CSE 575: Statistical Machine Learning (Spring 2021)
Instructor: Nupur Thakur

Linear Machines & SVM

Arizona State University

Table of contents

SVM - Problem Formulation

$$d = \frac{2}{||w||}$$

$$\{\mathbf{w}^*, b^*\} = \underset{\mathbf{w},b}{\operatorname{argmin}} ||\mathbf{w}|| \quad or \quad \{\mathbf{w}^*, b^*\} = \underset{\mathbf{w},b}{\operatorname{argmin}} \tfrac{1}{2}||\mathbf{w}||^2$$

Subject to

$$\mathbf{w}^t\mathbf{x}^{(i)} + b \geq 1 \quad \text{for } y^{(i)} = +1$$

$$\mathbf{w}^t\mathbf{x}^{(i)} + b \leq -1 \quad \text{for } y^{(i)} = -1$$

The constraints can be combined into:

$$\underline{y^{(i)}(\mathbf{w}^t\mathbf{x}^{(i)} + b) - 1} \geq 0 \quad \forall i$$

SVM Lagrangian Dual Formulation

$$L(w, b, \alpha) = \frac{1}{2}\underbrace{\|w\|^2}_{}^{w^Tw} - \sum_{i=1}^{n} \alpha_i \left[y^{(i)} (w^T x^{(i)} + b) - 1 \right] \quad — (1)$$

Lagrangian multipliers

$$\frac{\partial L}{\partial w} = w - \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)} \Rightarrow w = \sum_{i=1}^{n} \alpha_i y^{(i)} x^{(i)} \quad — (2)$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{n} \alpha_i y_i \Rightarrow \sum_{i=1}^{n} \alpha_i y_i = 0 \quad — (3)$$

Plugging (2) in (1),

$$L(w, b, \alpha) = \frac{1}{2} \sum_{i=1}^{n} \alpha_i y^{(i)} (x^{(i)})^T \sum_{j=1}^{n} \alpha_j y^{(j)} (x^{(j)})$$

For more on Lagrangian Formulation, refer (from page 11): http://cs229.stanford.edu/notes2020fall/notes2020fall/cs229-notes3.pdf

SVM Lagrangian Dual Formulation

$$- \sum_{i=1}^{n} \alpha_i y^{(i)} (x^{(i)})^T \sum_{j=1}^{n} \alpha_j y^{(j)} x^{(j)} - \underline{\underline{\sum_{i=1}^{n} \alpha_i y^{(i)}}}\, b = 0$$

$$+ \sum_{i=1}^{n} \alpha_i .$$

$$= -\frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} \underbrace{(x^{(i)})^T x^{(j)}} + \sum_{i=1}^{n} \alpha_i .$$

$$\boxed{L(w, b, \alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} \underbrace{\langle x^{(i)}, x^{(j)} \rangle}}$$

For more on Lagrangian Formulation, refer (from page 11): http://cs229.stanford.edu/notes2020fall/notes2020fall/cs229-notes3.pdf

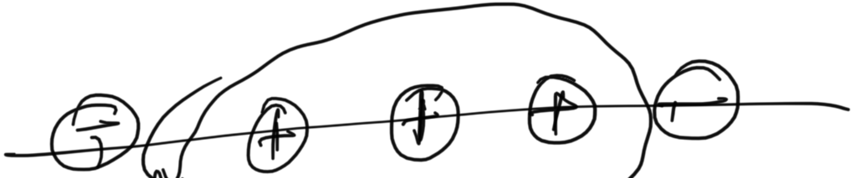
$$\begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = x_1 y_1 + x_2 y_2 + x_3 y_3$$
$$= \bar{x} \cdot \bar{y} \cdot \cos\theta$$

What if data is not linearly separable?

$2 \rightarrow 4$

$4 \rightarrow 16$

$8 \rightarrow 64$



$\phi(x)$ – Transformation on x

$$\boxed{L(w, b, \alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} \underbrace{\langle \phi(x^{(i)}), \phi(x^{(j)}) \rangle}_{\text{similarity between data points}}}$$

$K(x^{(i)}, x^{(j)})$

similarity between data points

What if data is not linearly separable?

- Kernel trick!

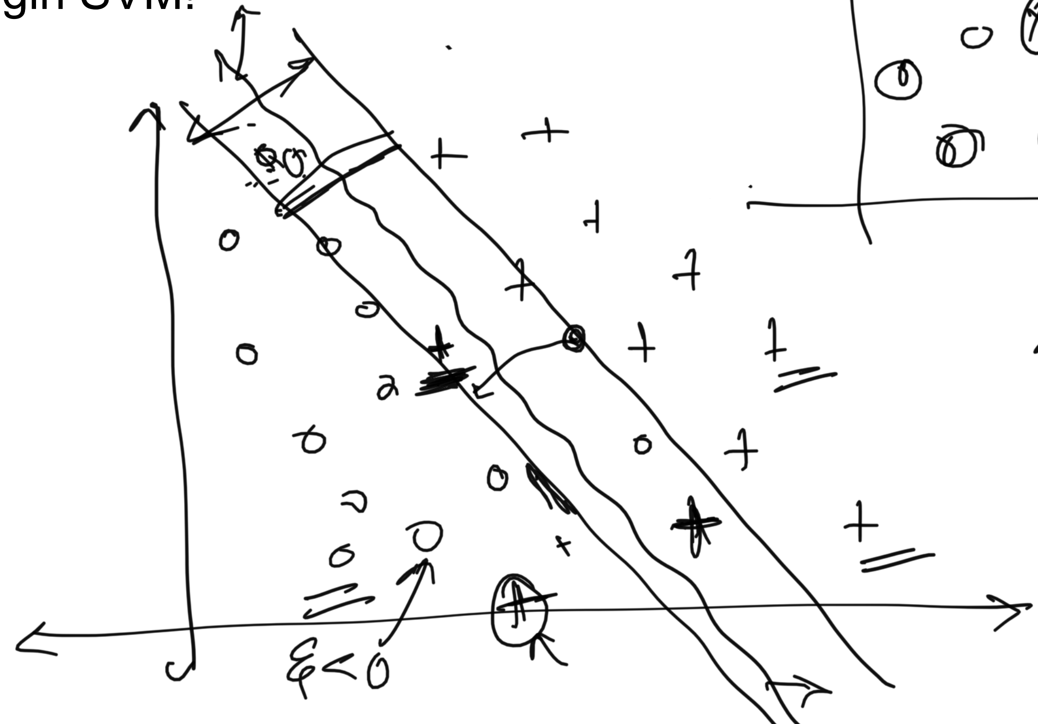
$$L(w, b, \alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} \underline{\underline{K(x^{(i)}, x^{(j)})}}$$

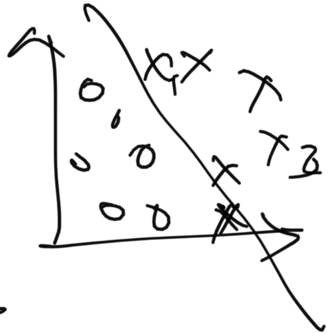
↳ Reduces computation

↳ Don't have to find the transformation ϕ

What if the data is still not separable?

- Soft-margin SVM!





Soft-Margin SVM Formulation

Hard-margin SVM

$$\boxed{\{\mathbf{w}^*, b^*\} = \underset{\mathbf{w},b}{\operatorname{argmin}} \frac{1}{2}||\mathbf{w}||^2} + C(\Sigma_i \xi_i)$$

subject to

$$w^t x^{(i)} + b \geq 1 - \xi_i \quad \text{for } y^{(i)} = +1$$

$$w^t x^{(i)} + b \leq -1 + \xi_i \quad \text{for } y^{(i)} = -1$$

$$\xi_i \geq 0, \forall i$$

$\xi = 0.8$

$\xi_i = 0$ – Correctly classified samples

$0 < \xi < 1$ – On the correct side of decision bound any but within the margin

$\xi > 1$ – completely misclassified samples.

$\xi = 0$ (Support vectors)

Effect of C parameter

C ⇒ Infinity ⇒ Hard-margin SVM

C = 0 ⇒ Decision boundary with very high misclassification.

Will SVM overfit?

↳ Robust to overfitting.

How to overcome overfitting?

- Tune C parameter.
- Tune the parameters used in kernels.
- Change the kernels

$$y = \frac{1}{1+e^{w^T x}}$$

$0-1$

Difference between SVM and Logistic Regression

Support Vector Machines	Logistic Regression
1. Margin, no probabilities	1. Probabilities
2. Sparse solution	2. Not sparse
3. Robust to overfitting	3. Can overfit
4. Hinge loss	4. Log likelihood loss

Questions?