Lecture 14

MFML Team

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Q4 Answer the following questions with justifications.

The figure below shows 4 points, representing some data in $\mathbb{R}^2$

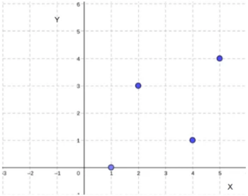


FIGURE 1. PCA

(A) Find the principal components for the given points.

NOTE: use

$$cov(X) = \frac{1}{N}(X-\mu)^T(X-\mu)$$

(5 marks)

(B) Find the components of the four points along their first principal component.

(2 marks)

(C) What is the percentage variance captured by the first principal component?

(1 mark)

(D) If the points are rotated anticlockwise by 90 degrees, what will the components (of the rotated points) along their first principal component be?

(2 marks)

Q4

The figure below shows 4 points, representing some data in $\mathbb{R}^2$

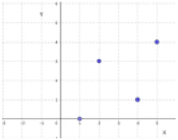


FIGURE 1. PCA

(A) The four points are $(2,3)$, $(4,1)$, $(5,4)$ and $(1,0)$ This data X is represented as:

$$\mathbf{X} = \begin{bmatrix} 2 & 4 & 5 & 1 \\ 3 & 1 & 4 & 0 \end{bmatrix}$$

$$\mu = \begin{bmatrix} 12/4 \\ 8/4 \end{bmatrix} = \begin{bmatrix} 3 \\ 2 \end{bmatrix}$$

(1 mark)

$$\mathbf{X} - \mu = \begin{bmatrix} -1 & 1 & 2 & -2 \\ 1 & -1 & 2 & -2 \end{bmatrix}$$

$$\mathbf{cov(X)} = \frac{\mathbf{1}}{\mathbf{4}} \begin{bmatrix} -1 & 1 & 2 & -2 \\ 1 & -1 & 2 & -2 \end{bmatrix} \begin{bmatrix} -1 & 1 \\ 1 & -1 \\ 2 & 2 \\ -2 & -2 \end{bmatrix} = \frac{\mathbf{1}}{\mathbf{4}} \begin{bmatrix} 10 & 6 \\ 6 & 10 \end{bmatrix}$$

(2 marks)

To find eigen values and eigen vectors:
eigen values:

$$\lambda^2 - 5\lambda + 4 = 0 \implies \lambda = 4, 1$$

eigen vectors:

$$\lambda = 4 \implies v_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$\lambda = 1 \implies v_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

Principal component directions:

$$\begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

and

$$\begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$

(2 marks)

(B) The components along first PC are:

$$\hat{x_1} = x_1^T * e_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 2 & 3 \end{bmatrix} * \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{5}{\sqrt{2}}$$

$$\hat{x_2} = \frac{5}{\sqrt{2}}$$

$$\hat{x_3} = \frac{9}{\sqrt{2}}$$

$$\hat{x_4} = \frac{1}{\sqrt{2}}$$

(2 marks)

(C) Percentage variance captured by first component $= \frac{4}{4+1} = 0.8$ (i.e., 80%)

(1 mark)

(D) Rotating all the points by same angle does not affect the components along the principal component. It will be smae as the answer in part (B)

(2 marks)

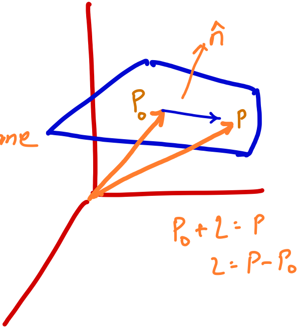
Agenda

Mathematical preliminaries for Support Vector Machines

→ Properties of planes + overview of SVM

- Constrained optimization and Lagrange multipliers.
- Primal and dual problems and how their solutions are related
- Karash-Kuhn-Tucker conditions.
- Definition of Kernel Functions
- Linear Classifiers

Equation of normal to a plane



$P_o \rightarrow$ a specific point on plane

$P \rightarrow$ any point on plane

$P - P_o \rightarrow$ line completely on the plane

$\hat{n} \rightarrow \perp^r$ dir. to plane

$$\Rightarrow \quad (P - P_o) \cdot \hat{n} = 0$$

$$\begin{bmatrix} x - x_o & y - y_o & z - z_o \end{bmatrix} \begin{bmatrix} n_x \\ n_y \\ n_z \end{bmatrix} = 0$$

$$(x - x_o) n_x + (y - y_o) n_y + (z - z_o) n_z = 0$$

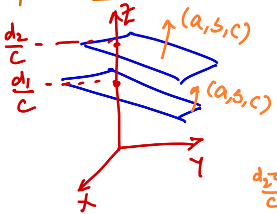
$$\underbrace{n_x}_{a} x + \underbrace{n_y}_{b} y + \underbrace{n_z}_{c} z + \underbrace{(-n_x x_o - n_y y_o - n_z z_o)}_{-d} = 0$$

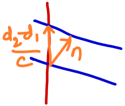
Plane eqn: $\quad ax + by + cz + d = 0$

$\therefore \quad n \equiv [a \quad b \quad c] \leftarrow \perp^r$ to plane

Point of intersection on z-axis = $\dfrac{d}{n_z}$, ...

Distance between two planes





$P1: \; ax + by + cz = d_1$

$P2: \; ax + by + cz = d_2$

$$\hat{n} = \frac{(a \quad b \quad c)}{\sqrt{a^2+b^2+c^2}}$$

$P1: \; x=0 \;\; y=0 \;\Rightarrow\; z = \frac{d_1}{c}$

$P2: \; x=0 \;\; y=0 \;\Rightarrow\; z = \frac{d_2}{c}$

$$\text{distance between planes} = \begin{bmatrix} 0 & 0 & \frac{d_2-d_1}{c} \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} \frac{1}{\sqrt{a^2+b^2+c^2}} = \frac{d_2-d_1}{\sqrt{a^2+b^2+c^2}}$$

For planes in higher dim:

$P1: \; a_1x + a_2y + a_3z + \; \cdots\cdots = d_1$

$P2: \; a_1x + a_2y + a_3z + \; \cdots\cdots = d_2$

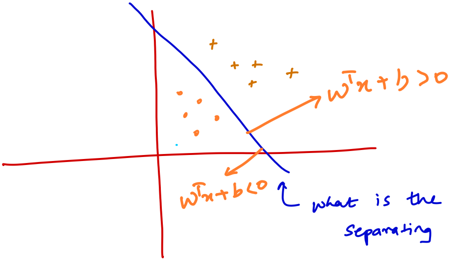
$$\text{distance b/w planes} = \frac{d_2-d_1}{\sqrt{a_1^2+a_2^2+\cdots\cdots}} = \frac{d_2-d_1}{\|a\|}$$

data	x_1	x_2	x_3	$\cdots$	x_N
feature 1	x_{11}	x_{12}	$\cdots$		
feature 2	x_{21}	x_{22}	$\cdots$		
$\vdots$	$\vdots$				
feature D					
label	$+1$	$+1$	-1	$\cdots$	-1

$\mathcal{X} = \{x_1, x_2 \ldots x_n\}$

$x_i \in \mathbb{R}^D$

$\leftarrow$ Supervised Binary Classification



Goal:

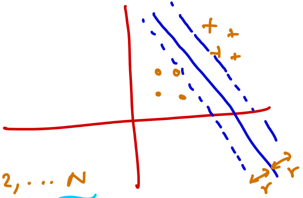
find w, b s.T $w^T x_i + b > 0$ for $y_i = +1$

and $w^T x_i + b < 0$ for $y_i = -1$

$$\Rightarrow \quad y_i (w^T x_i + b) \geq 0 \quad \text{for } i = \underbrace{1, 2, \ldots N}_{\text{dataset}}$$

Add a margin

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



Let $r = 1$

$$y_i(w^T x_i + b) \geq 1 \quad \text{for } i = \underbrace{1, 2, \dots N}_{\text{dataset}}$$

For +ve sample on boundary:

$$w^T x + b = 1$$

For -ve sample on boundary:

$$w^T x + b = -1$$

$$\Rightarrow \begin{aligned} a_1 x_1 + a_2 x_2 + \dots &= d_1 \\ a_1 x_1 + a_2 x_2 + \dots &= d_2 \end{aligned}$$

Distance between these planes $= \dfrac{(1-b) - (-1-b)}{\sqrt{w_1^2 + w_2^2 + \dots}} = \dfrac{2}{\|w\|}$

Goal:

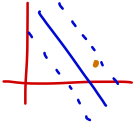
maximize $\dfrac{2}{\|w\|}$ subj to: $y_i(w^T x_i + b) \geq 1 \quad i = 1, 2, \dots N$

or $\boxed{\min \frac{1}{2}\|w\|^2 \quad \text{subj to} \quad y_i(w^T x_i + b) \geq 1 \quad i = 1, 2, \dots N}$ (1)

HW

Solution is unique

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i \quad \text{subj to} \quad y_i(w^T x_i + b) \geq 1 - \xi_i \quad -(2)$$
$$\xi_i \geq 0 \qquad i = 1, 2, \ldots N$$



$0 < \xi_i < 1 \Rightarrow$ for i^{th} sample, we are relaxing the constraint.
The sample can be inside the boundary

$\xi_i > 1 \Rightarrow$ Sample is misclassified.

$\therefore$ C now acts as a regularization term.

C is large $\Rightarrow$ all ξ_i are small (as we are minimizing $C\xi_i$)
$\Rightarrow$ all samples perfectly classified $\rightarrow$ Hard margin classifier.

C is small $\Rightarrow$ ξ_i can be large to allow some misclassifications
$\rightarrow$ soft margin classifier.

② ⇒ $$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\left[1-y_i(w^Tx_i+b)\right]_+ \quad — ③$$

Both terms are convex.

where $[z]_+ = \max(0, z)$

if $\underbrace{y_i(w^Tx_i+b) > 1}$ ⇒ $\underbrace{\left[1-\underbrace{y_i(w^Tx_i+b)}_{>1}\right]_+}_{<1} = 0$

↳ means correctly classified.

if $\underbrace{y_i(w^Tx_i+b) \leq 1}$ ⇒ $\underbrace{\left[1-\underbrace{y_i(w^Tx_i+b)}_{<1}\right]_+}_{>1} = \xi$

↓ means incorrectly classified (or inside boundary)

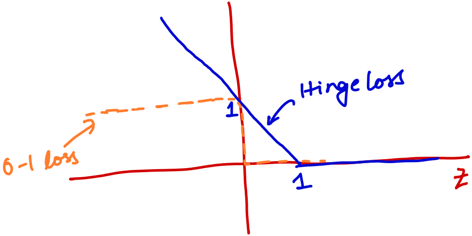
when $\xi > 0$

$y_i(w^Tx_i+b) = 1-\xi$

∴ second term in ③ (Regularization term) comes into picture for misclassified samples only.

$[1-z]_+ = \max(0, 1-z)$ ← Hinge loss

z	$1-z$	$[\;]_+$
-2	3	3
-1	2	2
0	1	1
1	0	0
2	-1	0
		⋮



0-1-Loss:

$y_i(w^Tx_i + b) < 0 \rightarrow$ loss = 1 wrong

$y_i(w^Tx_i + b) \geq 0 \rightarrow$ loss = 0 correct

} difficult to optimize

(*) So, (3) can be thought of:

balancing between Hinge loss term and a L2 Regularization (ie; $\frac{1}{2}\|w\|^2$)

Optimization problem

We shall work with the following optimization problem:

Objective fn $\longrightarrow$ $\min f(\boldsymbol{x})$ subject to

$g_i(\boldsymbol{x}) \leq 0 \; \forall i \in [m]$ $\longleftarrow$ inequality

$h_j(\boldsymbol{x}) = 0 \; \forall j \in [p]$ $\longleftarrow$ equality

constraint.

FOR SVM :

$\underbrace{\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i}_{f(x)}$ subj to $y_i(w^T x_i + b) \geq 1-\xi_i$ — $\longleftarrow g_1(x)$ (N eqns)

$\underbrace{\xi_i \geq 0}$ $\quad i = 1, 2, \ldots N$ $\longleftarrow g_2(x)$ (N eqns)

No $h_j(x)$

Optimization problem : Lagrangian

The Lagrangian associated with this optimization problem is

- $$\min f(\boldsymbol{x}) + \sum_{i=1}^{i=m} \lambda_i g_i(\boldsymbol{x}) + \sum_{j=1}^{j=p} \nu_j h_j(\boldsymbol{x})$$

- The λ_i's and $\overset{\nu_j}{\cancel{h_j}}$'s are called Lagrange multipliers.

$$\mathcal{L}(x, \lambda_i, v_j) = f(x) + \sum_{i=1}^{m} \underset{\geq 0}{\lambda_i} \underbrace{g_i(x)}_{\leq 0} + \underbrace{\sum_{j=1}^{p} v_j h_j(x)}_{0}$$

when x is feasible for ⊛ slide 12

$\underbrace{\lambda_i g_i(x)}_{\leq 0}$

$$\Rightarrow \quad \mathcal{L}(x_0, \lambda_i, v_j) \leq f(x_0)$$

$$\underbrace{\min_{x} \mathcal{L}(x, \lambda_i, v_i)}_{g(\lambda_i, v_i)} \leq f(x^*)$$

x^* is solution

as $g(\lambda, v)$ is Lower bound for $f(x^*)$ we can try to maximize it to find x^* (dual)

i.e;
$$\max_{\lambda, v} g(\lambda, v)$$
$$\lambda_i \geq 0 \quad i = 1..N$$

(NOTE ⊛ had $2N$ constraints)

weak duality: $g(\lambda^*, v^*) \leq f(x^*)$

Strong duality exists if :

- $f(x)$ is Convex
- $g_i(x)$ is Convex
- $h_i(x)$ are linear
- mild Regularity Condition holds

↳ one such is Slater condition

→ atleast one point exists in feasible region where

$g_i(x) < 0$ ← strict inequality

& $h_i(x) = 0$

Quadratic programming

Consider the following primal problem:

- We now consider the case of a quadratic objective function subject to affine constraints:

$$\min_{\boldsymbol{x}\in\mathbb{R}^d} \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{c}^T\boldsymbol{x}$$

$$\text{subject to } \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$$

- Here $\boldsymbol{A} \in \mathbb{R}^{m\times d}$, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^d$

Quadratic programming

- The Lagrangian $\mathfrak{L}(\boldsymbol{x}, \boldsymbol{\lambda})$ is given by $\frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{c}^T\boldsymbol{x} + \boldsymbol{\lambda}^T(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})$.
- Rearranging the above we have $\mathfrak{L}(\boldsymbol{x}, \boldsymbol{\lambda}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} + (\boldsymbol{c} + \boldsymbol{A}^T\boldsymbol{\lambda})^T\boldsymbol{x} - \boldsymbol{\lambda}^T\boldsymbol{b}$
- Taking the derivative of $\mathfrak{L}(\boldsymbol{x}, \boldsymbol{\lambda})$ and setting it equal to zero gives $\boldsymbol{Q}\boldsymbol{x} + (\boldsymbol{c} + \boldsymbol{A}^T\boldsymbol{\lambda}) = 0$.

Quadratic programming

We will now derive the dual problem

- If we take $\boldsymbol{Q}$ to be invertible, we have $\boldsymbol{x} = \boldsymbol{Q}^{-1}(\boldsymbol{c} + \boldsymbol{A}^T \boldsymbol{\lambda})$.
- Plugging this value of $\boldsymbol{x}$ into $\mathfrak{L}(\boldsymbol{x}, \boldsymbol{\lambda})$ gives us $\mathfrak{D}(\boldsymbol{\lambda}) = -\frac{1}{2}(\boldsymbol{c} + \boldsymbol{A}^T \boldsymbol{\lambda})\boldsymbol{Q}^{-1}(\boldsymbol{c} + \boldsymbol{A}^T \boldsymbol{\lambda}) - \boldsymbol{\lambda}^T \boldsymbol{b}$.
- This gives us the dual optimization problem: $\max_{\boldsymbol{\lambda} \in \mathbb{R}^m} -\frac{1}{2}(\boldsymbol{c} + \boldsymbol{A}^T \boldsymbol{\lambda})\boldsymbol{Q}^{-1}(\boldsymbol{c} + \boldsymbol{A}^T \boldsymbol{\lambda}) - \boldsymbol{\lambda}^T \boldsymbol{b}$ subject to $\boldsymbol{\lambda} \geq \boldsymbol{0}$.

Dual

Summary

The original problem is :

$$\min_{\boldsymbol{x} \in \mathbb{R}^d} \frac{1}{2} \boldsymbol{x}^T \boldsymbol{Q} \boldsymbol{x} + \boldsymbol{c}^T \boldsymbol{x}$$

$$\text{subject to } \boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{b}$$

The dual problem is

$$\max_{\boldsymbol{\lambda} \geq 0} -\frac{1}{2}(\boldsymbol{c} + \boldsymbol{A}^T \boldsymbol{\lambda}) \boldsymbol{Q}^{-1} (\boldsymbol{c} + \boldsymbol{A}^T \boldsymbol{\lambda}) - \boldsymbol{\lambda}^T \boldsymbol{b}$$

Weak duality

- Weak duality establishes an inequality connecting primal and dual problems
- Weak duality condition states that the optimal solution of the primal problem is greater than or equal to that of the dual problem.
- In the Quadratic Optimization problem discussed previously , weak duality exists

Strong duality

- Strong duality condition states that the optimal solution of the primal problem is equal to that of the dual problem
- One can solve the dual problem to get the same solution as solving the primal problem.
- In some optimization problems, solving the dual problem may be easier.
- Question: When does strong duality hold?

Slater's condition

- For a primal optimization problem we say that it obeys Slater's condition if
  1. the objective function f is convex, the constraints g_i are all convex ,the contraint functions h_i are all linear
  2. there exists a point $\bar{\boldsymbol{x}}$ in the interior of the region, i.e $g_i(\bar{\boldsymbol{x}}) < 0$ for all $i \in [m]$ and $h_j(\bar{\boldsymbol{x}}) = 0$ for all $j \in [p]$.
- Suppose Slater's condition holds then we have strong duality.
- Strong duality condition states that the optimal solution of the primal problem is equal to that of the dual problem

Example of Slater's condition

We will consider an optimization problem as given below

$$\min \; x^2 + y^2$$
$$\text{st } x + y - 3 \leq 0$$

- Here $f(x, y) = x^2 + y^2$ is a convex function and $g(x, y) = x + y - 3$ is a convex function
- We can find a point that satisfies the condition $x + y - 3 < 0$
- Slaters condition is satisfied

When strong duality holds., $\underline{(x^*, \lambda^*, \nu^*)}$ satisfy KKT

Primal is convex & Slater's condition is satisfied

$\hookrightarrow$ x^* is primal sol. λ^*, ν^* are dual sol^n.

if primal is:

$$\min_x f(x) \quad \text{subj to} \quad g_i(x) \leq 0 \quad i = 1, 2, \ldots m$$
$$h_j(x) = 0 \quad j = 1, 2, \ldots p$$

then K.K.T for (x^*, λ^*, ν^*) is:

1) Primal feasibility: $g_i(x^*) \leq 0$ & $h_j(x^*) = 0$ $i = [m], j = [p]$

2) Dual feasibility: $\lambda_i^* \geq 0$ $i = [m]$

3) Complementary slackness: $\lambda_i^* g_i(x^*) = 0$ $[m]$

$\Rightarrow$ either $\lambda_i^* = 0$ or $f_i(x^*) = 0$

4) Vanishing gradients (stationarity):

$$\nabla \mathcal{L} : \nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(x^*) + \sum_{j=1}^{p} \nu_j^* \nabla h_j(x^*) = 0$$

KKT conditions

$$\min\ f(\boldsymbol{x}) \quad \text{st} \quad g_i(\boldsymbol{x}) \leq 0\ \forall i \in [m], \quad h_j(\boldsymbol{x}) = 0\ \forall j \in [p]$$

We say that $\boldsymbol{x}^*$ and $(\lambda^*, \nu^*) \in \mathbb{R}^m \times \mathbb{R}^p$ respect the Karash-Kuhn-Tucker conditions if:

1. $g_i(\boldsymbol{x}^*) \leq 0\ \forall i \in [m]$, $h_i(\boldsymbol{x}^*) = 0\ \forall i \in [p]$.
2. $\lambda_i^* \geq 0\ \forall i \in [m]$.
3. $\lambda_i^* g_i(\boldsymbol{x}^*) = 0\ \forall i \in [m]$.
4. $\nabla f(\boldsymbol{x}^*) + \sum_{i=1}^{i=m} \lambda_i^* \nabla g_i(\boldsymbol{x}^*) + \sum_{i=1}^{i=p} \nu_i^* \nabla h_i(\boldsymbol{x}^*) = 0$.

If strong duality holds then any primal optimal solution $\boldsymbol{x}^*$ and dual optimal solution (λ^*, ν^*) satisfy the KKT conditions.

KKT condition

We will consider an optimization problem and will write its KKT conditions

$$\min \; x^2 + y^2$$
$$\text{st } x + y - 3 \leq 0$$

- Here $f(x, y) = x^2 + y^2$ and $g(x, y) = x + y - 3$

1. $x + y - 3 \leq 0$
2. $\lambda \geq 0$
3. $\lambda(x + y - 3) = 0$
4. $\nabla f + \lambda \nabla g = \mathbf{0}$

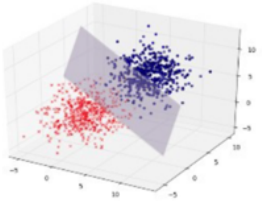
Classification Problem in Machine Learning

- Classification of data into different classes is one of the primary problems in machine learning
- Binary classification involves classifying data into exactly 2 classes
- There exists different algorithms for binary classification
- We will discuss a model called Support Vector Machine.
- SVM is a linear classifer model for binary classification

Linear Classifier

$$\mathbf{w}^T\mathbf{x} = 0$$

Hyperplane



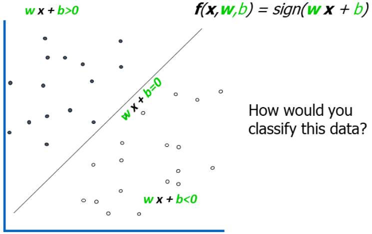
$$y = ax + b$$

Line

Linear Classifier and Hyperplane

- Consider line $w^T x + b = 0$. Let x_a and x_b lie on this line. So $w^T x_a + b = 0$ and $w^T x_b + b = 0$.
- This means $w^T (x_a - x_b) = 0$. $x_a - x_b$ lies on the line and is directed from x_b to x_a.
- Hence w is orthogonal to $x_a - x_b$ and in turn, to the line.

Linear Classifiers

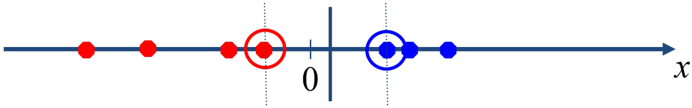


- denotes +1
- denotes -1

How would you classify this data?

Two examples of data

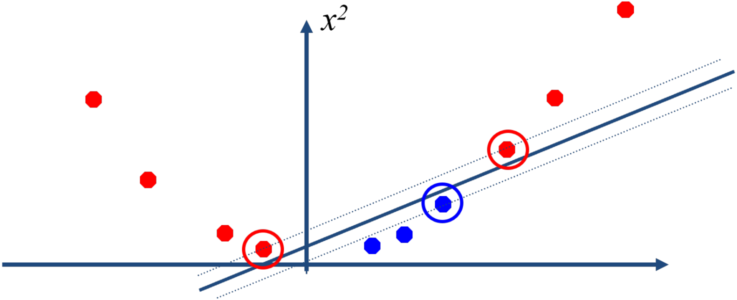
Dataset that are linearly separable with some noise



Dataset is not linearly separable

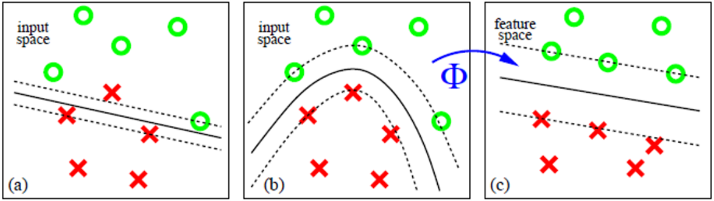
Mapping of Data

mapping data to a higher-dimensional space:



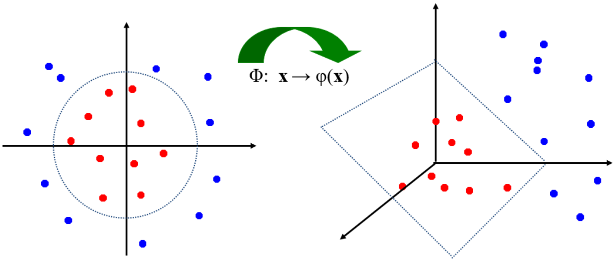
Mapping of Data

Find a feature space



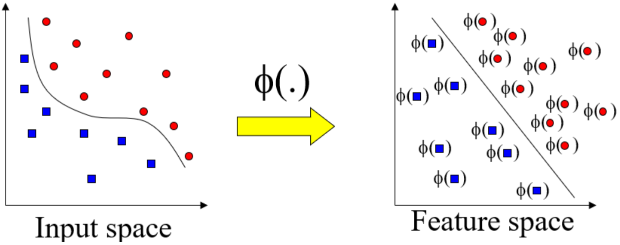
If every data point is mapped into high-dimensional space via some transformation $\phi : x \rightarrow \phi(x)$

Feature spaces



- General idea: the original input space can always be mapped to some higher-dimensional feature space where the training set is separable.

Transforming the Data



- Computation in the feature space can be costly because it is high dimensional.
- The feature space is typically infinite-dimensional.
- The kernel trick using kernel functions comes to rescue

Kernel Functions

Kernel is a continuous function $K(x, y)$

Kernel takes two arguments x and y

x and y could be real numbers, functions, vectors, etc

$K(x, y)$ maps x and y to a real value

Kernel value is independent of the order of the arguments, i.e.,

$$K(x, y) = K(y, x)$$

later

Kernel Functions

later

- A kernel function is some function that corresponds to an inner product in some expanded feature space.

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$$

- Linear classifier relies on dot product between vectors $x_i^T x_j$
- If every data point is mapped into high-dimensional space via some transformation $\phi : x \rightarrow \phi(x)$, the dot product becomes: $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$
- For some functions $K(x_i, x_j)$ checking $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ is difficult.
- Mercer's theorem: Every positive-semidefinite symmetric function is a kernel function.

Kernel Functions Construction

later

) We can *construct kernels from scratch*:

- For any $\varphi : \mathcal{X} \to \mathbb{R}^m$, $k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathbb{R}^m}$ is a kernel.
- If $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a *distance function*, i.e.
  - $d(x, x') \geq 0$ for all $x, x' \in \mathcal{X}$,
  - $d(x, x') = 0$ only for $x = x'$,
  - $d(x, x') = d(x', x)$ for all $x, x' \in \mathcal{X}$,
  - $d(x, x') \leq d(x, x'') + d(x'', x')$ for all $x, x', x'' \in \mathcal{X}$,

  then $k(x, x') := \exp(-d(x, x'))$ is a kernel.

Kernel Functions Construction

2) We can *construct kernels from other kernels*:

- if k is a kernel and $\alpha > 0$, then αk and $k + \alpha$ are kernels.
- if k_1, k_2 are kernels, then $k_1 + k_2$ and $k_1 \cdot k_2$ are kernels.

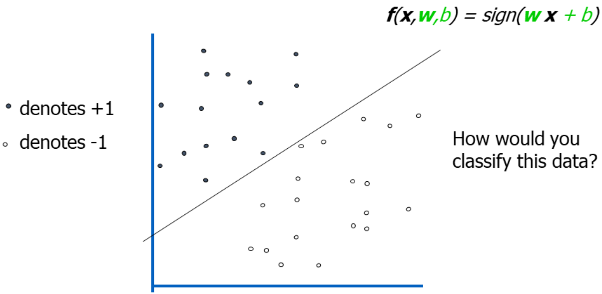
Examples of Kernels

- Linear: $K(x_i, x_j) = x_i^T x_j$
- Polynomial of power p : $K(x_i, x_j) = (1 + x_i^T x_j)^p$
- Sigmoid: $K(x_i, x_j) = tanh(\beta_0 x_i^T x_j + \beta_1)$

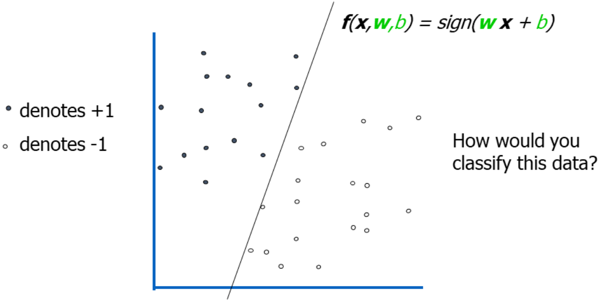
[illegible]

Linear Classifiers Revisited

Linear Classifiers



Linear Classifiers

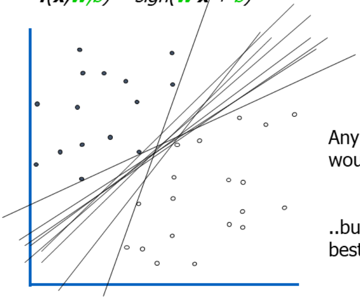


Linear Classifier

Linear Classifiers

$f(\mathbf{x},\mathbf{w},b) = sign(\mathbf{w}\,\mathbf{x} + b)$

- denotes +1
- denotes -1



Any of these would be fine..

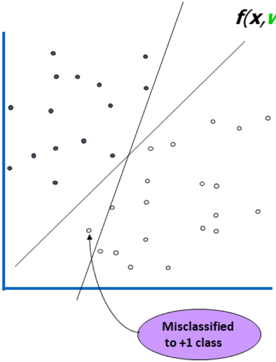
..but which is best?

Linear Classifier

Linear Classifiers

$$f(\mathbf{x},\mathbf{w},b) = sign(\mathbf{w}\,\mathbf{x} + b)$$

- • denotes +1
- ∘ denotes -1



How would you classify this data?