SVM Optimization Objective:

May 24, 2021

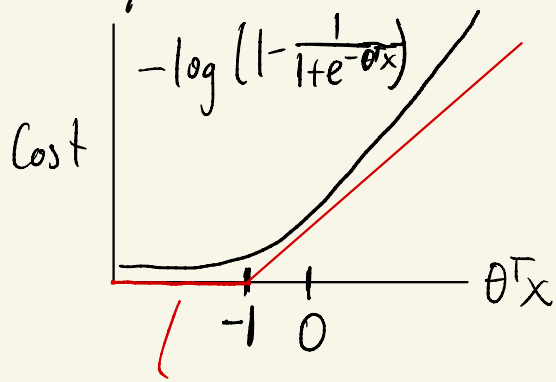
$$J(\theta) = -(y \log h_\theta(x) + (1-y)\log(1-h_\theta(x)))$$

$$= -y \log \frac{1}{1+e^{-\theta^T x}} - (1-y)\log\left(1 - \frac{1}{1+e^{-\theta^T x}}\right)$$

If $y=1$ (want $\theta^T x >> 0$):

let's call $\text{cost}_1(z)$

If $y=0$ (want $\theta^T x << 0$):



$\text{cost}_0(z)$

Logistic Regression:

$$\min_\theta \frac{1}{m}\left[\sum_{i=1}^{m} y^{(i)} \underbrace{(-\log h_\theta(x^{(i)}))}_{\text{cost}_1(\theta^T x^{(i)})} + (1-y^{(i)}) \underbrace{\left(-\log(1-h_\theta(x^{(i)}))\right)}_{\text{cost}_0(\theta^T x^{(i)})}\right]$$

$$+ \frac{\lambda}{2m}\sum_{j=1}^{n} \theta_j^2$$

Support Vector Machine:

$$\min_\theta \frac{1}{m} \sum_{i=1}^{m} y^{(i)} \text{cost}_1(\theta^T x^{(i)}) + (1-y^{(i)}) \text{cost}_0(\theta^T x^{(i)})$$

$$+ \frac{\lambda}{2m} \sum_{j=1}^{n} \theta_j^2$$

↳ doesn't change anything

$$A + \lambda B$$

↳ $CA + B \rightarrow C \approx \frac{1}{\lambda}$

$$\therefore \min_\theta C \sum_{i=1}^{m} \left[y^{(i)} \text{cost}_1(\theta^T x^{(i)}) + (1-y^{(i)}) \text{cost}_0(\theta^T x^{(i)}) \right] + \frac{1}{2} \sum_{j=1}^{n} \theta_j^2$$

↳ The objective cost function for SVMs.

Hypothesis Function for SVMs:

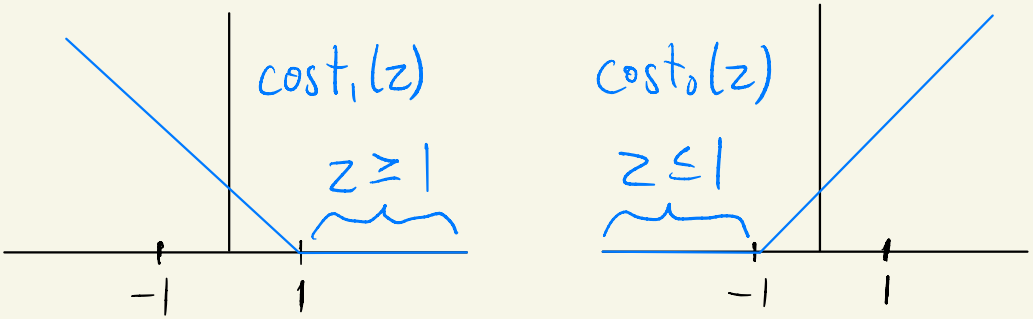
$$h_\theta(x) = \begin{cases} 1 & \text{if } \theta^T x \geq 0 \\ 0 & \text{if } \theta^T x < 0 \end{cases}$$

$$\text{cost}(\theta^T x, y) = \begin{cases} \max(0, 1-\theta^T x) & \text{if } y = 1 \\ \max(0, 1+\theta^T x) & \text{if } y = 0 \end{cases}$$

SVM Decision Boundary

$$\min_\theta C \sum_{i=1}^{m} \left[y^{(i)} \text{cost}_1(\theta^T x^{(i)}) + (1-y^{(i)}) \text{cost}_0(\theta^T x^{(i)}) \right] + \frac{1}{2} \sum_{j=1}^{n} \theta_j^2$$

We want this to be 0



If $y=1$, we want $\theta^T x \geq 1$.

↳ Then $\text{cost}_1(\theta^T x) = 0$ and cost is minimized.

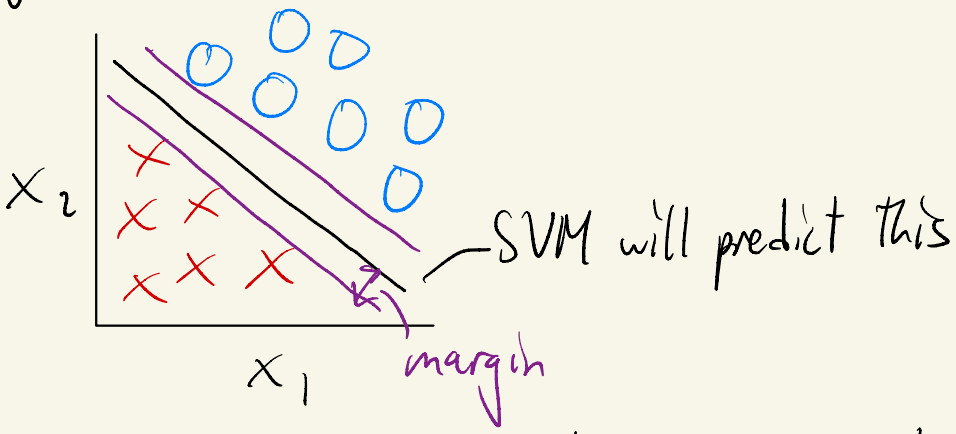
If $y=0$, we want $\theta^T x \leq -1$.

↳ Then $\text{cost}_0(\theta^T x) = 0$ and cost is minized.

Whenever $y^{(i)}=1$: $\theta^T x^{(i)} \geq 1$

Whenever $y^{(i)}=0$: $\theta^T x^{(i)} \leq -1$

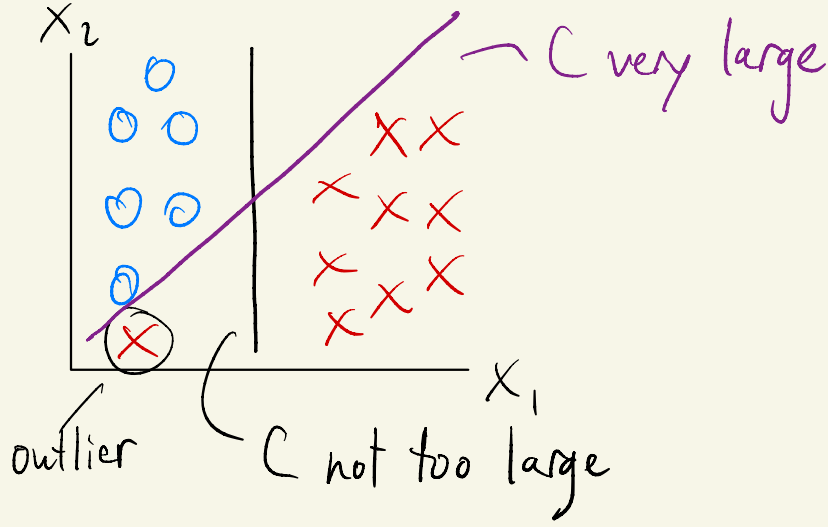
e.g. Linearly separable case



* SVM will try to separate the data w/ as large a margin as possible.

∴ SVMs are a.k.a. large margin classifiers.

<u>However, with Outliers:</u>



So why SVM over Logistic Regression?

SVMs try to find the best margin that separates two classes, which reduces the risk of error on the data.

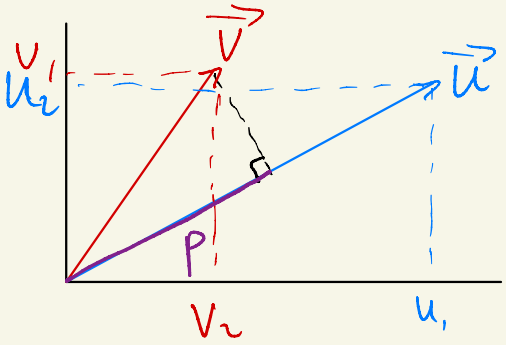
→ Works well w/ unstructured data like text and images.

→ Less prone to overfitting.

→ Uses a geometric model instead of a statistical model.

* But both work similarly in practice.

Vector Inner Product



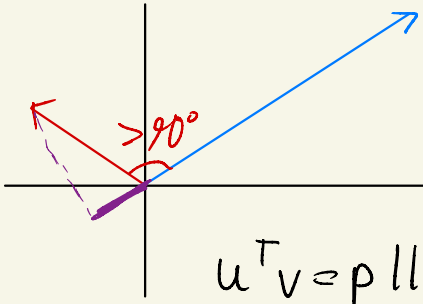
$$\vec{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \quad \vec{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

$$\|\vec{u}\| = \sqrt{u_1^2 + u_2^2} \in \mathbb{R}$$

$$p = \text{proj}_{\vec{u}} \vec{v}$$

$$u^T v = p\,\|\vec{u}\| = v^T u$$

$$= u_1 v_1 + u_2 v_2$$



$u^T v = p\,\|u\|$

$p < 0$

SVM Decision Boundary:

Simplify: $\theta_0 = 0$, $n = 2$ (i.e. x_1, x_2)

Goal: $\min_\theta \frac{1}{2} \sum_{j=1}^{n} \theta_j^2 = \frac{1}{2}(\theta_1^2 + \theta_2^2)$

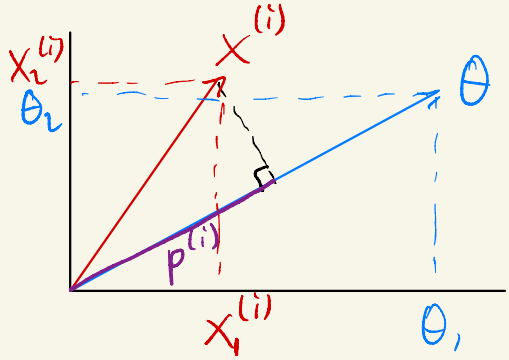
$$= \frac{1}{2}\left(\sqrt{\theta_1^2 + \theta_2^2}\right)^2$$

$$= \frac{1}{2}\|\theta\|^2$$

s.t. $p^{(i)}\|\theta\| \geq 1$ if $y^{(i)} = 1$

$p^{(i)}\|\theta\| \leq -1$ if $y^{(i)} = 0$

↳ $\theta = \begin{bmatrix} \theta_0 = 0 \\ \theta_1 \\ \theta_2 \end{bmatrix}$



$$\theta^T x^{(i)} = p^{(i)}\|\theta\| \quad \& \; p^{(i)} \in \mathbb{R}$$

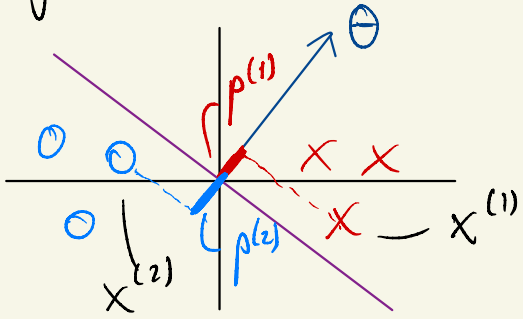
$$= \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)}$$

Goal: $\min_\theta \frac{1}{2}\sum_{j=1}^{n} \theta_j^2 = \frac{1}{2}(\theta_1^2 + \theta_2^2) = \frac{1}{2}\|\theta\|^2$

s.t. $p^{(i)}\|\theta\| \geq 1$ if $y^{(i)}=1$

$p^{(i)}\|\theta\| \leq -1$ if $y^{(i)}=0$

e.g.



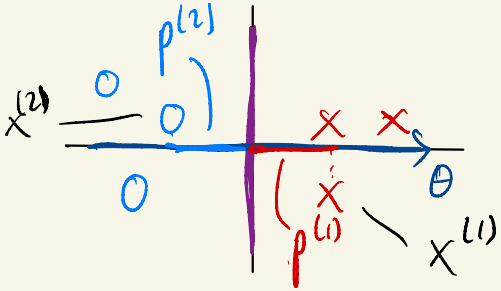
$p^{(1)}\|\theta\| \geq 1$

↳ $\|\theta\|$ large

$p^{(2)}\|\theta\| \leq -1$

↳ $\|\theta\|$ large

✡But we want $\|\theta\|$ small



$p^{(1)}\|\theta\| \geq 1$

↳ $\|\theta\|$ small

$p^{(2)}\|\theta\| \leq -1$

↳ $\|\theta\|$ small

✡Maximize p to minimize $\|\theta\|$ and maximize the margin.

✡You can see how the projection lengths are larger for a better margin.