

Learning From Data

Lecture 23

SVM's: Maximizing the Margin

A Better Hyperplane
 Maximizing the Margin
 Link to Regularization

M. Magdon-Ismail
 CSCI 4100/6100

RECAP: Linear Models, RBFs, Neural Networks

Linear Model with Nonlinear Transform

$$h(\mathbf{x}) = \theta \left(w_0 + \sum_{j=1}^d w_j \Phi_j(\mathbf{x}) \right)$$

Neural Network

$$h(\mathbf{x}) = \theta \left(w_0 + \sum_{j=1}^m w_j \theta(\mathbf{v}_j^T \mathbf{x}) \right)$$

gradient descent

k-RBF-Network

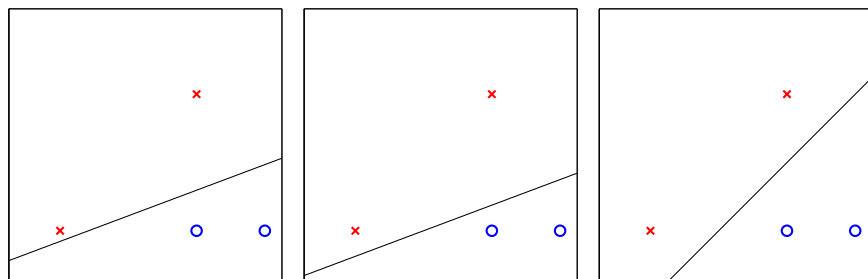
$$h(\mathbf{x}) = \theta \left(w_0 + \sum_{j=1}^k w_j \phi(\|\mathbf{x} - \boldsymbol{\mu}_j\|) \right)$$

k-means

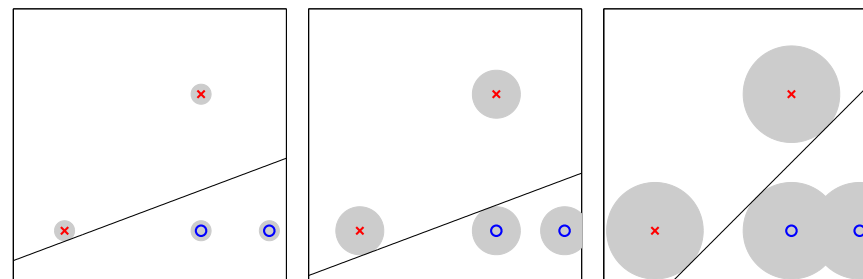
Neural Network: generalization of linear model by adding layers.

Support Vector Machine: more 'robust' linear model

Which Separator Do You Pick?

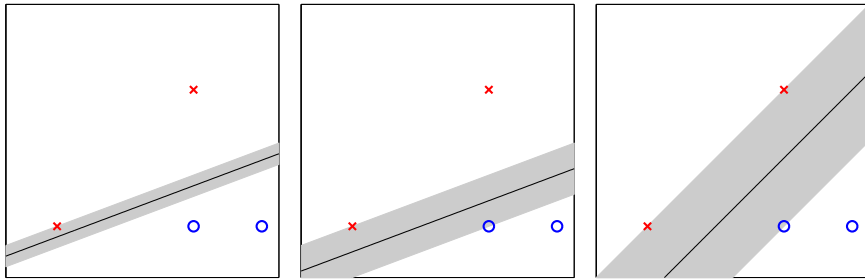


Robustness to Noisy Data



Being robust to noise (measurement error) is good (remember regularization).

Thicker Cushion Means More Robustness



We call such hyperplanes **fat**

Two Crucial Questions

1. Can we efficiently find the fattest separating hyperplane?
2. Is a fatter hyperplane better than a thin one?

Pulling Out the Bias

Before

$$\mathbf{x} \in \{1\} \times \mathbb{R}^d; \quad \mathbf{w} \in \mathbb{R}^{d+1}$$

$$\mathbf{x} = \begin{bmatrix} 1 \\ x_1 \\ \vdots \\ x_d \end{bmatrix}; \quad \mathbf{w} = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_d \end{bmatrix}.$$

$$\text{signal} = \mathbf{w}^T \mathbf{x}$$

Now

$$\mathbf{x} \in \mathbb{R}^d; \quad b \in \mathbb{R}, \quad \mathbf{w} \in \mathbb{R}^d$$

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}; \quad \mathbf{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_d \end{bmatrix}.$$

bias b

$$\text{signal} = \mathbf{w}^T \mathbf{x} + b$$

Separating The Data

Hyperplane $h = (b, \mathbf{w})$

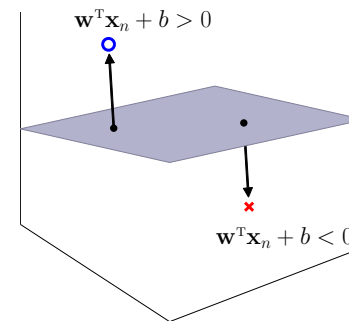
h separates the data means:

$$y_n(\mathbf{w}^T \mathbf{x}_n + b) > 0$$

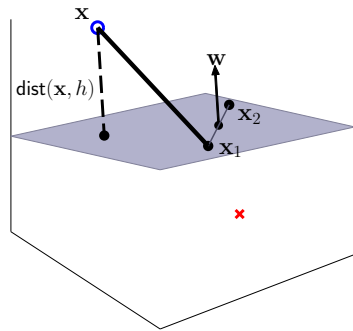
By rescaling the weights and bias,

$$\min_{n=1, \dots, N} y_n(\mathbf{w}^T \mathbf{x}_n + b) = 1$$

(renormalize the weights so that the signal $\mathbf{w}^T \mathbf{x} + b$ is meaningful)



Distance to the Hyperplane



\mathbf{w} is normal to the hyperplane:

$$\mathbf{w}^T(\mathbf{x}_2 - \mathbf{x}_1) = \mathbf{w}^T\mathbf{x}_2 - \mathbf{w}^T\mathbf{x}_1 = -b + b = 0.$$

(because $\mathbf{w}^T\mathbf{x} = -b$ on the hyperplane)

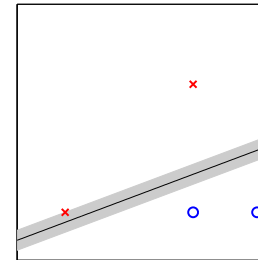
Unit normal $\mathbf{u} = \mathbf{w} / \|\mathbf{w}\|$.

$$\begin{aligned} \text{dist}(\mathbf{x}, h) &= |\mathbf{u}^T(\mathbf{x} - \mathbf{x}_1)| \\ &= \frac{1}{\|\mathbf{w}\|} \cdot |\mathbf{w}^T\mathbf{x} - \mathbf{w}^T\mathbf{x}_1| \\ &= \frac{1}{\|\mathbf{w}\|} \cdot |\mathbf{w}^T\mathbf{x} + b| \end{aligned}$$

Fatness of a Separating Hyperplane

$$\text{dist}(\mathbf{x}, h) = \frac{1}{\|\mathbf{w}\|} \cdot |\mathbf{w}^T\mathbf{x} + b|$$

Fatness = Distance to the closest point



Since $|\mathbf{w}^T\mathbf{x}_n + b| = |y_n(\mathbf{w}^T\mathbf{x}_n + b)| = y_n(\mathbf{w}^T\mathbf{x}_n + b)$,

$$\text{dist}(\mathbf{x}_n, h) = \frac{1}{\|\mathbf{w}\|} \cdot y_n(\mathbf{w}^T\mathbf{x}_n + b).$$

$$\begin{aligned} \text{Fatness} &= \min_n \text{dist}(\mathbf{x}_n, h) \\ &= \frac{1}{\|\mathbf{w}\|} \cdot \min_n y_n(\mathbf{w}^T\mathbf{x}_n + b) \quad \leftarrow \text{separation condition} \\ &= \frac{1}{\|\mathbf{w}\|} \quad \leftarrow \text{the margin } \gamma(h) \end{aligned}$$

Maximizing the Margin

$$\text{margin } \gamma(h) = \frac{1}{\|\mathbf{w}\|}$$

← bias b does not appear here

$$\text{minimize}_{b, \mathbf{w}} \frac{1}{2} \mathbf{w}^T \mathbf{w}$$

$$\text{subject to: } \min_{n=1, \dots, N} y_n(\mathbf{w}^T \mathbf{x}_n + b) = 1.$$

Maximizing the Margin

$$\text{margin } \gamma(h) = \frac{1}{\|\mathbf{w}\|}$$

← bias b does not appear here

$$\text{minimize}_{b, \mathbf{w}} \frac{1}{2} \mathbf{w}^T \mathbf{w}$$

$$\text{subject to: } \min_{n=1, \dots, N} y_n(\mathbf{w}^T \mathbf{x}_n + b) = 1.$$

$$\text{minimize}_{b, \mathbf{w}} \frac{1}{2} \mathbf{w}^T \mathbf{w}$$

$$\text{subject to: } y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1 \text{ for } n = 1, \dots, N.$$

Example – Our Toy Data Set

$$X = \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \\ 3 & 0 \end{bmatrix} \quad y = \begin{bmatrix} -1 \\ -1 \\ +1 \\ +1 \end{bmatrix}$$

$$y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1$$

$$\begin{aligned} -b &\geq 1 & (i) \\ -(2w_1 + 2w_2 + b) &\geq 1 & (ii) \\ 2w_1 + b &\geq 1 & (iii) \\ 3w_1 + b &\geq 1 & (iv) \end{aligned}$$

(i) and (iii) gives $w_1 \geq 1$

(ii) and (iii) gives $w_2 \leq -1$

So, $\frac{1}{2}(w_1^2 + w_2^2) \geq 1$ ($b = -1, w_1 = 1, w_2 = -1$)

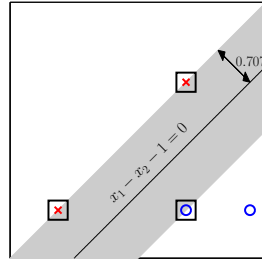
Optimal Hyperplane

$$g(\mathbf{x}) = \text{sign}(x_1 - x_2 - 1)$$

$$\text{margin: } \frac{1}{\|\mathbf{w}^*\|} = \frac{1}{\sqrt{2}} \approx 0.707.$$

For data points (i), (ii) and (iii) $y_n(\mathbf{w}^{*T} \mathbf{x}_n + b^*) = 1$

↑
Support Vectors



Quadratic Programming

$$\begin{aligned} &\text{minimize}_{\mathbf{u} \in \mathbb{R}^d} \quad \frac{1}{2} \mathbf{u}^T \mathbf{Q} \mathbf{u} + \mathbf{p}^T \mathbf{u} \\ &\text{subject to: } \mathbf{A} \mathbf{u} \geq \mathbf{c} \end{aligned}$$

$$\mathbf{u}^* \leftarrow \text{QP}(\mathbf{Q}, \mathbf{p}, \mathbf{A}, \mathbf{c})$$

($\mathbf{Q} = 0$ is linear programming)

Maximum Margin Hyperplane is QP

$$\text{minimize}_{b, \mathbf{w}} \quad \frac{1}{2} \mathbf{w}^T \mathbf{w}$$

subject to: $y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1$ for $n = 1, \dots, N$.

$$\begin{aligned} &\text{minimize}_{\mathbf{u} \in \mathbb{R}^{d+1}} \quad \frac{1}{2} \mathbf{u}^T \mathbf{Q} \mathbf{u} + \mathbf{c}^T \mathbf{u} \\ &\text{subject to: } \mathbf{A} \mathbf{u} \geq \mathbf{a} \end{aligned}$$

$$\mathbf{u} = \begin{bmatrix} b \\ \mathbf{w} \end{bmatrix} \in \mathbb{R}^{d+1}$$

$$\frac{1}{2} \mathbf{w}^T \mathbf{w} = [b \quad \mathbf{w}^T] \begin{bmatrix} 0 & \mathbf{0}_d^T \\ \mathbf{0}_d & \mathbf{I}_d \end{bmatrix} \begin{bmatrix} b \\ \mathbf{w}^T \end{bmatrix} = \mathbf{u}^T \begin{bmatrix} 0 & \mathbf{0}_d^T \\ \mathbf{0}_d & \mathbf{I}_d \end{bmatrix} \mathbf{u} \Rightarrow \mathbf{Q} = \begin{bmatrix} 0 & \mathbf{0}_d^T \\ \mathbf{0}_d & \mathbf{I}_d \end{bmatrix}, \mathbf{p} = \mathbf{0}_{d+1}$$

$$y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1 \equiv [y_n \quad y_n \mathbf{x}_n^T] \mathbf{u} \geq 1 \Rightarrow \begin{bmatrix} y_1 & y_1 \mathbf{x}_1^T \\ \vdots & \vdots \\ y_N & y_N \mathbf{x}_N^T \end{bmatrix} \mathbf{u} \geq \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \Rightarrow \mathbf{A} = \begin{bmatrix} y_1 & y_1 \mathbf{x}_1^T \\ \vdots & \vdots \\ y_N & y_N \mathbf{x}_N^T \end{bmatrix}, \mathbf{c} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

Back To Our Example

Exercise:

$$X = \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \\ 3 & 0 \end{bmatrix} \quad y = \begin{bmatrix} -1 \\ -1 \\ +1 \\ +1 \end{bmatrix}$$

$$y_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1$$

$$\begin{aligned} -b &\geq 1 & (i) \\ -(2w_1 + 2w_2 + b) &\geq 1 & (ii) \\ 2w_1 + b &\geq 1 & (iii) \\ 3w_1 + b &\geq 1 & (iv) \end{aligned}$$

Show that

$$\mathbf{Q} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \mathbf{p} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad \mathbf{A} = \begin{bmatrix} -1 & 0 & 0 \\ -1 & -2 & -2 \\ 1 & 2 & 0 \\ 1 & 3 & 0 \end{bmatrix} \quad \mathbf{c} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Use your QP-solver to give

$$(b^*, w_1^*, w_2^*) = (-1, 1, -1)$$

Primal QP algorithm for linear-SVM

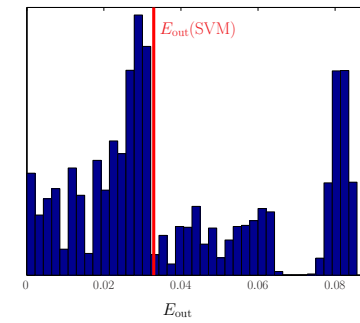
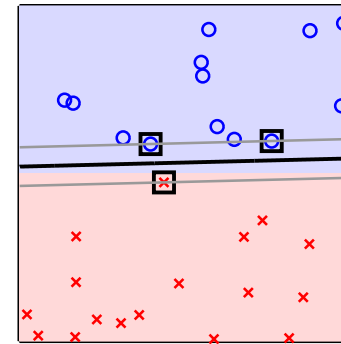
1. Let $\mathbf{p} = \mathbf{0}_{d+1}$ be the $(d+1)$ -vector of zeros and $\mathbf{c} = \mathbf{1}_N$ the N -vector of ones. Construct matrices Q and A , where

$$A = \underbrace{\begin{bmatrix} y_1 & -y_1 \mathbf{x}_1^T \\ \vdots & \vdots \\ y_N & -y_N \mathbf{x}_N^T \end{bmatrix}}_{\text{signed data matrix}}, \quad Q = \begin{bmatrix} 0 & \mathbf{0}_d^T \\ \mathbf{0}_d & I_d \end{bmatrix}$$

2. Return $\begin{bmatrix} b^* \\ \mathbf{w}^* \end{bmatrix} = \mathbf{u}^* \leftarrow \text{QP}(Q, \mathbf{p}, A, \mathbf{c})$.

3. The final hypothesis is $g(\mathbf{x}) = \text{sign}(\mathbf{w}^{*T} \mathbf{x} + b^*)$.

Example: SVM vs PLA



PLA depends on the ordering of data (e.g. random)

Link to Regularization

$$\begin{aligned} & \underset{\mathbf{w}}{\text{minimize}} && E_{\text{in}}(\mathbf{w}) \\ & \text{subject to:} && \mathbf{w}^T \mathbf{w} \leq C. \end{aligned}$$

	optimal hyperplane	regularization
minimize	$\mathbf{w}^T \mathbf{w}$	E_{in}
subject to	$E_{\text{in}} = 0$	$\mathbf{w}^T \mathbf{w} \leq C$

