CS 307 Lecture 5

Linear Classification

Training Set $= \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$

$x^{(i)} \in \mathbb{R}^{d+1}$ ← d features + add^n 1

$y^{(i)} \in \{+1, -1\}$

Goal: Compute a hypothesis $h_\theta : \mathbb{R}^{d+1} \to \{+1, -1\}$ ($\theta \in \mathbb{R}^{d+1}$) such that

$$h_\theta(x) = \begin{cases} +1 & \text{if } x^T\theta > 0 \\ -1 & \text{if } x^T\theta \le 0. \end{cases}$$

$x^T\theta \to \theta_0 x_0 + \theta_1 x_1 + \dots \theta_d x_d$, $x_0 = 1$

and

$$J(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, x^{(i)}, y^{(i)})$$

is minimized

0-1 Loss function

$$\ell_{01}(\theta, x, y) = \begin{cases} 1 & \text{if } h_\theta(x) \neq y \\ 0 & \text{o.w.} \end{cases}$$

- Very difficult to compute (NP-hard)

Desire: A loss function that is "continuous" so that we can use SGD.

$\ell_h(\theta, x, y) = 1[y\theta^T x \le 0]$ — $[1(b) = \begin{cases} 0 & \text{if } b = \text{false} \\ 1 & \text{if } b = \text{true} \end{cases}$.

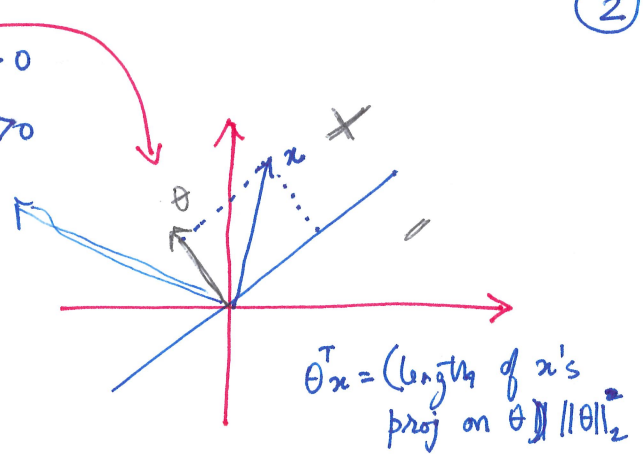
If $\ell_{01}(\theta, x, y) = 1$ then $\ell_h(\theta, x, y) = 1$

If $\theta^T x = 0$ then $\ell_h(\theta, x, y) = 1$ no matter what y is.

For x s.t. $\theta^T x \neq 0$, $\ell_{01}(\theta, x, y) = \ell_h(\theta, x, y)$.

$$\ell(\theta, x, y) = \begin{cases} -y\theta^T x & y\theta^T x \le 0 \\ 0 & y\theta^T x > 0 \end{cases}$$

$$\nabla_\theta \ell(\theta, x, y) = \begin{cases} -yx & y\theta^T x \le 0 \\ 0 & y\theta^T x > 0 \end{cases}$$



Perceptron: $\theta = 0$

Repeat until θ does not change

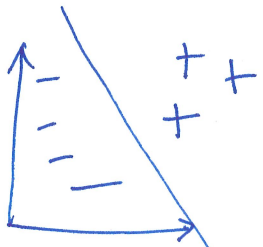
for $i \in \{1, 2, \ldots n\}$

$$\theta \leftarrow \theta - \alpha \frac{1}{n} \nabla_\theta \ell(\theta, x^{(i)}, y^{(i)})$$

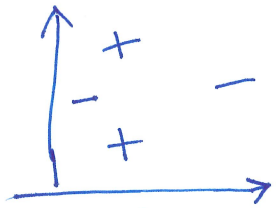
if $(y\theta^T x \le 0)$

$$\theta \leftarrow \theta + y^{(i)} x^{(i)}$$

Linear Separability: A training set is linearly separable if there is a hyperplane that separates the +1 and -1 examples.



Linearly separable



Not Linearly separable.

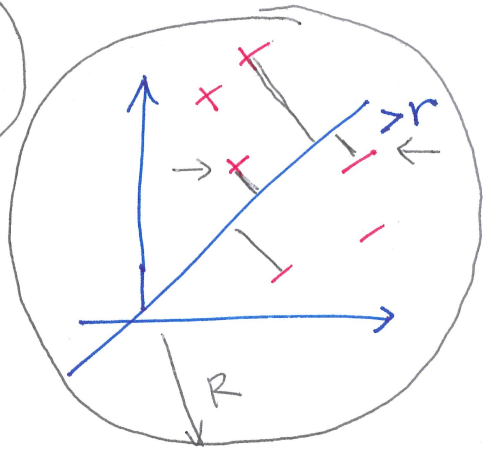
$\{(x^{(i)}, y^{(i)})\}_{i=1}^n$ is linearly separable if there is θ s.t. $y^{(i)}\theta^T x^{(i)} > 0 \quad \forall i \in \{1, 2, \ldots n\}$.

Signed distance of x w.r.t θ: $\dfrac{\theta^T x}{\|\theta\|_2}$

If θ is such that $y^{(i)}\theta^T x^{(i)} > 0 \quad \forall i \in \{1, 2 \ldots n\}$ then distance of $x^{(i)}$ from θ is $\dfrac{y^{(i)}\theta^T x^{(i)}}{\|\theta\|_2}$

Margin: Let θ be such that $y^{(i)}\theta^T x^{(i)} > 0$. The margin of θ w.r.t. $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$

$$= \min_{i \in \{1,2,..n\}} \frac{y^{(i)}\theta^T x^{(i)}}{\|\theta\|_2}$$



Perceptron Theorem: Suppose $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$ that is linearly separable, i.e., $\exists\ \theta$ s.t.

$$y^{(i)}\theta^T x^{(i)} > 0 \quad \forall i.$$

Let $\forall i$ $\dfrac{y^{(i)}\theta^T x^{(i)}}{\|\theta\|_2} > \gamma$.

Let us assume $R \geq \max_{i \in \{1,2,..n\}} \|x^{(i)}\|_2$

Then the number of times θ is changed in the perceptron algorithm is $\leq \left(\frac{R}{\gamma}\right)^2$.

Logistic Regression

$$l_{log}(\theta, x, y) = \underbrace{\ln}_{\text{natural log } \log_e} \left(1 + \underbrace{e}_{\text{euler's constant}}^{-y\theta^T x}\right)$$