Lecture 17: Convergence Proof of Perceptron Algo, Kernel perceptron, Logistic Regression (Begin)

Instructor: Prof. Ganesh Ramakrishnan

Recap: Perceptron Update Rule

- Perceptron works for two classes ($y = \pm 1$). A point is misclassified if $y\mathbf{w}^T(\phi(\mathbf{x})) < 0$ (error is on those x for which the "unsigned distance" turns out be "signed" !!
- Perceptron Algorithm:
  - INITIALIZE: w=ones()
  - REPEAT: for each $< \mathbf{x}, y >$
    - If $y\mathbf{w}^T\Phi(\mathbf{x}) < 0$
    - then, $\mathbf{w} = \mathbf{w} + \eta\phi(\mathbf{x}).y$
    - endif

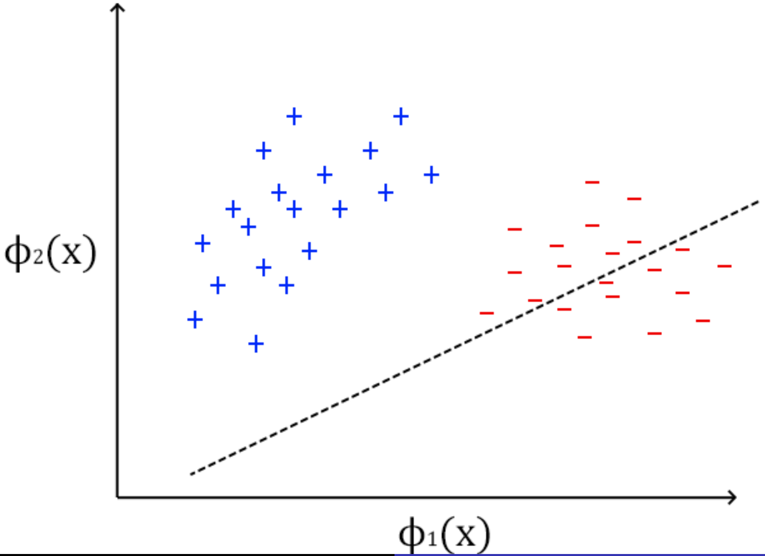
→ By design wont converge if dataset is NOT linearly seperable! In practice you run this for some max # iterations

- **Intuition:**

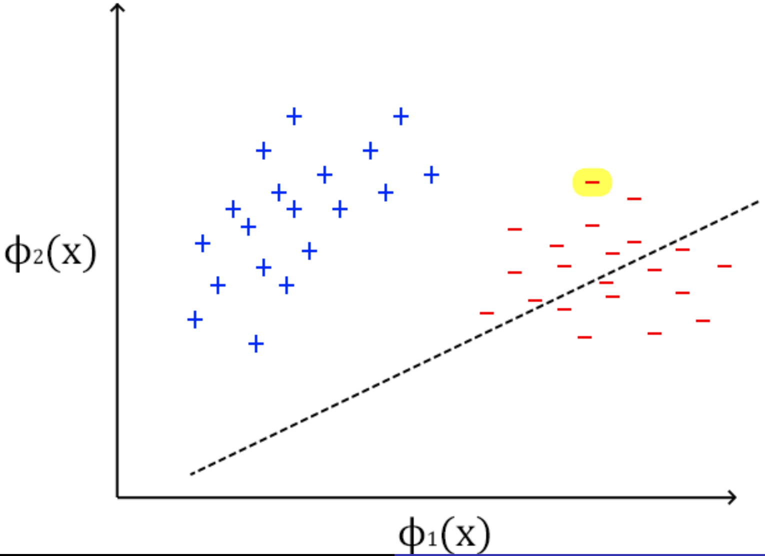
$$
\begin{aligned}
y(\mathbf{w}^{(k+1)})^T\phi(\mathbf{x}) &= y\left(\mathbf{w}^k + \eta y\phi^T(\mathbf{x})\right)\phi(\mathbf{x}) \\
&= y(\mathbf{w}^k)^T\phi(\mathbf{x}) + \eta y^2\|\phi(\mathbf{w})\|^2 \\
&> y(\mathbf{w}^k)^T\phi(\mathbf{x})
\end{aligned}
$$

unsigned distance becomes increasingly less neg. ←

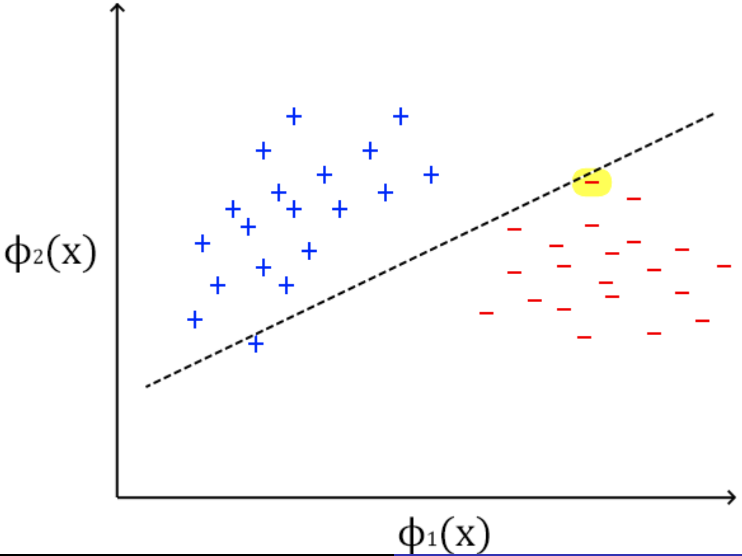
Since $y(\mathbf{w}^k)^T\phi(\mathbf{x}) \leq 0$, we have $y(\mathbf{w}^{(k+1)})^T\phi(\mathbf{x}) > y(\mathbf{w}^k)^T\phi(\mathbf{x}) \Rightarrow$ more hope that this point is classified correctly now.

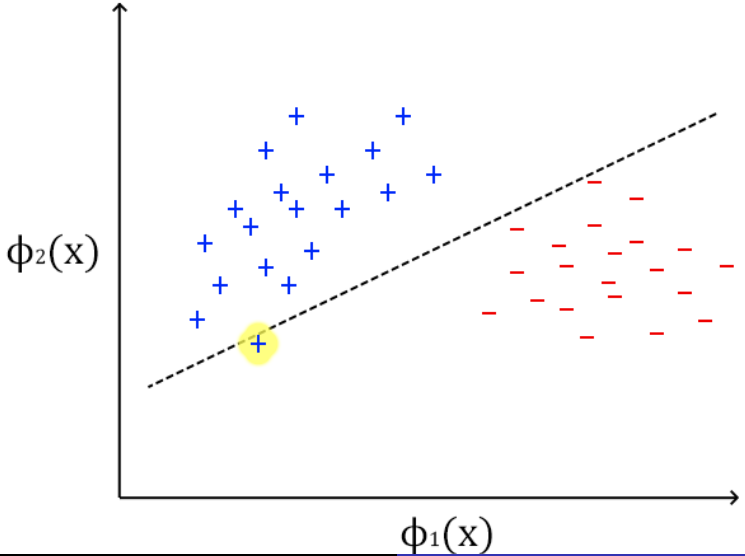


$\phi_2(x)$
$\phi_1(x)$



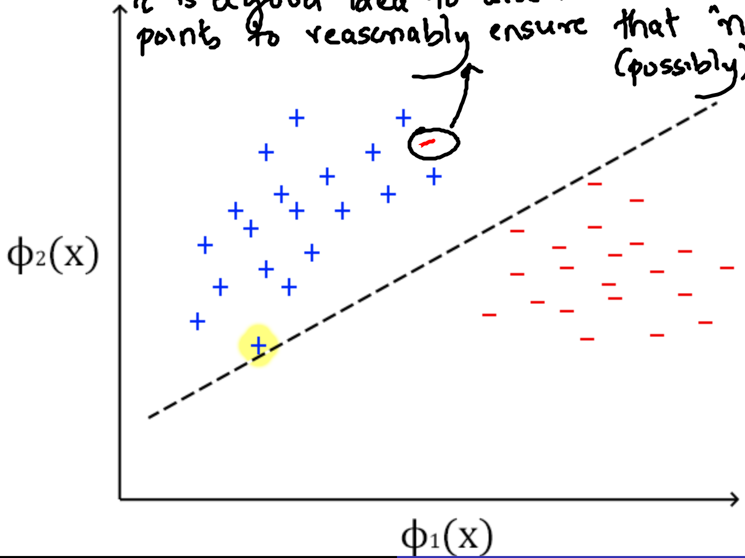
$\phi_2(x)$
$\phi_1(x)$


φ2(x)
φ1(x)



$\phi_2(x)$
$\phi_1(x)$

Especially in the presence of such noise, it is a good idea to alternate between + and − points to reasonably ensure that "noise" pts are (possibly) left alone.



Efficiently computing # of misclassified pts

requires

$$\mathbb{R}^m\left\{\underbrace{\begin{bmatrix} \phi(x^{(1)})\,y^{(1)} \\ \phi(x^{(2)})\,y^{(2)} \\ \phi(x^{(m)})\,y^{(m)} \end{bmatrix}}_{\mathbb{R}^n}\right. \left.\begin{bmatrix} w^k \\ \\ \\ \end{bmatrix}\right\}\mathbb{R}^n$$

$\#\text{multiplications} = mn$ → Can be empirically reduced using decompositions on ϕy matrix.

$\#\text{additions} = m(n-1)$ → Please use inbuilt libraries for matrix vector mult in scipy & numpy & tensorflow!

Convergence of Perceptron Algorithm

Perceptron Update Rule: Further analysis

(An intuitive sufficient condition)

- **Formally,**:- If $\exists$ an optimal separating hyperplane with parameters $\mathbf{w}^*$ such that,

$$\forall\ (\mathbf{x}, y),\ \ y\phi^T(\mathbf{x})\mathbf{w}^* \geq 0$$

then the perceptron algorithm converges.

Proof:- We want to show that

Atleast θ^2 (independent of k) reduction in each step!

$gap(k) = \|w^{(k)} - \rho w^*\|_2^2$

$gap(k+1) < gap(k) - \theta^2 \rightarrow$ 3

$$\lim_{k\to\infty} \|\mathbf{w}^{(k+1)} - \rho\mathbf{w}^*\|^2 = 0 \quad (1)$$

(If this happens for some constant ρ, we are fine.) independent of k

$\|w^{(k+1)} - \rho w^*\|_2^2 = \|w^{(k)} + y\phi(x) - \rho w^*\|_2^2 \rightarrow$ 1

In our discussion, let $\eta = 1$

Note: (x, y) is such that $y(w^{(k)})^T\phi(x) < 0 \rightarrow$ 2

Perceptron Update Rule: Further analysis

- **Formally,**:- If $\exists$ an optimal separating hyperplane with parameters $\mathbf{w}^*$ such that,

$$\forall\ (\mathbf{x}, y),\ \ y\phi^T(\mathbf{x})\mathbf{w}^* \geq 0$$

then the perceptron algorithm converges.

Proof:- We want to show that

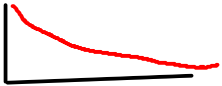
$$\lim_{k\to\infty} \|\mathbf{w}^{(k+1)} - \rho\mathbf{w}^*\|^2 = 0 \tag{1}$$

(If this happens for some constant ρ, we are fine.)

$$\underbrace{\|\mathbf{w}^{(k+1)} - \rho\mathbf{w}^*\|^2}_{gap(k+1)} = \underbrace{\|\mathbf{w}^k - \rho\mathbf{w}^*\|^2}_{gap(k)} + \boxed{\|y\phi(\mathbf{x})\|^2 + 2y(\mathbf{w}^k - \rho\mathbf{w}^*)^T\phi(\mathbf{x})} \tag{2}$$

Desire to be $\leq -\theta^2$

Substitute $w^{(k+1)} = w^{(k)} + y\phi(x)$



$1/k > 0$ but no constant bounds it

Perceptron Update Rule: Further analysis

- **Formally,**:- If $\exists$ an optimal separating hyperplane with parameters $\mathbf{w}^*$ such that,

$$\forall\ (\mathbf{x}, y),\ \ y\phi^T(\mathbf{x})\mathbf{w}^* \geq 0$$

then the perceptron algorithm converges.
Proof:- We want to show that

$$\lim_{k\to\infty} \|\mathbf{w}^{(k+1)} - \rho\mathbf{w}^*\|^2 = 0 \tag{1}$$

(If this happens for some constant ρ, we are fine.)

$$\|\mathbf{w}^{(k+1)} - \rho\mathbf{w}^*\|^2 = \|\mathbf{w}^k - \rho\mathbf{w}^*\|^2 + \underline{\|y\phi(\mathbf{x})\|^2 + 2y(\mathbf{w}^k - \rho\mathbf{w}^*)^T\phi(\mathbf{x})} \tag{2}$$

- For convergence of perceptron, we **need** L.H.S. to be less than R.H.S. at every step, although by some small but non-zero value (with $\theta \neq 0$)

$$\|\mathbf{w}^{(k+1)} - \rho\mathbf{w}^*\|^2 \leq \|\mathbf{w}^k - \rho\mathbf{w}^*\|^2 - \theta^2 \tag{3}$$

Perceptron Update Rule: Further analysis

- **Need** that $\|\mathbf{w}^{(k+1)} - \rho\mathbf{w}^*\|^2$ reduces by atleast θ^2 at every iteration.

we know $\rightarrow \|w^{(k+1)} - \delta w^*\|^2 = \|w^k - \delta w^*\|^2 + 2y(w^k - \delta w^*)^T\phi(x) + \|y\phi(x)\|^2$

we need $\rightarrow$

$$\|\mathbf{w}^{(k+1)} - \rho\mathbf{w}^*\|^2 \leq \|\mathbf{w}^k - \rho\mathbf{w}^*\|^2 - \theta^2 \quad (4)$$

- Based on (2) and (4), we **need** to find θ such that,

$$-\theta^2 \geq 2y(w^{(k)} - \delta w^*)^T\phi(x) + \|y\phi(x)\|^2$$

If I show such a θ^2 exists, I know (4) will hold!

Perceptron Update Rule: Further analysis

- **Need** that $\|\mathbf{w}^{(k+1)} - \rho\mathbf{w}^*\|^2$ reduces by atleast θ^2 at every iteration.

$$\|\mathbf{w}^{(k+1)} - \rho\mathbf{w}^*\|^2 \leq \|\mathbf{w}^k - \rho\mathbf{w}^*\|^2 - \theta^2 \tag{4}$$

- Based on (2) and (4), we **need** to find θ such that,

$$\|\phi(\mathbf{x})\|^2 + 2y(\mathbf{w}^k - \rho\mathbf{w}^*)^T\phi(\mathbf{x}) \leq -\theta^2$$

$(\|y\phi(\mathbf{x})\|^2 = \|\phi(\mathbf{x})\|^2$ since $y = \pm 1)$

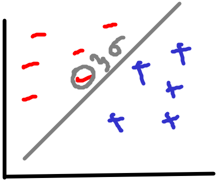
- The number of iterations would be: $O\left(\frac{\|\mathbf{w}^{(0)} - \rho\mathbf{w}^*\|^2}{\theta^2}\right)$
- Tutorial 6, Problem 4 has more concerning the number of iterations. But first we will discuss how convergence holds in the first place!

$O\left(\frac{\Gamma^2}{\theta^2}\right)$ independent of arbit δ & w^*

Perceptron Update Rule: Further analysis

$$\delta = -\min_{x \in \mathcal{D}} 2yw^{*T}\phi(x) = \max_{x \in \mathcal{D}} -2yw^{*T}\phi(x) \geq y'w^{*}\phi(x') \quad \forall\, x', y'$$

- **Observations**:-
  1. $y(\mathbf{w}^k)^T\phi(\mathbf{x}) < 0$ ($\because$ $\mathbf{x}$ was misclassified)
  2. $\Gamma^2 = \max_{\mathbf{x}\in\mathcal{D}} \|\phi(\mathbf{x})\|^2$
  3. $\delta = -\min_{\mathbf{x}\in\mathcal{D}} 2y{\mathbf{w}^*}^T\phi(\mathbf{x})$ (if σ = unsigned dist of closest pt in $\mathcal{D}$ to w^*, then $\delta = -2\sigma$)
- Here, negative margin $\delta = -2y{\mathbf{w}^*}^T\phi(\widehat{\mathbf{x}})$ is the negative of unsigned distance of closest point $\widehat{\mathbf{x}}$ from separating hyperplane :
  $\widehat{\mathbf{x}} = \underset{\mathbf{x}\in\mathcal{D}}{\operatorname{argmax}} -2y{\mathbf{w}^*}^T\phi(\mathbf{x}) = \underset{\mathbf{x}\in\mathcal{D}}{\arg\min}\, y{\mathbf{w}^*}^T\phi(x)$
- Since the data is linearly separable,



Margin = 2 * distance of closest point
(will revisit for support vector classification)

Perceptron Update Rule: Further analysis

- **Observations**:-
  1. $y(\mathbf{w}^k)^T\phi(\mathbf{x}) < 0$ ($\because$ $\mathbf{x}$ was misclassified)
  2. $\Gamma^2 = \max_{\mathbf{x}\in\mathcal{D}} \|\phi(\mathbf{x})\|^2$
  3. $\delta = -\min_{\mathbf{x}\in\mathcal{D}} 2y{\mathbf{w}^*}^T\phi(\mathbf{x}) \geq -2y'{w^*}^T\phi(x') \quad \forall x', y'$

$-\theta^2 \geq 2y(w^k - \rho w^*)^T\phi(x) + \|\phi(x)\|^2$

need ρ s.t. this is negative

- Here, negative margin $\delta = -2y{\mathbf{w}^*}^T\phi(\widehat{\mathbf{x}})$ is the negative of unsigned distance of closest point $\widehat{\mathbf{x}}$ from separating hyperplane :
$$\widehat{\mathbf{x}} = \operatorname*{argmax}_{\mathbf{x}\in\mathcal{D}} -2y{\mathbf{w}^*}^T\phi(\mathbf{x}) = \arg\min_{\mathbf{x}\in\mathcal{D}} y{\mathbf{w}^*}^T\phi(x)$$
- Since the data is linearly separable, $\widehat{y}{\mathbf{w}^*}^T\phi(\widehat{\mathbf{x}}) \geq 0$, so, $\delta \leq 0$. Consequently, we **need** from ρ that:

$$2y(w^k - \rho w^*)^T\phi(x) + \|\phi(x)\|^2 < -2\rho y(w^*)^T\phi(x) + \|\phi(x)\|^2$$
$$< \rho\delta + \Gamma^2 \leq 0$$

$y(w^k)^T\phi(x) < 0$

Perceptron Update Rule: Further analysis

- **Observations**:-
  1. $y(\mathbf{w}^k)^T\phi(\mathbf{x}) < 0$ ($\because$ $\mathbf{x}$ was misclassified)
  2. $\Gamma^2 = \max\limits_{\mathbf{x}\in\mathcal{D}} \|\phi(\mathbf{x})\|^2$
  3. $\delta = -\min\limits_{\mathbf{x}\in\mathcal{D}} 2y{\mathbf{w}^*}^T\phi(\mathbf{x})$
- Here, negative margin $\delta = -2y{\mathbf{w}^*}^T\phi(\widehat{\mathbf{x}})$ is the negative of unsigned distance of closest point $\widehat{\mathbf{x}}$ from separating hyperplane :
  $\widehat{\mathbf{x}} = \underset{\mathbf{x}\in\mathcal{D}}{\operatorname{argmax}} -2y{\mathbf{w}^*}^T\phi(\mathbf{x}) = \arg\min\limits_{\mathbf{x}\in\mathcal{D}} y{\mathbf{w}^*}^T\phi(x)$
- Since the data is linearly separable, $\widehat{y}{\mathbf{w}^*}^T\phi(\widehat{\mathbf{x}}) \geq 0$, so, $\delta \leq 0$. Consequently, we **need** from ρ that:

$$0 \leq \|\mathbf{w}^{(k+1)} - \rho\mathbf{w}^*\|^2 < \|\mathbf{w}^k - \rho\mathbf{w}^*\|^2 + \Gamma^2 + \rho\delta$$

Perceptron Update Rule: Further analysis

- Since, $\mathbf{w}^{*T}\phi(\hat{\mathbf{x}}) \geq 0$, so, $\delta \leq 0$. Consequently, restating what we **need** from ρ is:

$$0 \leq \|\mathbf{w}^{(k+1)} - \rho\mathbf{w}^*\|^2 < \|\mathbf{w}^k - \rho\mathbf{w}^*\|^2 + \underline{\Gamma^2 + \rho\delta}$$

to look like

$$0 \leq \|\mathbf{w}^{(k+1)} - \rho\mathbf{w}^*\|^2 < \|\mathbf{w}^k - \rho\mathbf{w}^*\|^2 - \theta^2$$

- Taking, $\rho =$

$$\rho\delta + \Gamma^2 \leq 0 \Rightarrow \rho\delta \leq -\Gamma^2 \Rightarrow \rho \geq \frac{-\Gamma^2}{\delta}$$

$(\because \delta < 0)$

$\rho = \frac{-2\Gamma^2}{\delta}$ is fine!

Perceptron Update Rule: Further analysis

- Since, $\mathbf{w^*}^T\phi(\widehat{\mathbf{x}}) \geq 0$, so, $\delta \leq 0$. Consequently, restating what we **need** from ρ is:

$$0 \leq \|\mathbf{w}^{(k+1)} - \rho\mathbf{w}^*\|^2 < \|\mathbf{w}^k - \rho\mathbf{w}^*\|^2 + \underbrace{\Gamma^2 + \rho\delta}$$

$$\rho = -\frac{2\Gamma^2}{\delta}$$

to look like

$$0 \leq \|\mathbf{w}^{(k+1)} - \rho\mathbf{w}^*\|^2 < \|\mathbf{w}^k - \rho\mathbf{w}^*\|^2 + \theta^2$$

- Taking, $\rho = \frac{2\Gamma^2}{-\delta}$, $0 \leq \|\mathbf{w}^{(k+1)} - \rho\mathbf{w}^*\|^2 \leq \|\mathbf{w}^k - \rho\mathbf{w}^*\|^2 - \Gamma^2$
- Thus, we get, $\Gamma^2 = \theta^2$, that we were looking for in eq.(3).
  $\therefore \|\mathbf{w}^{(k+1)} - \rho\mathbf{w}^*\|^2$ decreases by atleast Γ^2 at every iteration. $\rightarrow$ gap(k) - gap(k+1) $\geq \Gamma^2$
- Summarily: $\mathbf{w}^k$ converges to $\rho\mathbf{w}^*$ by making a minimum θ^2 decrement at each step.
- Thus, for $k \rightarrow \infty$, $\|\mathbf{w}^k - \rho\mathbf{w}^*\| \rightarrow 0$. This proves convergence.

Number of iterations: (Tutorial 6, Problem 4)

- A statement on number of iterations for convergence:
  If $||\mathbf{w}^*|| = 1$ and if there exists $\delta > 0$ such that for all $i = 1, \ldots, n$, $y_i(\mathbf{w}^*)^T\phi(\mathbf{x}_i) \geq \delta$ and $||\phi(\mathbf{x}_i)||^2 \leq \Gamma^2$ then the perceptron algorithm will make atmost $\frac{\Gamma^2}{\delta^2}$ errors (that is take atmost $\frac{\Gamma^2}{\delta^2}$ iterations to converge)