CS 307 Lecture 13

Recap:

Convex set C: $\forall x, y \in C$, then $\{\alpha x + (1-\alpha) y \mid \alpha \in [0,1]\} \subseteq C$

Convex function $f: \mathbb{R}^d \to \mathbb{R}$: $\forall \alpha \in [0,1]$, $f(\alpha x + (1-\alpha) y) \leq \alpha f(x) + (1-\alpha) f(y)$

Theorem: Every local minimum of a convex function is a global minimum

— If gradient of f exists then we can use gradient descent to minimize f.

Convex Programming:
$$\min_{x \in C} f(x)$$

C ← Convex set; f → convex fn

There are efficient algorithms to solve a convex program.

$$\min f(x)$$
$$\text{s.t.} \quad C(x) \leftarrow \text{constraints on } x.$$

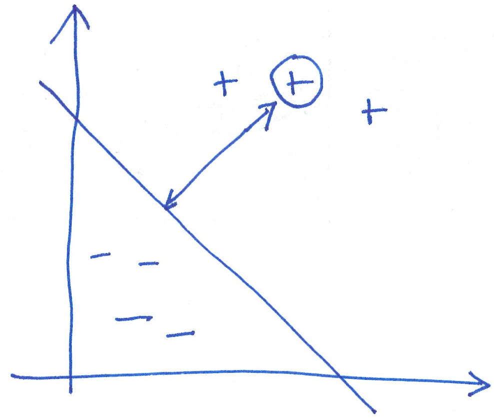
Linear Classification: Training set $S = \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$, $x^{(i)} \in \mathbb{R}^d$, $y^{(i)} \in \{-1, +1\}$.

Find (θ, θ_0) s.t.
$$h_\theta(x) = \begin{cases} +1 & \theta^T x + \theta_0 > 0 \\ -1 & \theta^T x + \theta_0 \leq 0 \end{cases}$$

θ_0 $\theta_1, \theta_2 \ldots \theta_d$

that minimizes $\cancel{J(\theta)}$ $J(\theta, \theta_0)$, $\theta \in \mathbb{R}^d$, $\theta_0 \in \mathbb{R}$

$$P[y=1 \mid x; \theta, \theta_0] = \frac{1}{1+e^{-(\underbrace{\theta^T x + \theta_0}_{\gg 0})}}$$

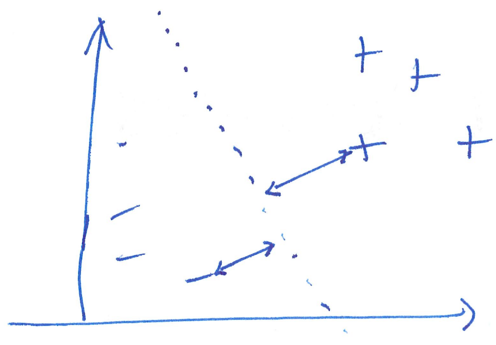


$$P[y=-1 \mid x; \theta, \theta_0] = 1 - \frac{1}{1+e^{-(\theta^T x + \theta_0)}}$$

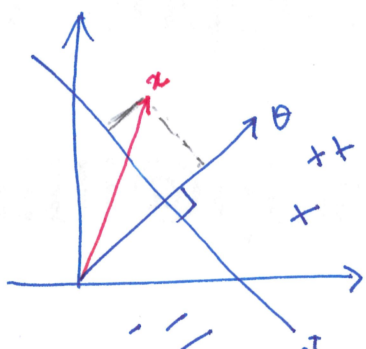
$$= \frac{e^{-(\theta^T x + \theta_0)}}{1+e^{-(\theta^T x + \theta_0)}}$$

$$= \frac{1}{e^{\underbrace{\theta^T x + \theta_0}_{\ll 0}} + 1}$$

Linearly Separable Training Set.



Distance of a point from hyperplane.



$$\frac{\theta^T x}{\|\theta\|} + \frac{\theta_0}{\|\theta\|_2}$$

$$= \underbrace{\frac{\theta^T x + \theta_0}{\|\theta\|_2}}_{\text{Signed distance}}$$

If $\theta^T x + \theta_0 = 0$ separates the traing set $S = \{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$.

then distance of $x^{(i)}$ from $\theta^T x + \theta_0 = 0$ is

$$\frac{y^{(i)}(\theta^T x^{(i)} + \theta_0)}{\|\theta\|_2}.$$

Goal: Find θ, θ_0 s.t. minimum distance of traing samples from $\theta^T x + \theta_0$ is maximized.

$$\max_{\theta, \theta_0} \min_{i \in \{1,..n\}} \frac{y^{(i)}(\theta^T x^{(i)} + \theta_0)}{\|\theta\|_2}$$

Formulation 1:

$$\left.\begin{aligned} &\max \ \gamma \\ &\text{s.t.} \ \forall i \in \{1,..n\} \ \frac{y^{(i)}(\theta^T x^{(i)} + \theta_0)}{\|\theta\|_2} \geq \gamma \end{aligned}\right\} \quad \begin{aligned} &\max \frac{\hat{\gamma}}{\|\theta\|} \\ &\text{s.t.} \\ &y^{(i)}(\theta^T x^{(i)} + \theta_0) \geq \hat{\gamma} \\ &\forall i \in \{1,..n\} \end{aligned}$$

Suppose $(\theta^T x^{(i)} + \theta_0) y^{(i)} \geq \hat{\gamma}$

Take $\hat{\theta} = \frac{\theta}{\hat{\gamma}}$, $\hat{\theta}_0 = \frac{\theta_0}{\hat{\gamma}}$ $\Rightarrow$ $\|\hat{\theta}\| = \frac{1}{\hat{\gamma}} \|\theta\|$.

$$y^{(i)}(\hat{\theta}^T x^{(i)} + \hat{\theta}_0) \geq 1$$

$$\left.\begin{aligned} &\max \frac{1}{\|\hat{\theta}\|} \\ &\text{s.t} \quad y^{(i)}(\hat{\theta}^T x^{(i)} + \hat{\theta}_0) \geq 1 \end{aligned}\right\}$$

Formulation 2:

$$\max \frac{1}{\|\theta\|}$$

$$\text{s.t} \ \forall i \in \{1,..n\} \quad y^{(i)}(\theta^T x^{(i)} + \theta_0) \geq 1 \ \left\{ \begin{aligned} &y^{(1)}(\theta^T x^{(1)} + \theta_0) \geq 1 \\ &y^{(2)}(\theta^T x^{(2)} + \theta_0) \geq 1 \\ &\vdots \\ &y^{(n)}(\theta^T x^{(n)} + \theta_0) \geq 1 \end{aligned} \right\} \text{Convex set.}$$

Standard Form for SVM:

$$\min \|\theta\|^2 \quad \} \rightarrow \min \theta^T\theta \text{ (convex fn)}$$

s.t $\forall i \in \{1..n\}$ $\quad y^{(i)}(\theta^T x^{(i)} + \theta_0) \geq 1$

Hard SVM (since it assumes training set is separable)

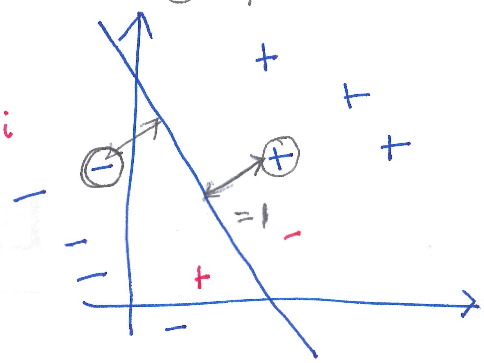
Soft SVM:

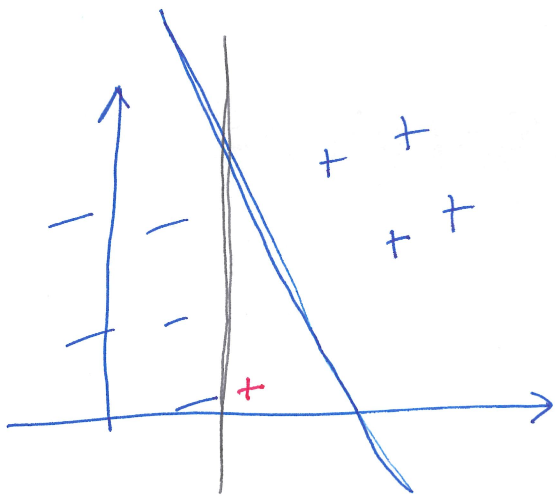
$$\min \|\theta\|^2 + C \sum_{i=1}^{n} \xi_i$$

hyperparameter

s.t $\quad y^{(i)}(\theta^T x^{(i)} + \theta_0) \geq \underline{1} - \xi_i$

$$\xi_i \geq 0$$





$$\xi_i \geq 1 - y^{(i)}(\theta^T x^{(i)} + \theta_0)$$

$$\xi_i \geq 0.$$

Optimal solution: $\xi_i = \max\{0, 1 - y^{(i)}(\theta^T x^{(i)} + \theta_0)\}$.

$$l_{hinge}(\theta, \theta_0, x, y) = \max(0, \underbrace{1 - y(\theta^T x + \theta_0)}_{})$$

perceptron: $\max(0, -y(\theta^T x + \theta_0))$

Soft SVM:

$$\min \|\theta\|^2 + C \sum_{i=1}^{n} l_{hinge}(x^{(i)}, y^{(i)}, \theta, \theta_0)$$

$$\min J_{softSVM}(\theta, \theta_0) = \frac{\lambda}{2}\|\theta\|^2 + \sum_{i=1}^{n} l_{hinge}(x^{(i)}, y^{(i)}, \theta, \theta_0)$$

$$J_{softSVM}(\theta, \theta_0) = \frac{\lambda}{2}\sum_{i=1}^{d}\theta_i^2 + \sum_{i=1}^{n} \max\Big(0, 1 - y^{(i)}\big(\underbrace{\theta^T x^{(i)}}_{\sum_{j=1}^{d}\theta_j x_j^{(i)}} + \theta_0\big)\Big)$$

$j \neq 0$
$$\frac{\partial}{\partial \theta_j} J_{softSVM}(\theta, \theta_0) = \lambda\theta_j + \sum_{i:\, y^{(i)}(\theta^T x^{(i)} + \theta_0) < 1} -y^{(i)} x_j^{(i)}$$

$$\frac{\partial}{\partial \theta_0} J_{softSVM}(\theta, \theta_0) = \sum_{i:\, y^{(i)}(\theta^T x^{(i)} + \theta_0) < 1} -y^{(i)}.$$