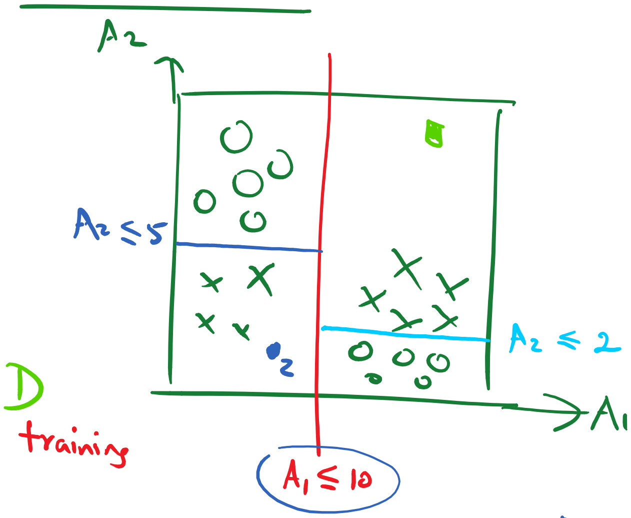
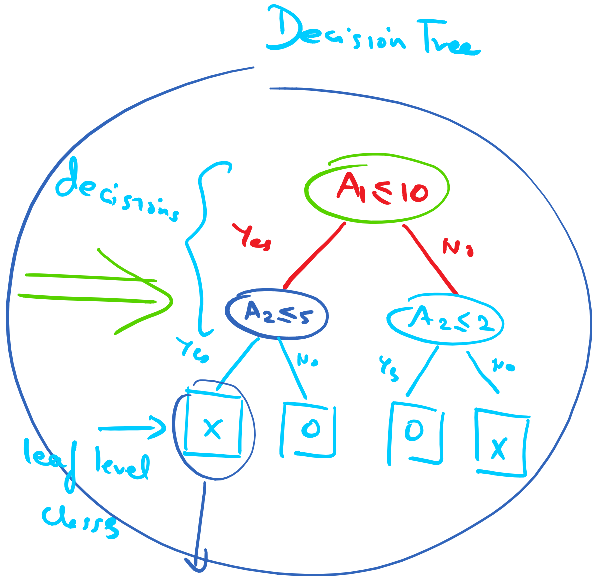
Tuesday, October 4, 2016 10:07 AM

Decision Trees



Decision Tree



disjunction over conjunctions

(DNF)

Disjunctive Normal Form

R1: If $A_1 \le 10$ and $A_2 \le 5$ then class = X
OR
R2: If $A_1 \le 10$ and $A_2 > 5$ then class = O
OR
R3: If $A_1 > 10$ and $A_2 \le 2$ then class = O
OR
R4: If $A_1 > 10$ and $A_2 > 2$ then class = X

$\vec{z} = (\underset{A_1}{6}, \underset{A_2}{1})$ $\hat{y} = X$

Greedy strategy: locally optimal decision

Sample D train

	A1	A2	class
x_1	2	6	H
x_2	5	1	L
x_3	10	3	H
x_4	1	10	H
x_5	20	15	L

$A_1 \le v_1$ $A_2 \le v_2$

$A_1 \in \mathbb{R}$

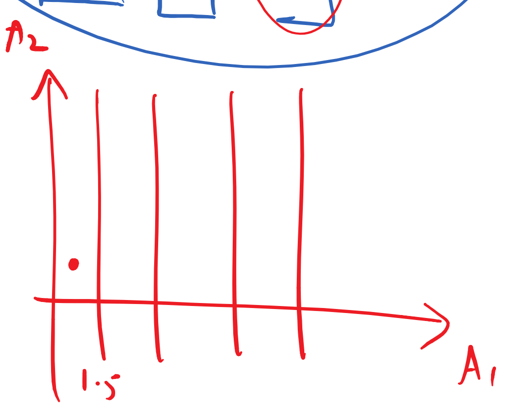
$\{1, 2, 5, 10, 20\}$ — 1.5, 3.5, 7.5, 15

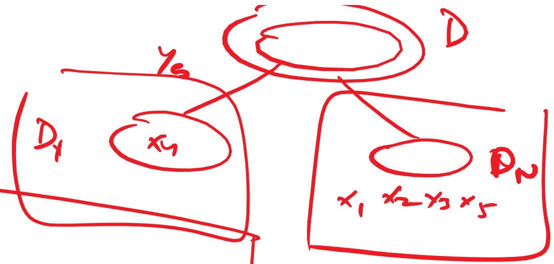
$A_2 \in \mathbb{R}$

$\{1, 3, 6, 10, 15\}$ — 2, 4.5, 8, 12.5

$A_1 \le 1.5$
$A_1 \le 3.5$
$A_1 \le 7.5$
$A_1 \le 15$

$A_2 \le 2$
$A_2 \le 4.5$
$A_2 \le 8$
$A_2 \le 12.5$





① Entropy of $\underline{D}$

$$\to H(D) = -\sum_{i=1}^{K} \hat{P}_i \log \hat{P}_i$$

$\hat{P}_H : 3/5$

$\hat{P}_L : 2/5$

② Compute the Entropy after the decision/$\underline{\text{split}}$

$$H(D_Y, D_N) = \underbrace{\frac{n_Y}{n} H(D_Y) + \frac{n_N}{n} H(D_N)}_{\text{weighted sum of resulting regions}}$$

$A_1 \le 7.5$ D

$A_1 > v$

D_Y

x_1	H
x_2	L
x_4	H

$\frac{n_Y}{n} = \frac{3}{5}$

$H(D_Y)$

$P_H : 2/3$

$P_L : 1/3$

D_N

x_3	H
x_5	L

$\frac{2}{5} = \frac{n_N}{n}$

$\downarrow$

$H(D_N)$

$P_H = 1/2$

$P_L = 1/2$

③ Information Gain

$$IG(A_1 \le v_1) = \underbrace{H(D) - H(D_Y, D_N)}$$

Out of all splits/Decisions pick the one that $\underline{\text{maximizes}}$

Out of all splits/Decisions pick the one that $\underline{\text{maximizes}}$ IG

Gini Index

$\rightarrow$ $$G(D) = 1 - \sum_{i=1}^{k} \underline{p_i^2}$$

$\rightarrow$ $$G(D_Y, D_N) = \frac{n_Y}{n} G(D_Y) + \frac{n_N}{n} G(D_N)$$

Categorical Decisions

	A_1	A_2	Class
x_1	1	R	H
x_2	3	G	L
x_3	10	R	H
x_4	5	B	H
x_5	20	G	L

A_1: 1 3 5 10 20

A_2: ?

$$A_2 = \underbrace{\{R, G, B\}}_{\text{Set of Values}}$$

$$\underbrace{A_2 \in V}$$

Yes / No

$A_2 \in \{R\}$

$A_2 \in \{G\}$

$A_2 \in \{B\}$

V

Yes	No
$\{R\}$	$\{B, G\}$
$\{G\}$	$\{R, B\}$
$\{B\}$	$\{R, G\}$
$\{R, G\}$	$\{B\}$
$\{R, B\}$	$\{G\}$
$\{B, G\}$	$\{R\}$

X

m values for A

$A_2 = \{R, G, B, Y\}$

R	GBY
G	RBY
B	RGY

m values for A

$O(2^{m-1})$ choices

$A \in R$

D_N: x_1 H, x_4 H

D_Y: x_2 L, x_3 H, x_5 L

$H(D)$

$H(D_Y, D_N)$

$A \in \{B, G\}$

D_Y: x_2, x_3, x_5

D_N: x_1, x_4

R	GBY
G	RBY
B	RGY
Y	RGB
RG	BY
RB	GY
RY	GB
GB	RY
GY	RB
BY	RG
RGB	Y
RGY	B
RBY	G
GBY	R

Chapter 1: Geometry + Algebra + Stats (eigen decomposition)

Chapter 2: Numeric / Normal (Σ)

Chapter 3: mapping categorical $\rightarrow$ one-hot

Bernoulli $\rightarrow$ Binomial

Multivariate Bernoulli $\rightarrow$ Multinomial

Prob.

Counts

PMF $\leftarrow$ Joint Distribution

Chapter 5: K, types of kernels

PSD

operations on K (5.3)

Chapter 6: vol, angle, prob as $d \rightarrow \infty$

Chapter 7: PCA, KPCA, SVD

Chap 20: LDA

Objective + eigen

KLDA

Chap 18:

Bayes

KNN

Chapter 7: PCA, KPCA, SVD

Var MSE

Eigen on K

$\vec{\omega} = \sum c_i \phi(x_i)$

KNN

Chap 19: Decision Trees

SVM: Support Vector Machine

globally optimal decision objective

vs truth

Not the Best classifier

+ kernels

Structural Risk Minimization

large Margin classifier

Binary classification

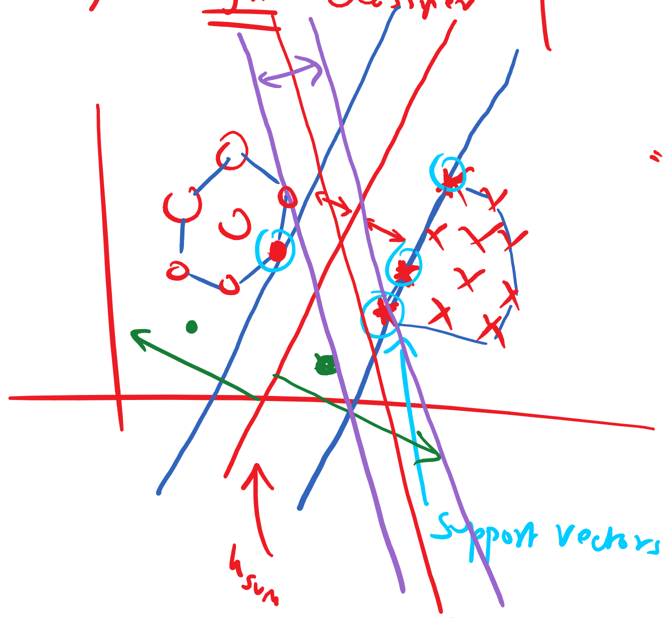
$y \in \{P, N\}$

$y \in \{+1, -1\}$

"linear" method

best "soft" hyperplane

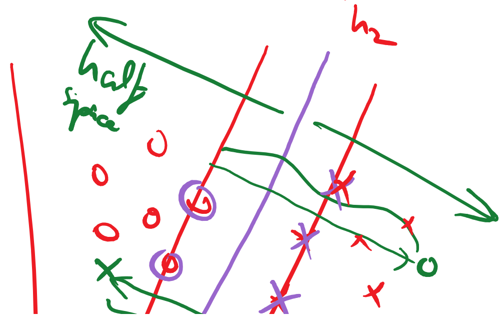
Separating hyperplane



Soft margin

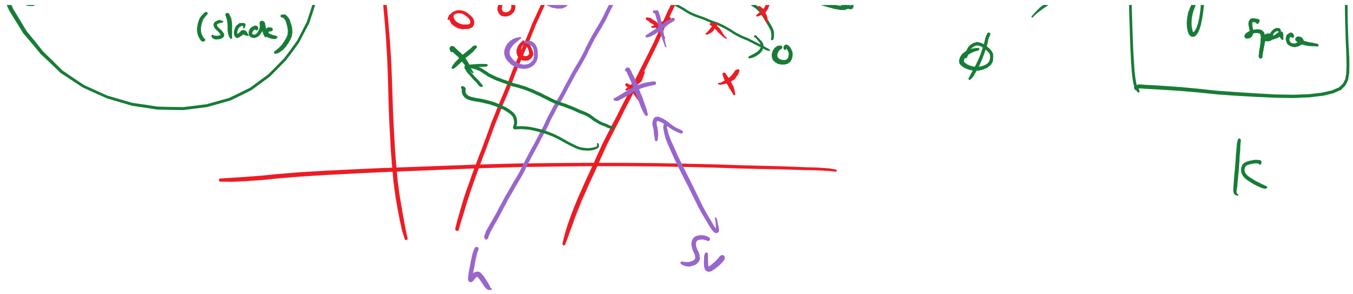
① Max margin

② Min error (slack)



ϕ

feature space



$$\boxed{h(x): \vec{w}^T x + b = 0}$$

All the points that comprise the hyperplane

$\boxed{\vec{x_i}} \in \mathbb{R}^d$ $\quad \vec{w} = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix} \quad \vec{x_i} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix}$

$$w_1 x_1 + w_2 x_2 + \dots + w_d x_d + b = 0$$

bias (offsets in each dimension)

in 2D $\boxed{y = mx + b}$

$$mx - y + b = 0$$

$$\boxed{\underbrace{\begin{pmatrix} m \\ -1 \end{pmatrix}}_{\vec{w}}^T \cdot \underbrace{\begin{pmatrix} x \\ y \end{pmatrix}}_{\vec{x}} + b = 0}$$

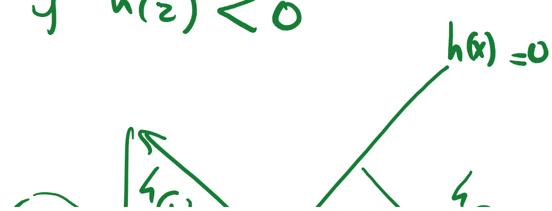
SVM classifier (test point $\vec{z}$)

$$\hat{y} : \begin{cases} +1 & \text{if } h(\vec{z}) > 0 \\ -1 & \text{if } h(\vec{z}) < 0 \end{cases}$$

model:

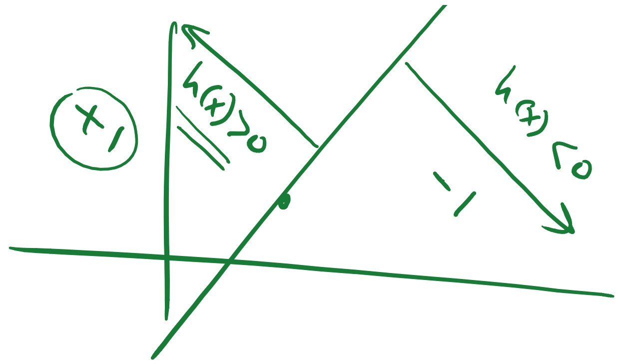
best $h(x)$

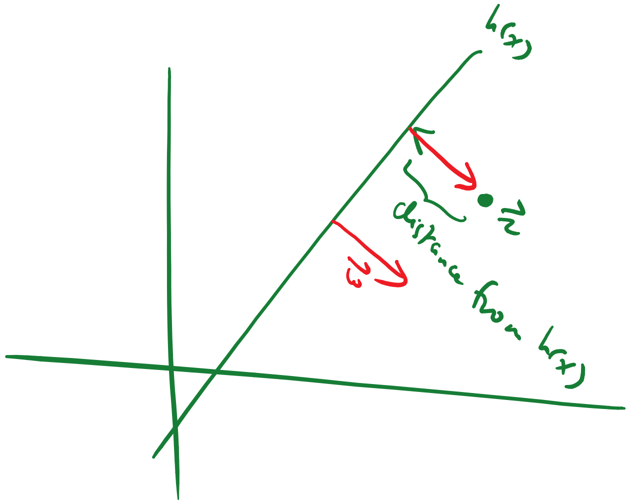
$$\boxed{\hat{y} = \text{sign}(h(\vec{z}))}$$

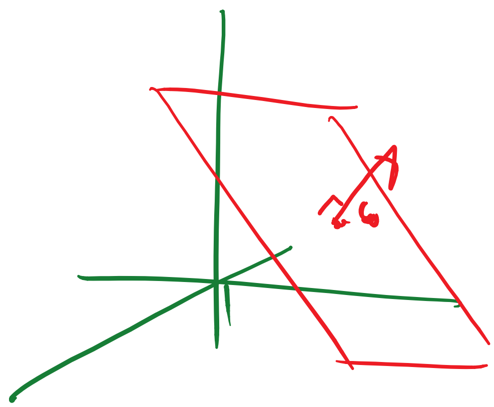


$$\boxed{\hat{y} = \text{sign}(h(\vec{z}))}$$

SVM







$\vec{\omega}$ is orthogonal to $h(x)$

$$\boxed{\text{distance} = \frac{h(\vec{z})}{\|\omega\|}} = \frac{\omega^T z + b}{\|\omega\|}$$