Classification#1

Discriminant functions

Machine learning, 2021

Mansoor Rezghi

Department of Computer science, TMU

Ref: CB, AW, DU

References

- Bishop: chap4
- R. O. Duda, P.E. Hart, D. G. Stork, Pattern Classification, Second Edition, Wiley, 2001.(DU)-chap5

* Linear regression: norm 1& norm 2

* Logistic Regression as discriminant function classifier

* Perceptron classifer

* Support vector machine classifier

supervised | Regres $\longrightarrow$ line $\min \sum \text{loss}(\hat{y}_i, y_i)$

$\hookrightarrow$ optmt < $\boxed{\text{GD , SGD   Min-}}$ $x_{k+1} = x_k - \alpha \boxed{\nabla f_k}$

$\alpha\, g$ — subg —

Classification $\{(x_i, y_i)\}$ (x, y)

$\leftarrow$ feat $y_i \to$ Label $\downarrow$? $y \in \{C_1, \ldots, C_k\}$

Classification

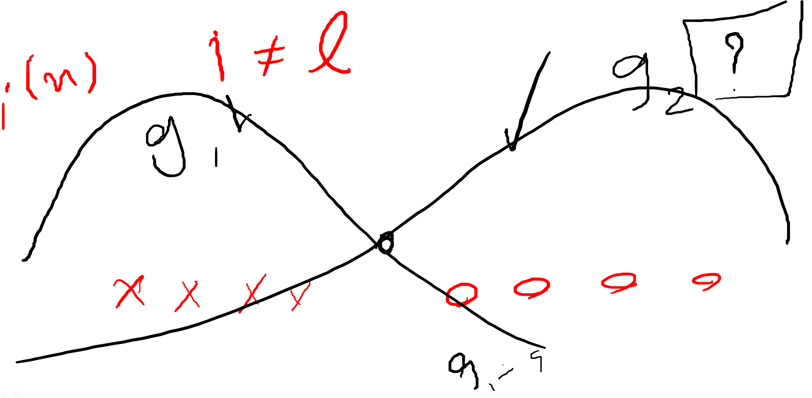
Given Data set x_i with known labels t_i

Predict the label of test data x

$\{C_1 \ldots C_n\}$

$C_i \quad g_i(x)$

$x \in C_l \quad g_l(x) > g_i(x) \quad i \neq l$



$C_1 \rightarrow g_1$

$\vdots$

$C_K \rightarrow g_K$

$g_i(x) = g_j(x)$

$\{x_i, y_i\}$

$P(x,C)$

Discriminant function

For each class find discriminat function $g_i(x)$

Assign x to class C_l if $l = \arg\max_{1 \le i \le K} g_i(x)$

P(x,C) or P(C|x)

- Generative Models
- Discriminative Models

$\bar{x}$

$P(\bar{x}, C_1) > P(\bar{x}, C_2)$

$P(x, C)$

$P(x, C)$

$\bar{x} \quad P(\bar{x}, C_1) > P(\bar{x}, C_2)$

$P(x, C_1) > P(x, C_2)$ Assign $X \longrightarrow C_1$

$P(A,B) = P(A|B)\,P(B)$

$\Updownarrow$

$P(C_1|x)\cancel{P(x)} > P(C_2|x)\cancel{P(x)}$

$$P(C_1|x) = \frac{P(x|C_1)\,P(C_1)}{P(x)}$$

$\Downarrow$

$\boxed{P(C_1|x)} > P(C_2|x)$

$\Updownarrow$

$$\frac{P(x|C_1)P(C_1)}{P(x)} > \frac{P(x|C_2)P(C_2)}{P(x)}$$

$\Updownarrow$

$P(x|C_1)\,P(C_1) > P(x|C_2)\,P(C_2)$

Generative.

$P(x|C_1)$

$P(x=1..|C_1) =$

$P(x) = P(x|C_1)P(C_1) + P(x|C_2)P(C_2)$

Classification

Generative models

Discriminative functions

- Least squares, Perceptron
- Logistic regression classifier
- Optimization based methods
- Support Vector machine
- ....

Discriminative models: Logistic Regression

- KNN
- Nave Bayes
- Decision Tree
- Random forest

Discriminant function

For each class find discriminat function $g_i(x)$

Assign x to class C_l if $l = \arg\max_{1i \leq K} g_i(x)$

Linear $g_i(x)$ $\Rightarrow$ Linear Classfier

g

Discriminant Function

Linear separable classes : classes can be separated via linear function(Linear decision surface)

$$g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + w_0,$$

Kernel $\Leftarrow \Phi(x)$

$\uparrow$

x

By increasing the dimension

?

Nonlinear separable classes : classes can not be separated via linear function(NonLinear decision surface)

Nonlinear to Linear

Quadratic discriminant function

$x \in R^d$

$$g(\mathbf{x}) = w_0 + \sum_{i=1}^{d} w_i x_i + \sum_{i=1}^{d} \sum_{j=1}^{d} w_{ij} x_i x_j.$$

d=2

$$\bar{x} = \begin{pmatrix} x_0 \\ x_1 \\ x_1 x_2 \\ x_1^2 \\ x_2^2 \end{pmatrix} \quad \bar{W} = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \\ w_{12} + w_{21} \\ w_{11} \\ w_{22} \end{pmatrix}$$

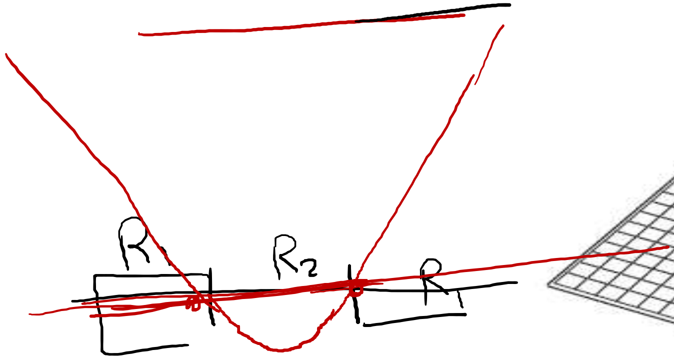
$$g(\mathrm{x}) = \mathrm{g}(\bar{\mathrm{x}}) = \bar{\mathrm{w}}_0 + \sum_{i=1}^{5} \bar{w}_i \bar{x}_i$$

$y = x^2$ $\quad x \to \begin{pmatrix} x \\ x^2 \end{pmatrix}$

Nonlinear to Linear separable(R. O. Duda)

$\boxed{\text{kernel.}}$

ϕ



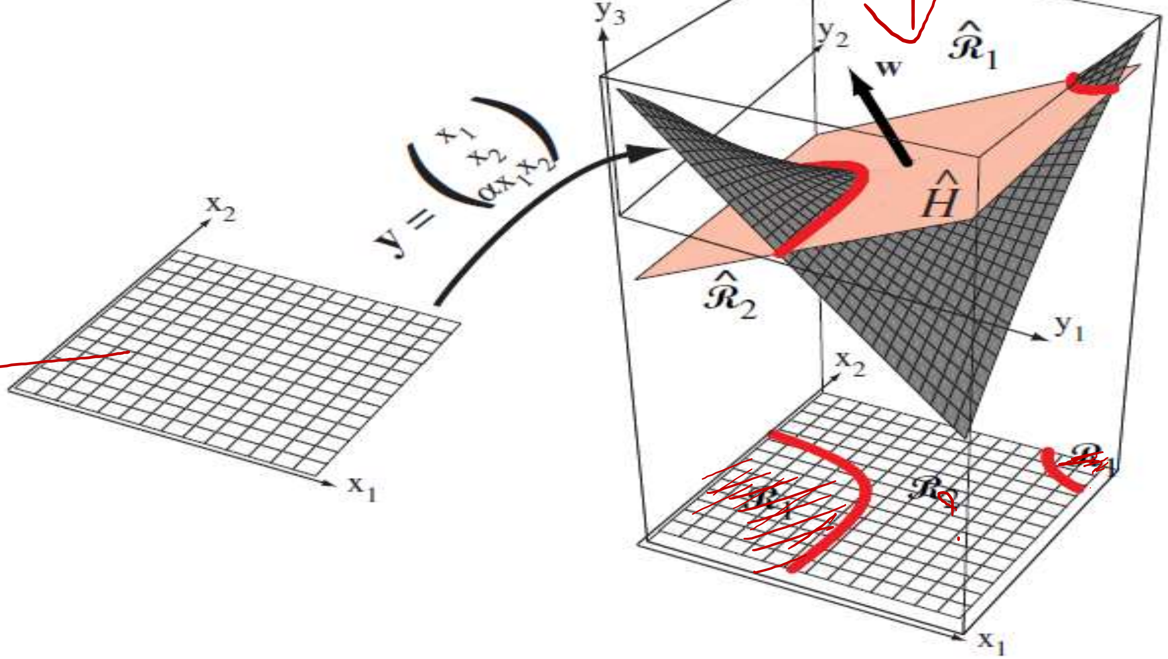


Figure 5.6: The two-dimensional input space x is mapped through a polynomial function f to y. Here the mapping is $y_1 = x_1$, $y_2 = x_2$ and $y_3 \propto x_1 x_2$. A linear discriminant in this transformed space is a hyperplane, which cuts the surface. Points to the positive side of the hyperplane $\hat{H}$ correspond to category ω_1, and those beneath it ω_2. Here, in terms of the x space, $\mathcal{R}_1$ is a not simply connected.

$$\boxed{x \in C_i \quad g_i(x) \geq g_j(x)}$$

$x \in C_j \rightarrow (x_i, t_i) \quad t_i = \begin{cases} 1 & x_i \in C_1 \\ -1 & x_i \in C_2 \end{cases}$

$(x_i, t_i) \quad t_i = \begin{cases} 1 & x \in C_1 \\ 0 & x \in C_2 \end{cases}$

$(x, y_i) \quad y_i \in \{C_1 \ldots C_k\} \rightarrow t_i = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \rightarrow l^{\text{th}}$ position

$x_i \in C_l$

$g(x) \rightarrow t_i$

$x_i \rightarrow t_i$

$g(x_i) \rightarrow t_i$

$C_1 \quad g_1$

$C_2 \quad g_2$

$g_1(x) > g_2(x)$

$h(x) = g_1(x) - g_2(x)$

$h(x) > 0$

Linear Regression Based classifier(Bishop-chap4)

$\boxed{\text{Regression Based.}}$ Two-class

$$\{(x_i, y_i)\} \quad y_i \in C_1 \text{ or } \in C_2 \Rightarrow t_i = \begin{cases} 1 & x_i \in C_1 \\ -1 & x_i \in C_2 \end{cases}$$

$\boxed{316}$

$g_i(x)$ if $\underline{g(x_i) \approx t_i}$

$g(x) = W^T x \qquad W^T x_i \approx t_i$

$$\frac{1}{N} \sum \text{loss}(W^T x_i, t_i) = \sum_i (W^T x_i - t_i)^2 + \lambda \|w\|_2$$

$$|W^T x_i - t_i| \qquad \lambda \|w\|_1$$

$h(x) > 0 \rightarrow x \in C_1$

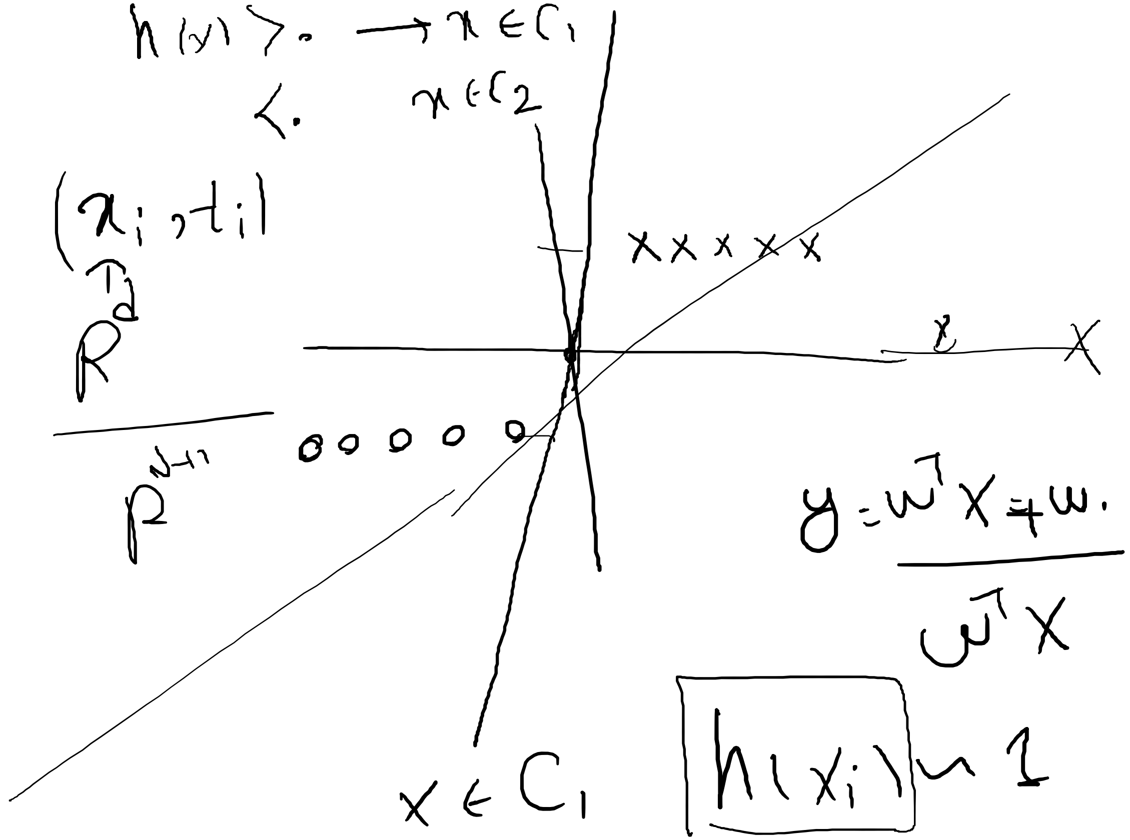
$< 0 \quad x \in C_2$

(x, t_i) $\boxed{\text{line Separable}}$ $\Pi(h(x_i)) = \begin{cases} 1 \\ -1 \end{cases}$

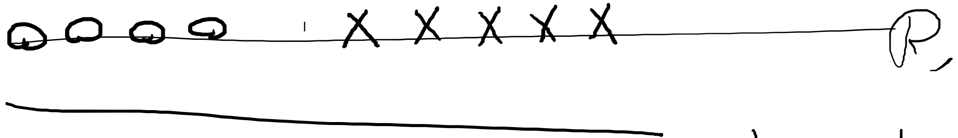
(x_i, t_i)

R^d

R^{d+1}



$\Leftarrow$



C_2 C_1 $x_i \in R^d$

$y = w^T x + w_0$

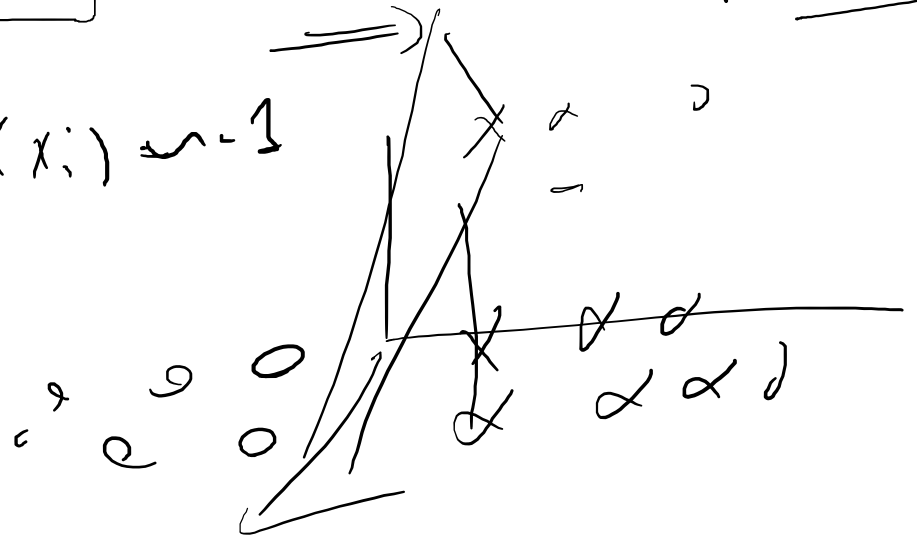
$w^T x = h(x)$

$x \in C_1$ $\boxed{h(x_i)} \sim 1$

$\forall x \; \boxed{h(x)} > 0 \rightarrow x \in C_1$

$< 0 \rightarrow x \in C_2$

$x \in C_2$ $h(x_i) \sim -1$



$$\boxed{\Pi(t) = \begin{cases} 1 & t > 0 \\ -1 & t < 0 \end{cases}}$$

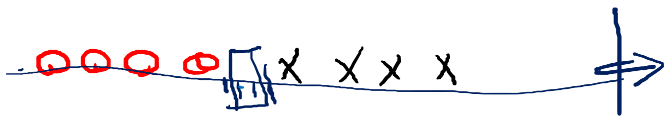
$\Pi(h(x_i))$

(x_i, y_i) $y_i = \begin{cases} 1 & x \in C_1 \\ -1 & x \in C_2 \end{cases}$

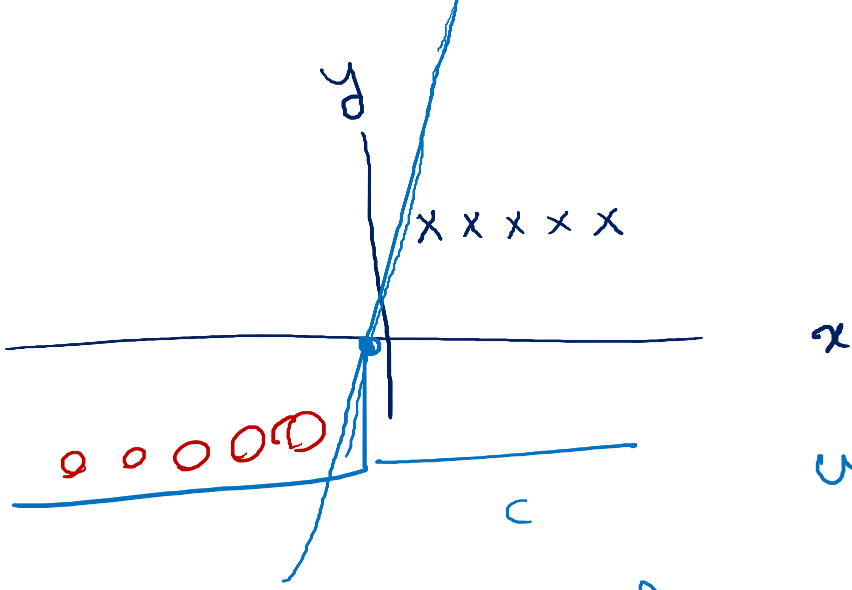
l_i

$y_i = \begin{cases} 1 & x \in C_1 \\ 0 & x \in C_2 \end{cases}$

$w^T x + w_0 >$



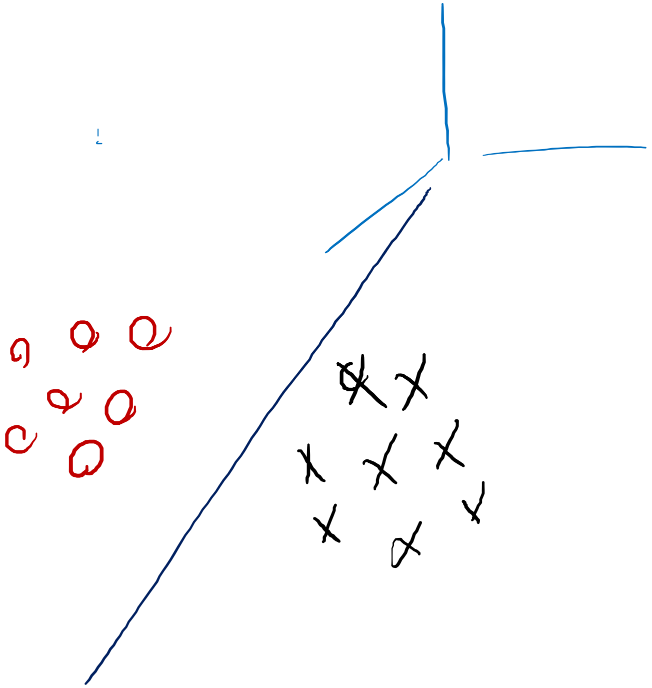
$x > a$



$w^T x_i + w \sim t_i$

$f(x) = w^T x_i + w_0$

$w^T x + w_0 > 0 \Rightarrow$



$ax + b \sim t_i$

$C_1 \Leftarrow ax + b > 0$

$\boxed{x > -\frac{b}{a}}$

$a_1 x_1 + a_2 x_2 + b \geq 0$

$\boxed{a_1 x_1 + a_2 x_2 = -b}$

Regression Based Classifier:

Input $\{X_i\}$ $\quad X_i \in \{C_1 \dots C_k\}$

if $X_i \in C_l$ $\quad t_i = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$ lth position.

$$X_i \longrightarrow t_i = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$C_1, C_2, \ldots, C_k$

$\downarrow \quad \downarrow \quad\quad \downarrow$

$g_1 \quad g_2 \quad\quad g_k$

$x \to A \quad \forall x \in C_\ell \quad g_\ell(x) > g_i(x)$

$i \neq \ell$

$$g_1(x) = w_1^T x$$
$$g_2(x) = w_2^T x$$
$$\vdots$$
$$g_k(x) = w_k^T x$$

$$g(x) = \begin{pmatrix} g_1(x) \\ \vdots \\ g_k(x) \end{pmatrix} = \begin{pmatrix} w_1^T x \\ w_2^T x \\ \vdots \\ w_k^T x \end{pmatrix} = W^T x$$

$$W \in R^{d \times k}$$

R_k

$$t = \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$g(x) =$

$$\begin{pmatrix} w_1^T x_i \\ w_2^T x_i \\ \vdots \\ w_k^T x_i \end{pmatrix} \sim \begin{pmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$x \in C_2$

$$g(x_i) \simeq t_i$$

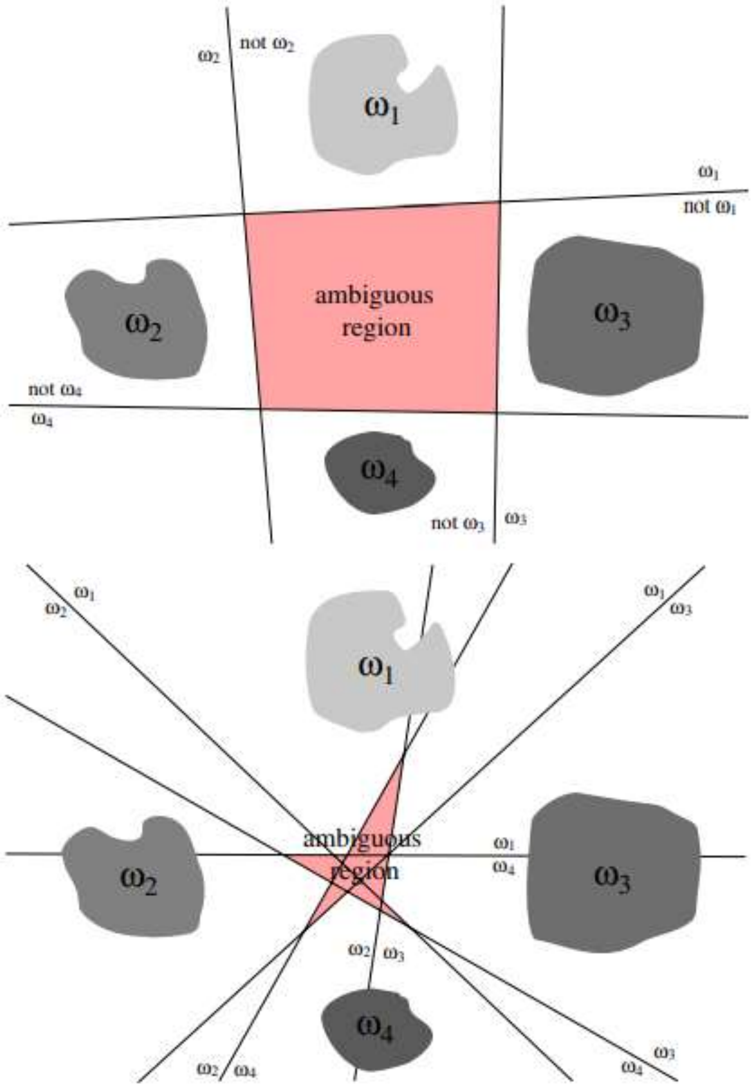
$x_i \in R^d \to t_i \in R^k$ $\quad g(x) = W^T x$ $\quad \lambda \Omega(W)$

$W \in R^{d \times k}$

$$\sum \text{loss}(g(x_i), t_i) = \sum \| W^T x_i - t_i \|_2^2 \;\Big|\; + \boxed{\begin{array}{l} \lambda \|W\|_F^2 \\ \lambda \|W\|_1 \text{ — m} \\ \lambda \|W\|_{2,1} \end{array}}$$

$$= \sum \| W^T x_i - t_i \|_1$$

Regressio



not ω2
ω2
ω1
ω1
not ω1
ambiguous region
ω2
ω3
not ω4
ω4
ω4
not ω3
ω3
ω1
ω2
ω1
ω1
ω3
ambiguous region
ω1
ω4
ω2
ω3
ω2
ω3
ω4
ω2
ω4
ω3
ω4

Drawbacks of LS method

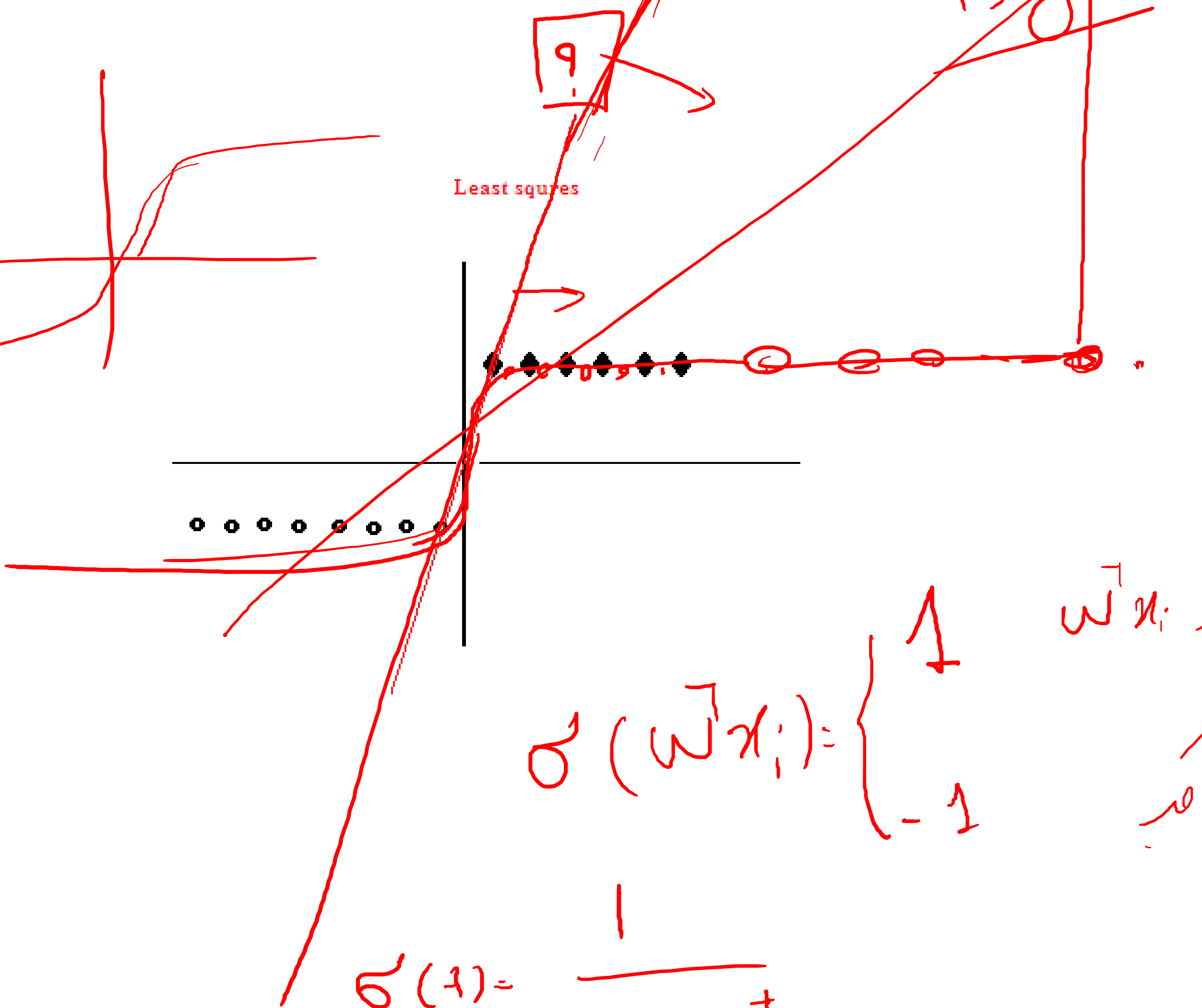
Regression

$\lim_{t \to -\infty} \sigma(t) = 0$

Sensitive to

- Number of train data of each class
- Outlier

$w^T x_i$ in $y_i = \{1$



$$\sum \text{loss}(w^T x_i, y_i) \; ?$$

$$= \sum \text{loss}(\sigma(w^T x_i), y_i)$$

$$\sigma(w^T x_i) = \begin{cases} 1 \\ -1 \end{cases}$$

$w^T x_i$ خیلی بزرگ

خیلی کوچک

$$\sigma(t) = \frac{1}{1+e^{-t}}$$

$\lim_{t \to -\infty} \sigma(t) =$

$\lim_{t \to \infty} \sigma(t) = 1$

Logistic Regression

$$\min \sum \text{loss}(\sigma(w^T x_i), t_i)$$

$$\sum (\sigma(w^T x_i) - t_i)^2$$

(x_i, t_i)

$$t_i = \begin{cases} 1 & x \in C_1 \\ 0 & x \in C_2 \end{cases}$$

$$\sum |\sigma(w^T x_i) - t_i|$$

$$\sigma(t) = \frac{1}{1+e^{-t}}$$

- Deterministic viewpoint
- Statistical viewpoint Discriminative model

$$\sum \text{loss}(\sigma(w^T x_i), t_i)$$

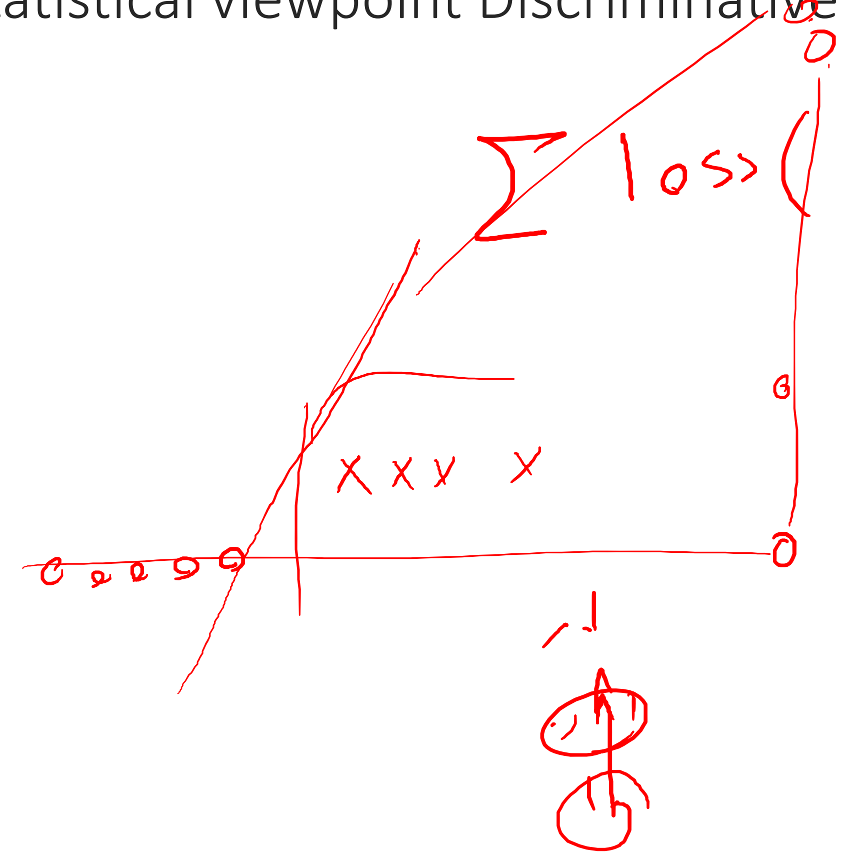
$\sigma(0) = \frac{1}{2}$

$\lim \sigma(t) = 1$, $t \to \infty$

$\lim \sigma(t) = 0$, $t \to -\infty$

$\sigma(w^T x_i) = 0.9$, $t_i = 1$

$\sigma(w^T x_i) = 0.1$, $t_i = 1$

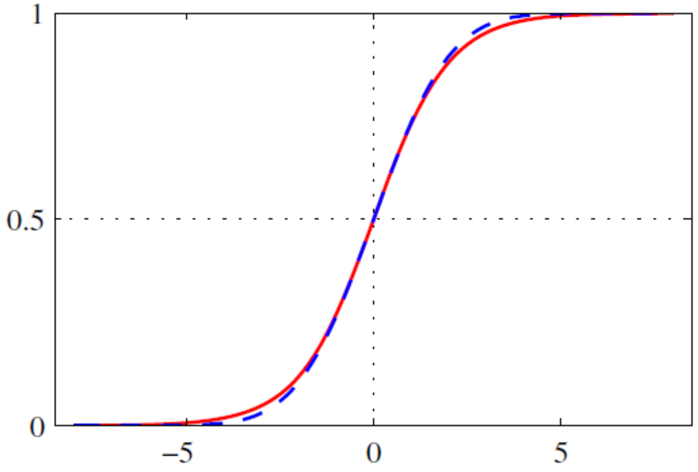


x.

$w^T \Phi$

Logistic Regression classifier-Two class

$$y(\mathbf{x}) = f\left(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x})\right) \quad f(a) = \sigma(a) = \frac{1}{1+\exp(-a)}$$



$$y(\mathbf{x}) = f\left(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x})\right) \quad f(a) = \sigma(a) = \frac{1}{1+\exp(-a)}$$

$$f_{\mathbf{w}}(\phi(\mathbf{x})) = f\left(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x})\right)$$

Conv

Non-convex objective function

$$J(\mathbf{w}) = \sum_{i=1}^{N}\left(y^{(i)} - f_{\mathbf{w}}(\phi(x^{(i)}))\right)^2$$

$$J(\mathbf{w}) = \sum_{i=1}^{N} \text{Cost}\left(y^{(i)}, f_{\mathbf{w}}(\phi(x^{(i)}))\right)$$

$y^{(i)}$

$\sigma(w^T x_i) \simeq 1 \to \infty$

$f_w(\phi(x^{i})) \simeq 1$

$-y^{(i)}$

$s = \log(t)$

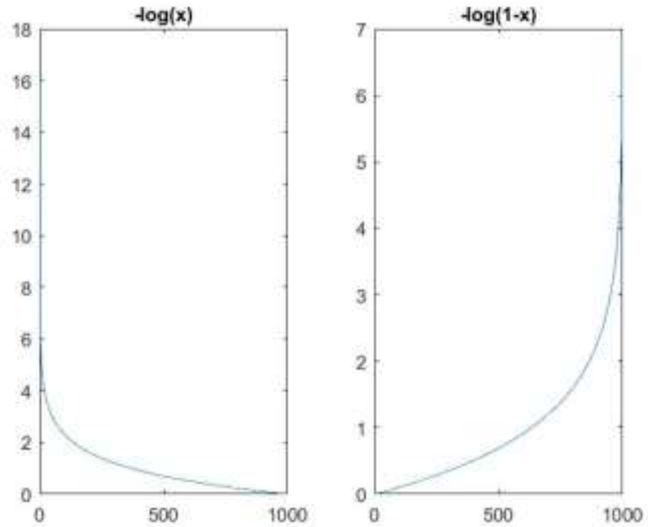
$t = e^s$

For $y^{(i)} = 0$	$f_{\mathrm{w}}(\phi(\mathbf{x}^{(i)})) \simeq 0 \Rightarrow$	$\text{Cost}\left(y^{(i)}, f_{\mathrm{w}}(\phi(\mathbf{x}^{(i)}))\right) \simeq 0$
	$f_{\mathrm{w}}(\phi(\mathbf{x}^{(i)})) \simeq 1 \Rightarrow$	$\text{Cost}\left(y^{(i)}, f_{\mathrm{w}}(\phi(\mathbf{x}^{(i)}))\right) \simeq$ **Inf**

For $y^{(i)} = 1$	$f_{\mathrm{w}}(\phi(\mathbf{x}^{(i)})) \simeq 1 \Rightarrow$	$\text{Cost}\left(y^{(i)}, f_{\mathrm{w}}(\phi(\mathbf{x}^{(i)}))\right) \simeq 0$
	$f_{\mathrm{w}}(\phi(\mathbf{x}^{(i)})) \simeq 0 \Rightarrow$	$\text{Cost}\left(y^{(i)}, f_{\mathrm{w}}(\phi(\mathbf{x}^{(i)}))\right) \simeq$ **Inf**

$$\text{Cost}\left(y^{(i)}, f_{\mathrm{w}}(\phi(\mathbf{x}^{(i)}))\right) =$$

$$-\left[y^{(i)} \ln f_{\mathrm{w}}(\phi(\mathbf{x}^{(i)})) + (1 - y^{(i)}) \ln(1 - f_{\mathrm{w}}(\phi(\mathbf{x}^{(i)})))\right]$$



$y^{(i)} = 0 \Rightarrow$

$y^{(i)} = 1$

Convex $$J(\mathrm{w}) = \sum_{i=1}^{N} \text{Cost}\left(y^{(i)}, f_{\mathrm{w}}(\phi(\mathbf{x}^{(i)}))\right) \to$$

x^i ,

$\text{if } y_i = 0 \xrightarrow{\text{انتظار}} \sigma(w^T x_i) \simeq 0$

$\sigma(w^T x_i) = 0 \rightarrow (e) = 0$

$\sigma(w^T x_i) \simeq 1$

$\sigma(w^T x_i) - 1 \simeq 0$

$\ln \frac{1}{\sigma(w^T x_i)}$

$\text{if } y_i = 0 \quad \boxed{\ln \frac{1}{1-\sigma(w^T x_i)}}$

$\sigma(w^T x_i) \simeq 0 \qquad \ln \frac{1}{1-\sigma(w^T x_i)} \simeq 0$

$\sigma(w^T x_i) \simeq 1 \qquad \ln \frac{1}{1-\sigma(w^T x_i)} \simeq \infty$

$\text{if } y_i = 1 \quad \ln \frac{1}{\sigma(w^T x_i)}$

$\sigma(w^T x_i) \simeq 1 \rightarrow \ln \frac{1}{\sigma(w^T x_i)} = 0$

$\sigma(w^T x_i) \simeq 0 \qquad \ln \frac{1}{\sigma(w^T x_i)} = \infty$

$$\operatorname{cost}(\sigma(w^T x_i), y_i) = \begin{cases} \ln \dfrac{1}{\sigma(w^T x_i)} & y_i = 1 \\ \ln \dfrac{1}{1-\sigma(w^T x_i)} & y_i = 0 \end{cases}$$

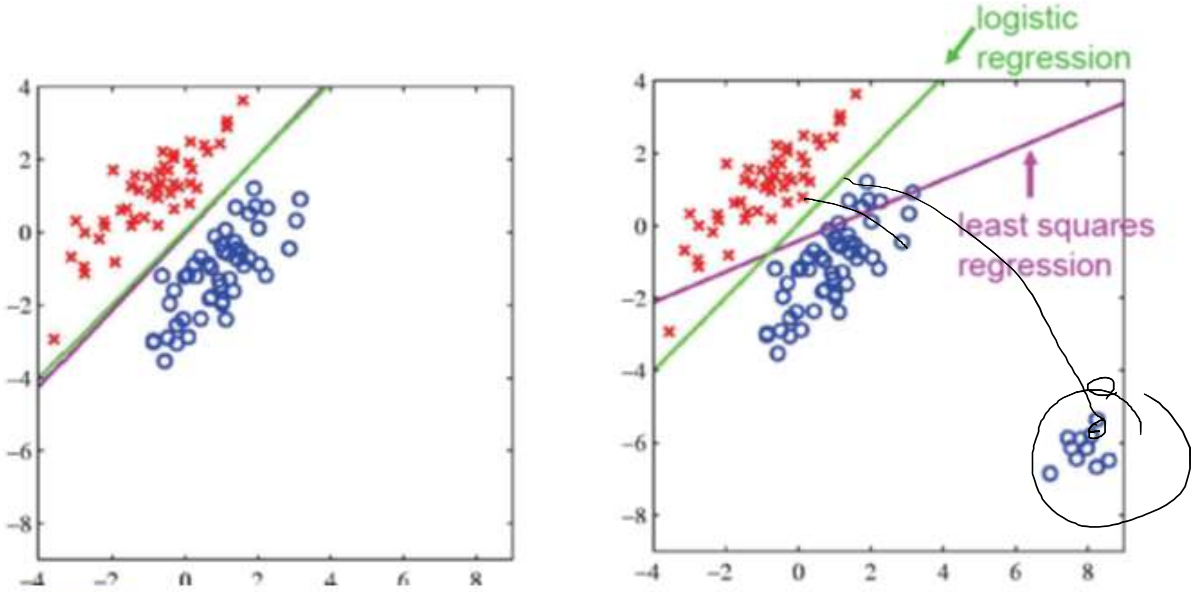
$$y_i \ln \frac{1}{\sigma(w^T x_i)} + (1-y_i) \ln \frac{1}{1-\sigma(w^T x_i)} = \operatorname{loss}(\sigma(w^T x_i), y_i)$$

Convex

$$\min \sum_{i=1}^{N} \operatorname{loss}(\sigma(w^T x_i), y_i) = \sum y_i \ln \frac{1}{\sigma(w^T x_i)} + (1-y_i) \ln\left(\frac{1}{1-\sigma(w^T x_i)}\right)$$

$$\max \sum y_i \ln \sigma(w^T x_i) + (1-y_i) \ln(1-\sigma(w^T x_i))$$

Concave.



Zemel

Logistic Regression-Multiclass objective function: Other Viewpoint

$$\max \; Loss(y, \hat{y}) = \Pi_j \hat{y}_j^{y_j},$$

$$\Downarrow$$

$$\max \; Loss(y, \hat{y}) = \ln \Pi_j \hat{y}_j^{y_j} = \sum_j y_j \ln \hat{y}_j \quad \Rightarrow \quad \max \; \sum_i Loss(y_i, \hat{y}_i) = \sum_i \sum_j y_{ji} \ln \hat{y}_{ji}$$

Perceptron

Input data $\{(x_i, t_i)\},\qquad t_i = \begin{cases} 1 & x_i \in C_1 \\ 0 & x_i \in C_2 \end{cases}$

Model: Find hyperplane $w^t x$ s.t

$$\forall x_i, \quad t_i(w^T x_i) \geq 0$$

Loss functions:

- Loss(w)= $|M|$,

  M={miss classified train data by $w^T x$ hyperplane}

- Loss(w)=$-\sum_{x_i \in M} t_i(w^T x_i)$ Perceptron

$\mathrm{Loss}(w) = -\sum_{x_i \in M} t_i(w^T x_i)$ Perceptron

$\Downarrow$ GD

$$w_{k+1} = w_k + \eta \sum_{x_n \in M} t_n x_n,$$

$\Downarrow$ SGD

$$w_{k+1} = w_k + \eta t_n x_n \quad w_{k+1} = w_k + t_n x_n$$

$\Rightarrow$

Perceptron

For i=1 … maxiter

For train data like $(\phi_i, t_i), i = 1, .., N$ train data

$y_i = \mathbf{w}^T \phi_i$

If $t_i y_i < 0$ Then

$w = w + t_i \phi_i$

However, the *perceptron convergence theorem* states that if there exists an exact solution (in other words, if the training data set is linearly separable), then the perceptron learning algorithm is guaranteed to find an exact solution in a finite number of steps. Proofs of this theorem can be found for example in Rosenblatt (1962),

Frank Rosenblatt. Principles of Neurodynamics: Perceptron and the Theory of Brain Mechanisms. Spartan Books, Washington, D.C., 1962

Albert B. J. Novikoff. On convergence proofs for perceptrons. In Proceedings of the Symposium on Mathematical Theory of Automata, volume 12, Brooklyn, New York, 1962.

http://ciml.info/dl/v0_99/ciml-v0_99-ch04.pdf

$$margin(\mathbf{D}, \boldsymbol{w}, b) = \begin{cases} \min_{(x,y)\in\mathbf{D}} y(\boldsymbol{w} \cdot \boldsymbol{x} + b) & \text{if } \boldsymbol{w} \text{ separates } \mathbf{D} \\ -\infty & \text{otherwise} \end{cases}$$

$$margin(\mathbf{D}) = \sup_{\boldsymbol{w},b} margin(\mathbf{D}, \boldsymbol{w}, b)$$

Theorem 2 (Perceptron Convergence Theorem). *Suppose the perceptron algorithm is run on a linearly separable data set* $\mathbf{D}$ *with margin* $\gamma > 0$. *Assume that* $||\boldsymbol{x}|| \leq 1$ *for all* $\boldsymbol{x} \in \mathbf{D}$. *Then the algorithm will converge after at most* $\frac{1}{\gamma^2}$ *updates.*

- Loss(w)= $-\sum_{x_i \in M} (w^T x_i)^2$
- Loss(w)= $-\sum_{x_i \in M} \frac{(w^T x_i - b)^2}{\|w\|^2}$

Multicategory Generalizations

$$g_i(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + w_{i0} \quad i = 1, ..., c,$$

$g_i(\mathbf{x}) = \mathbf{a}_i^t\mathbf{y} \quad i = 1, ..., c,$ where again $\mathbf{x}$ is assigned to ω_i if $g_i(\mathbf{x}) > g_j(\mathbf{x})$ for all $j \neq i$.

$\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n$, with n_i in the subset $\mathcal{Y}_i$ labelled ω_i.

Duda: 5.12.1

For the data from linear separable multi-class, there exist a set of vectors $W_i, \quad i = 1, ..., k$ such that if $\phi_k \in \mathcal{C}_i$, then

$$\hat{w}_i^T \phi_k > \hat{w}_j^T \phi_k \text{ For } i \neq j$$

$$\hat{\alpha} = \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_c \end{bmatrix}$$

$$\eta_{ij} = \begin{array}{@{}c@{}l} \begin{bmatrix} 0 \\ \vdots \\ \mathbf{y} \\ 0 \\ \vdots \\ -\mathbf{y} \\ 0 \\ \vdots \\ 0 \end{bmatrix} & \begin{array}{l} \\ \\ \Rightarrow i \\ \\ \\ \Rightarrow j \\ \\ \\ \\ \end{array} \end{array}$$

$$\boxed{\begin{array}{c} \hat{\alpha}^t \eta_{ij} > 0 \\[2em] \mathbf{a}_i^t \mathbf{y}_k - \mathbf{a}_j^t \mathbf{y}_k > 0 \end{array}}$$

Support Vector Machine

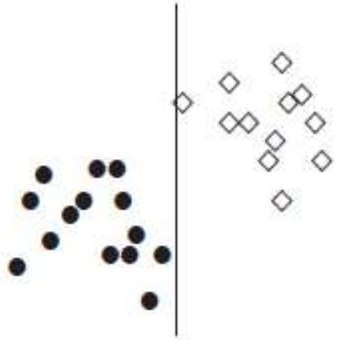
SVM

Support vector machine(Maximum marginal classification) Webb (P.249)

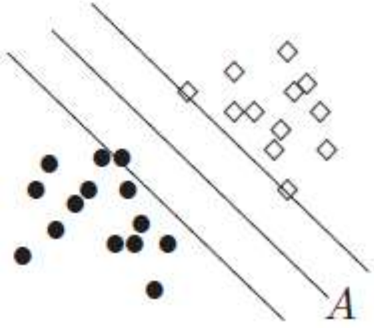
$$g(x) = w^T x + w_0$$

$$w^T x + w_0 \begin{cases} > & 0 \\ < & 0 \end{cases} \Rightarrow x \in \begin{cases} \omega_1 \text{ with corresponding numeric value, } y_i = +1 \\ \omega_2 \text{ with corresponding numeric value, } y_i = -1 \end{cases}$$

$y_i(w^T x_i + w_0) > 0$ for all i



(a)



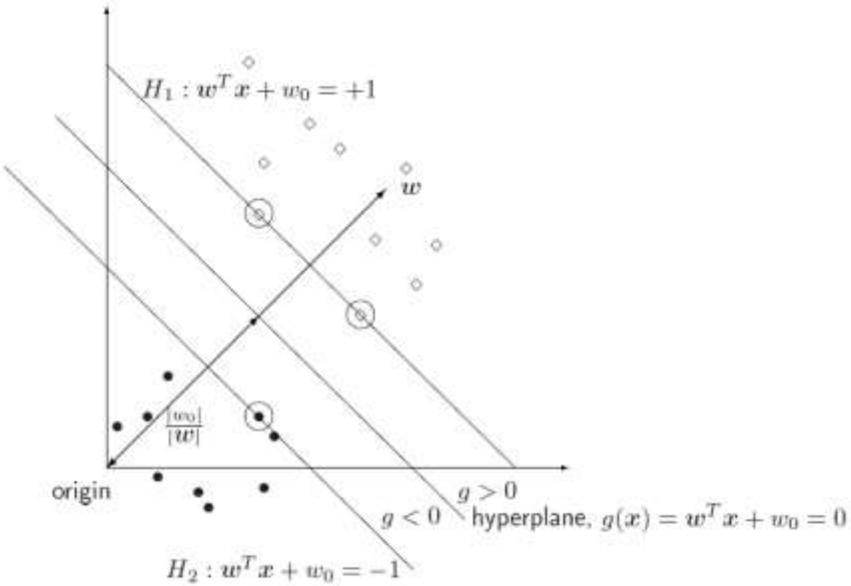
(b)

$$y_i(w^T x_i + w_0) \geq b$$

$$\begin{aligned} w^T x_i + w_0 &\geq +1 \quad \text{for } y_i = +1 \\ w^T x_i + w_0 &\leq -1 \quad \text{for } y_i = -1 \end{aligned}$$

$$\min \ \frac{1}{2} \| W \|^2$$

$$y_i(w^T x_i + w_0) \geq 1 \quad i = 1, \ldots, n$$



Convex Optimization

Optimality Conditions

$$\frac{\partial L_p}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i = 0$$

$$\frac{\partial L_p}{\partial w_0} = -\sum_{i=1}^{n} \alpha_i y_i = 0$$

$$y_i(\boldsymbol{x}_i^T \boldsymbol{w} + w_0) - 1 \geq 0$$

$$\alpha_i \geq 0$$

$$\alpha_i(y_i(\boldsymbol{x}_i^T \boldsymbol{w} + w_0) - 1) = 0$$

Dual Problem

$$L_p = \frac{1}{2} w^T w - \sum_{i=1}^{n} \alpha_i (y_i (w^T x_i + w_0) - 1)$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

Dual Form

max $L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$

s.t

$$\alpha_i \geq 0 \qquad \sum_{i=1}^{n} \alpha_i y_i = 0$$

Discrimination

W?

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

W0 : By Slackness

$$\alpha_i (y_i (x_i^T w + w_0) - 1) = 0$$

$$\alpha_i = 0$$

$$\alpha_i \neq 0 \Rightarrow (y_i (x_i^T w + w_0) - 1) = 0$$

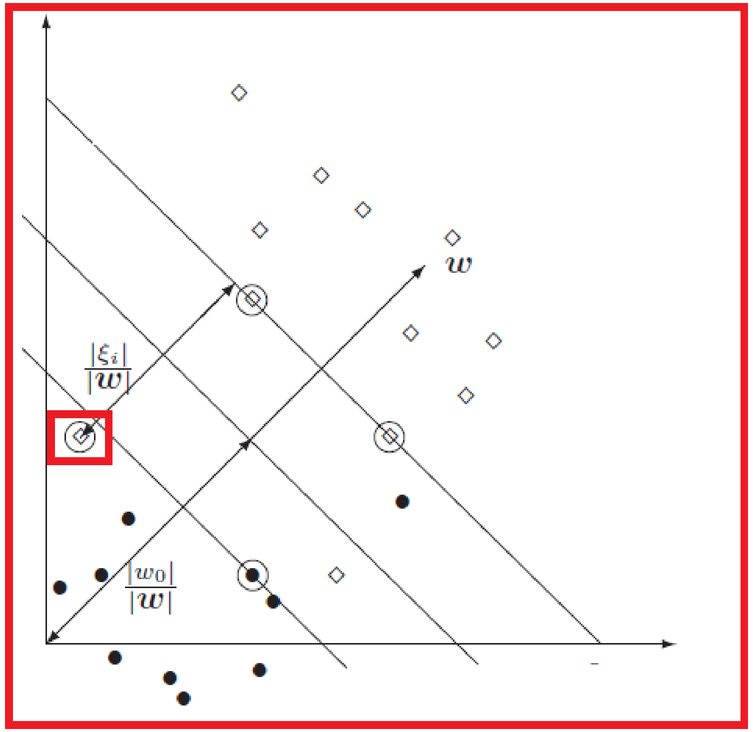
Support Vector

$$n_{\mathcal{SV}} w_0 + w^T \sum_{i \in \mathcal{SV}} x_i = \sum_{i \in \mathcal{SV}} y_i$$

$$w^T x + w_0 = \sum_{i \in \mathcal{SV}} \alpha_i y_i x_i^T x - \frac{1}{n_{\mathcal{SV}}} \sum_{i \in \mathcal{SV}} \sum_{j \in \mathcal{SV}} \alpha_i y_i x_i^T x_j + \frac{1}{n_{\mathcal{SV}}} \sum_{i \in \mathcal{SV}} y_i$$

$$\sum_{i \in \mathcal{SV}} \alpha_i y_i x_i^T x - \frac{1}{n_{\mathcal{SV}}} \sum_{i \in \mathcal{SV}} \sum_{j \in \mathcal{SV}} \alpha_i y_i x_i^T x_j + \frac{1}{n_{\mathcal{SV}}} \sum_{i \in \mathcal{SV}} y_i > 0 \quad \Rightarrow \quad \text{assign } \boldsymbol{x} \text{ to } \omega_1$$

SVM for Linear non-separable



Constraints:

$$\boxed{\begin{array}{ll} w^T x_i + w_0 \geq +1 - \xi_i & \text{for } y_i = +1 \\ w^T x_i + w_0 \leq -1 + \xi_i & \text{for } y_i = -1 \\ \xi_i \geq 0 & i = 1, \ldots, n \end{array}} \Rightarrow \boxed{\begin{array}{ll} y_i(w^T x_i + w_0) \geq 1 - \xi_i & i = 1, \ldots, n \\ \xi_i \geq 0 & i = 1, \ldots, n \end{array}}$$

Objective function

$$\boxed{\frac{1}{2} w^T w + C \sum_i \xi_i}$$

Dual Problem

Max $L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$

s.t

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C$$

Complementarity conditions

$$\alpha_i(y_i(x_i^T w + w_0) - 1 + \xi_i) = 0$$

$$r_i\xi_i = (C - \alpha_i)\xi_i = 0$$

Patterns for which $\alpha_i > 0$ are termed the support vectors

$$y_i(x_i^T w + w_0) - 1 + \xi_i = 0$$

$0 < \alpha_i < C \Rightarrow \xi_i = 0$

$\xi_i \neq 0 \Rightarrow \alpha_i = C$

$\Downarrow$

x_i are misclassified if $\xi_i > 1$.

If $\xi_i < 1$, they are classified correctly, but lie closer to the separating hyperplane than $1/|w|$

$\mathcal{SV}$ is the set of support vectors with associated values of α_i satisfying $0 < \alpha_i \leq C$

$\widetilde{\mathcal{SV}}$ is the set of $n_{\widetilde{\mathcal{SV}}}$ support vectors satisfying $0 < \alpha_i < C$

$$\sum_{i \in \mathcal{SV}} \alpha_i y_i \boldsymbol{x}_i^T \boldsymbol{x} + \frac{1}{n_{\widetilde{\mathcal{SV}}}} \left\{ \sum_{j \in \widetilde{\mathcal{SV}}} y_j - \sum_{i \in \mathcal{SV}, j \in \widetilde{\mathcal{SV}}} \alpha_i y_i \boldsymbol{x}_i^T \boldsymbol{x}_j \right\} > 0$$