Foundations of Machine Learning
AI2000 and AI5000

FoML-33
Support Vector Machines

Dr. Konda Reddy Mopuri
Department of AI, IIT Hyderabad
July-Nov 2025

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

DiL
Data-driven Intelligence
& Learning Lab

So far in FoML

- Intro to ML and Probability refresher
- MLE, MAP, and fully Bayesian treatment
- Supervised learning
  a. Linear Regression with basis functions
  b. Bias-Variance Decomposition
  c. Decision Theory - three broad classification strategies
  d. Neural Networks
- Unsupervised learning
  a. K-Means, Hierarchical, and GMM for clustering
- Kernelizing linear Models
  a. Dual representation, Kernel trick

For today

- SVM

Support Vector Machines

- Kernel method with sparse a solution
    - Inference needs kernel function values only at a subset of training data

Support Vector Machines

- Solution for a convex optimization problem
- Applications
  - Classification
  - Regression
  - Anomaly detection

SVM for binary classification

SVM for binary classification

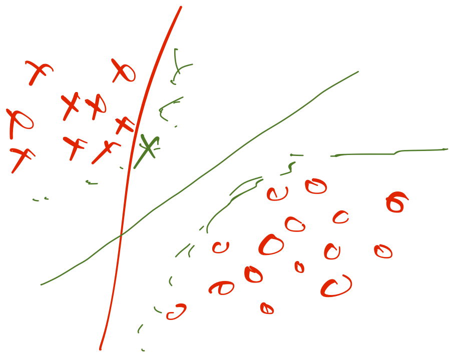
- Setting: linearly separable data with two labels {-1, 1}
- Model: linear model with fixed basis functions

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)^T \quad \mathbb{R}^D$$

$$\mathbf{t} = (t_1, t_2, \ldots, t_N)^T$$

$t_i = +1$ or -1

$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}) + b$$



SVM for binary classification

- $\exists$ at least one choice of model parameters to classify the training data correctly $\{\underline{w}, b\}$

Recap:Perceptron for linearly separable data

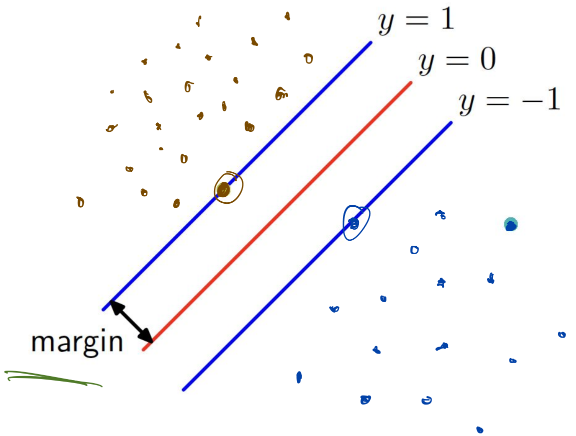
- Finds a solution in finite steps
- One of infinite solutions
- May not be best (in some sense)

Which one to pick?

- Among the multiple solutions
- We must pick the one that generalizes well
  - how?

Margin

- SVM approach through 'margin'
- Smallest distance between the decision boundary and any of the samples

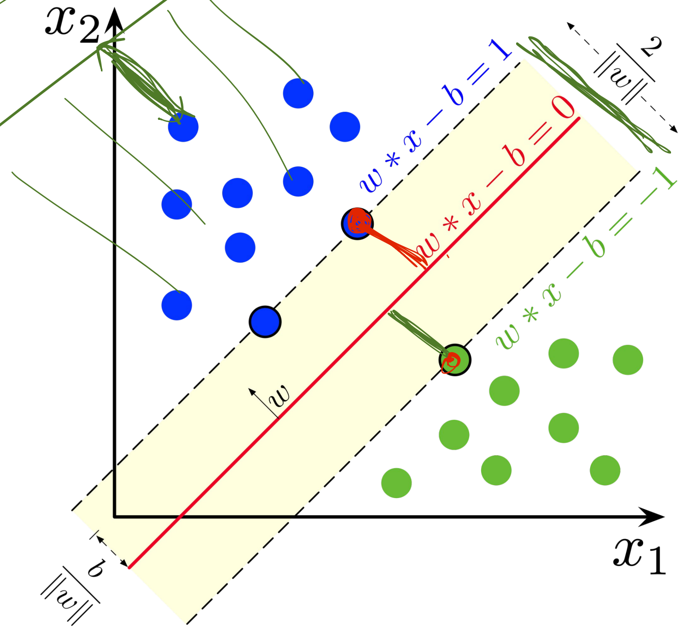


Margin

- Decision boundary is chosen to maximize the margin

$(w^T x + b) >> 0$

$<< 0$



SVM for binary classification

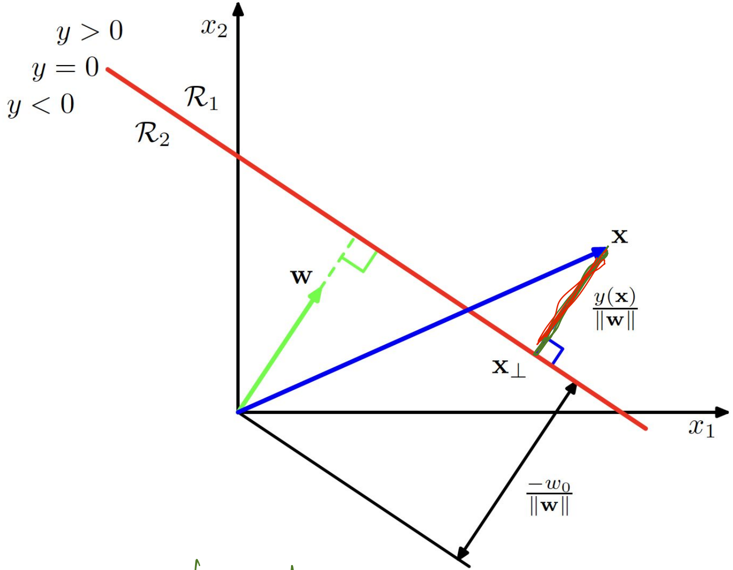
$\underline{a \cdot b}$ $\underline{a \cdot b}$

$\perp$ distance to the Solution

$$\frac{t_n y(\mathbf{x}_n)}{\|\mathbf{w}\|} = \frac{t_n(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n) + b)}{\|\mathbf{w}\|}.$$

$$\arg\max_{\mathbf{w},b} \left\{ \frac{1}{\|\mathbf{w}\|} \min_n \left[t_n \left(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n) + b \right) \right] \right\}$$

$(\alpha \underline{w} \quad \alpha b) \leftarrow (\underline{w}, b)$



SVM for binary classification

$$\arg\max_{\mathbf{w},b}\left\{\frac{1}{\|\mathbf{w}\|}\min_n\left[t_n\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)+b\right)\right]\right\}$$

Need to consider the scaling invariance

-1 $+1$ >0 <0

$$t_n\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}_n)+b\right) \geqslant 1, \qquad n=1,\ldots,N.$$

→ Scale invariance and correct classification of training data

$\|10(w_1^T)x^* + b_1 10\| \neq 0$

$w_2 = 10 w_1$

$b_2 = 10 b_1$

& restricts the set $\underline{w}$ to be the correct classifiers

SVM for binary classification

$$\arg\max_{\mathbf{w},b} \left\{ \frac{1}{\|\mathbf{w}\|} \overbrace{\min_n \left[t_n \left(\mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}_n) + b \right) \right]}^{1} \right\} \Rightarrow \arg\min_{\mathbf{w},b} \frac{1}{2} \|\mathbf{w}\|^2$$

$$t_n \left(\mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}_n) + b \right) \geqslant 1, \qquad n = 1, \ldots, N.$$

SVM for binary classification

$$\arg\max_{\mathbf{w},b}\left\{\frac{1}{\|\mathbf{w}\|}\min_n\left[t_n\left(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)+b\right)\right]\right\} \Rightarrow \arg\min_{\mathbf{w},b}\frac{1}{2}\|\mathbf{w}\|^2$$

$$t_n\left(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)+b\right)\geqslant 1, \qquad n=1,\ldots,N.$$

$$L(\mathbf{w},b,\mathbf{a})=\frac{1}{2}\|\mathbf{w}\|^2-\sum_{n=1}^{N}a_n\left\{t_n(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n)+b)-1\right\}$$

Next

- SVM (continued)