier

$\longrightarrow$ Perceptron

Recap

Given Data: $x_1, \dots, x_n \in \mathcal{X}$

$y_1, \dots, y_n$

$y_i = f(x_i)$ (true function) target

Binary classification: $y_i = \pm 1$

Linear model:

$$h(x) = \mathrm{sgn}(w \cdot x + b)$$

learned classifier

$$x = [x^{(1)}, \dots, x^{(d)}, 1]$$

$$w = [w, b]$$

$$h(x) = \mathrm{sgn}(w \cdot x)$$

(simplification: affine $\to$ linear in $\mathbb{R}^{d+1}$)

Bayes classifier

based on prob. assumptions on data

- $y = \pm 1$ with eq. prob

$$P_Y(1) = P_Y(-1) = \frac{1}{2}$$

- $P_{X|Y}(x \mid Y=1)$, $P_{X|Y}(x \mid Y=-1)$

class conditional densities

$$P_{X|Y}(x_i \mid Y=1) = \begin{cases} 0 & y_i, \; f(x_i) = -1 \\ \frac{1}{n_+} & f(x_i) = y_i = 1 \end{cases}$$

$$S^+ = \{ i : i \in [n], f(x_i) = 1 \}$$

$$S^- = \{ i : i \in [n], f(x_i) = -1 \}$$

$n^+ = |S^+|$ $\quad n^- = |S^-|$

(# +ve pts)

$$P_{X|Y}(x_i \mid Y=-1) = \begin{cases} 0 & y_i = 1 \\ \frac{1}{n_-} & y_i = -1 \end{cases}$$

Bayes rule

$$h(x) = \begin{cases} 1 & P_{Y|X}(Y=1 \mid X=x) > P_{Y|X}(Y=-1 \mid X=x) \\ -1 & P_{Y|X}(Y=-1 \mid X=x) > P_{Y|X}(Y=1 \mid X=x) \end{cases}$$

$$P_{Y|X}(Y=1 \mid X=x) = \frac{P_Y(Y=1) \cdot \hat{P}_{X|Y}(X=x \mid Y=1)}{P_X(x)}$$

$$P_{Y|X}(Y=-1 \mid X=x) = \frac{P_Y(Y=-1) \cdot \hat{P}_{X|Y}(X=x \mid Y=-1)}{P_X(x)}$$

$$P_Y(Y=1) = P_Y(Y=-1) = \frac{1}{2}$$

$$h(x) = \begin{cases} 1 & P_{X|Y}(X=x \mid Y=1) > P_{X|Y}(X=x \mid Y=-1) \\ -1 & \text{o.w.} \end{cases}$$

Total probability

$$P_{X|Y}(X=x \mid Y=1) = \int \hat{P}_{X|Y,X'}(X=x \mid Y=1, X'=x') \, P_{X'|Y}(X'=x' \mid Y=1) \, dx'$$

$$= \int \mathcal{K}(x, x') \, P_{X|Y}(x' \mid Y=1) \, dx'$$

(expectation of $\mathcal{K}(x, \cdot)$ wrt $P_{X|Y}(\cdot \mid Y=1)$).

$$\approx \frac{1}{n_+} \sum_{i \in S_+} \mathcal{K}(x, x_i)$$

$$\approx \sum_{i=1}^{n} \mathcal{K}(x, x_i) \, \underline{P_{X|Y}(x_i \mid Y=1)}$$

$$h(x) = \begin{cases} 1 & \underbrace{\frac{1}{n_+} \sum_{i \in S_+} \mathcal{K}(x, x_i)}_{P_{X|Y}(x \mid Y=1)} > \underbrace{\frac{1}{n_-} \sum_{i \in S_-} \mathcal{K}(x, x_i)}_{P_{X|Y}(x \mid Y=-1)} \\ -1 & \text{o.w.} \end{cases}$$

"Geometric" point of view

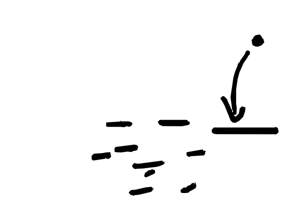
$$m_+ = \frac{1}{n_+} \sum_{i \in S_+} x_i \qquad m_- = \frac{1}{n_-} \sum_{i \in S_-} x_i$$

(mean) of +ve pts

$$h_S(x) = \begin{cases} 1 & d(x, m_+) < d(x, m_-) \\ -1 & \text{o.w.} \end{cases}$$

For $x \in \mathbb{R}^n$,

$x \to h_S(x)$ turns out to be a linear model



$$h_S(x) = \mathrm{sgn}\Big((m_+ - m_-) \cdot x + \underbrace{\tfrac{1}{2}\left(\|m_-\|^2 - \|m_+\|^2 \right)}_{b} \Big)$$

$$= \mathrm{sgn}\Big(\frac{1}{n_+} \sum_{i \in S_+} x_i \cdot x - \frac{1}{n_-} \sum_{i \in S_-} x_i \cdot x + b \Big)$$

$h_S(x)$ is in

- Dot product form

- Kernelizing: Replacing dot product with kernel evaluations.

$$h_S^{\mathcal{K}}(x) = \mathrm{sgn}\Big(\frac{1}{n_+} \sum_{i \in S_+} \mathcal{K}(x_i, x) - \frac{1}{n_-} \sum_{i \in S_-} \mathcal{K}(x_i, x) + \frac{1}{2}\Big(\frac{1}{n_-^2} \sum_{i,j \in S_-} \mathcal{K}(x_i, x_j) - \frac{1}{n_+^2} \sum_{i,j \in S_+} \mathcal{K}(x_i, x_j) \Big) \Big)$$

$$\|m^-\|^2 = m^- \cdot m^-$$

$$= \Big(\frac{1}{n_-} \sum_{i \in S_-} x_i \Big) \cdot \Big(\frac{1}{n_-} \sum_{i \in S_-} x_i \Big)$$

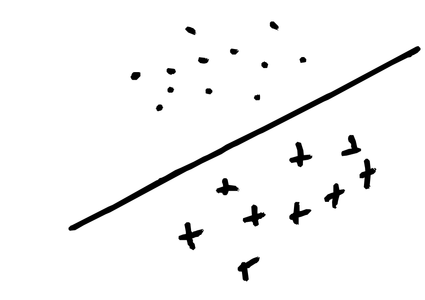
$$\to \frac{1}{(n_-)^2} \sum_{i,j \in S_-} \mathcal{K}(x_i, x_j)$$

$$h_{bayes}(x) = \begin{cases} 1 & \frac{1}{n_+} \sum_{i \in S^+} \mathcal{K}(x, x_i) > \frac{1}{n_-} \sum_{i \in S^-} \mathcal{K}(x, x_i) \\ -1 & \text{o.w.} \end{cases}$$

$$h_{bayes}(x) = \mathrm{sgn}\Big(\frac{1}{n_+} \sum_{i \in S^+} \mathcal{K}(x, x_i) - \frac{1}{n_-} \sum_{i \in S_-} \mathcal{K}(x, x_i) \Big)$$

- Mean-based geometric classifier $h_S^{\mathcal{K}}$ is nonlinear generalization of h_S (linear)

- $h_S^{\mathcal{K}}$ is also Bayes classifier when $b = 0$



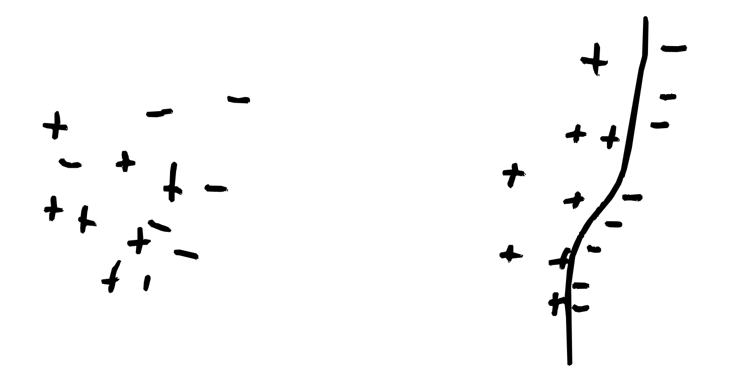
Linear separability

$y_i = f(x_i)$

if true function f has the form

$$f(x) = \mathrm{sgn}(w \cdot x + b)$$

for some $w \in \mathcal{X}$



Perceptron algorithm

$$h_p(x) = \mathrm{sgn}(w \cdot x)$$
(basic neuron)

$w_0 = 0 \in \mathcal{X}$

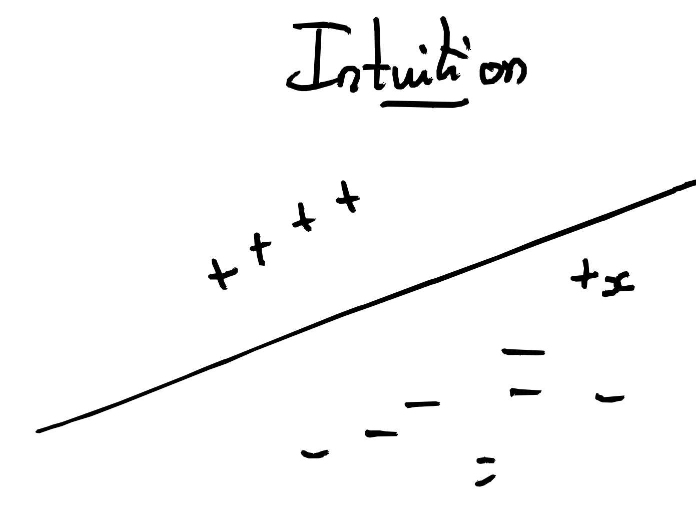
If i^{th} data point, x_i, is misclassified, then,

$$w_{i+1} = w_i + \underset{\text{learning rate}}{\eta}\, y_i x_i$$

o.w.

$$w_{i+1} = w_i$$

Intuition



$$\mathrm{sgn}(w \cdot x) \neq y$$

$$\mathrm{sgn}(y\, w \cdot x) = -1$$

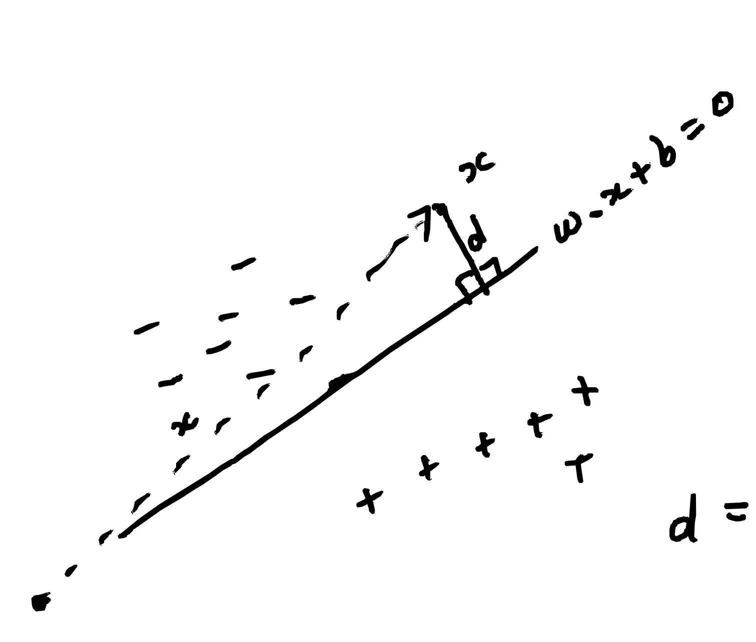
$$\underbrace{y_i\, w_{i+1} \cdot x_i}_{} = \underset{<0}{y_i\, w_i \cdot x_i} + \underset{>0}{\eta\, \|x_i\|^2}$$

Convergence proof

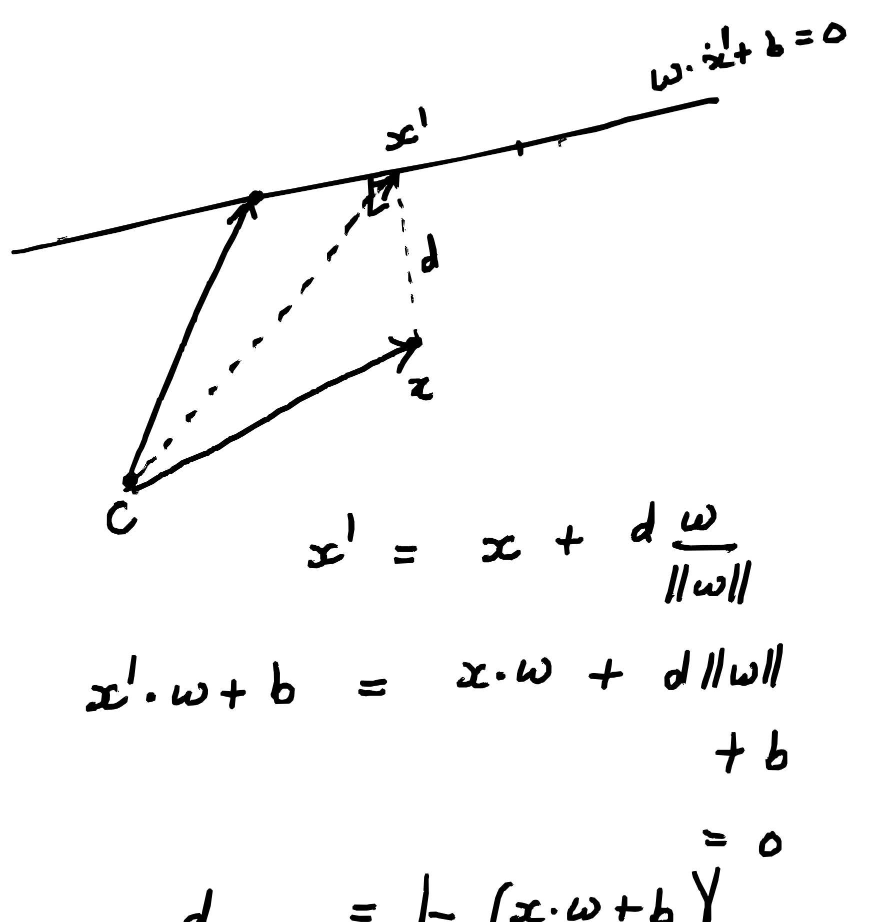
If $(x_i, y_i)_{i=1}^{n}$ are linearly separable, and $\|x_i\| < R \ \forall i$, then, the total number of updates, U, over n points satisfies (# of errors)

$$U \leq \frac{R^2}{\rho^2}$$

where $\rho = \min_i y_i \frac{w^* \cdot x_i}{\|w^*\|} > 0$ for some $w^* \in \mathcal{X}$.



$$d = \frac{|w \cdot x + b|}{\|w\|}$$



$$x' = x + d \frac{w}{\|w\|}$$

$$x' \cdot w + b = x \cdot w + d\|w\| + b = 0$$

$$d = \left| -\frac{(x \cdot w + b)}{\|w\|} \right|$$

Consequence of linear separability: $\exists\, w \in \mathcal{X}$ s.t.

$$\min_i \frac{|w \cdot x_i + b|}{\|w\|} > 0$$

when $b = 0$,

$\exists\, w \in \mathcal{X}$ s.t. $\min_i \frac{|w \cdot x_i|}{\|w\|} > 0$

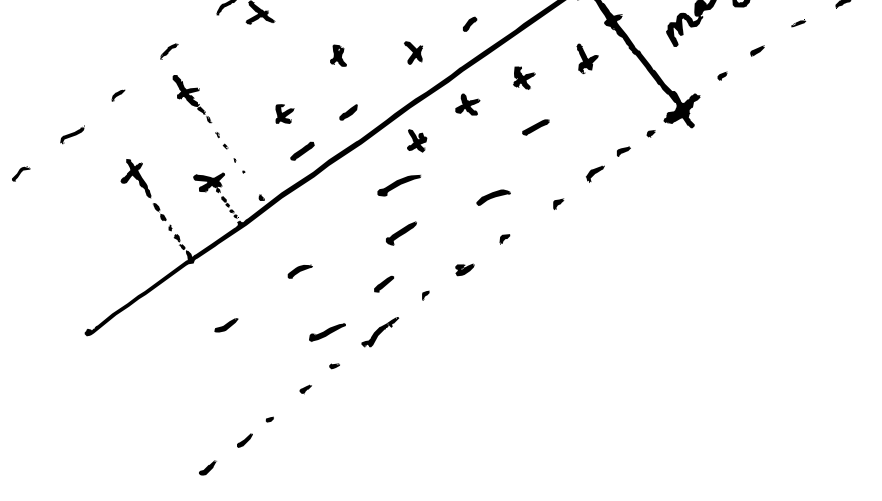
For a correctly classified pt,

$$y_i \frac{w \cdot x_i}{\|w\|} = \frac{|w \cdot x_i|}{\|w\|}$$
$$= \text{distance of } x_i \text{ from plane.}$$

Under linear separability,

$$\rho = \min_i \frac{|w \cdot x_i + b|}{\|w\|} = \min_i y_i \frac{w \cdot x_i}{\|w\|}$$

margin



$$\rho = \min_i \; y_i \frac{w \cdot x_i}{\|w\|}$$

$$U \rho \leq \left| \sum_{i \in I} y_i \frac{w_i \cdot x_i}{\|w_i\|} \right|$$

U: # of updates or # of errors

I: set of incorrect of indices

(Cauchy Schwarz)

$$\leq \left\| \sum_{i \in I} y_i x_i \right\|$$

$(w_{i+1} = w_i + \eta y_i x_i)$

$$\leq \left\| \frac{1}{\eta} \sum_{i \in I} (w_{i+1} - w_i) \right\|$$

$$= \frac{1}{\eta} \| w_{n+1} \|$$

$$= \frac{1}{\eta} \sqrt{\|w_{n+1}\|^2}$$

$$= \frac{1}{\eta} \sqrt{\sum_{i \in I} \left(\|w_{i+1}\|^2 - \|w_i\|^2 \right)}$$

$$\leq \frac{1}{\eta} \sqrt{\sum_{i \in I} \left(\|w_i + \eta x_i y_i\|^2 - \|w_i\|^2 \right)}$$

$$= \frac{1}{\eta} \sqrt{\sum_{i \in I} \left(\eta^2 \|x_i\|^2 + 2 w_i \cdot x_i y_i \right)}$$

$$\leq \sqrt{\sum_{i \in I} \|x_i\|^2}$$

$$\leq \sqrt{U R^2} = \sqrt{U} R$$

$$U \rho \leq \sqrt{U} R$$

$$U \leq \frac{R^2}{\rho^2}$$

iterations for conv : independent