CS540 Introduction to Artificial Intelligence
Lecture 5

Young Wu
Based on lecture slides by Jerry Zhu, Yingyu Liang, and Charles Dyer

June 1, 2020

Survey Question

Admin

Socrative room : CS540E

student login → CS540E

enter ID

- Which prerecorded lecture videos have you watched?
- A: Yes
- B: Lectures $1, 2, 3, 4, 5, 6$
- C: Lectures $1, 2, 3, 4$
- D: Lectures $1, 2$
- E: No

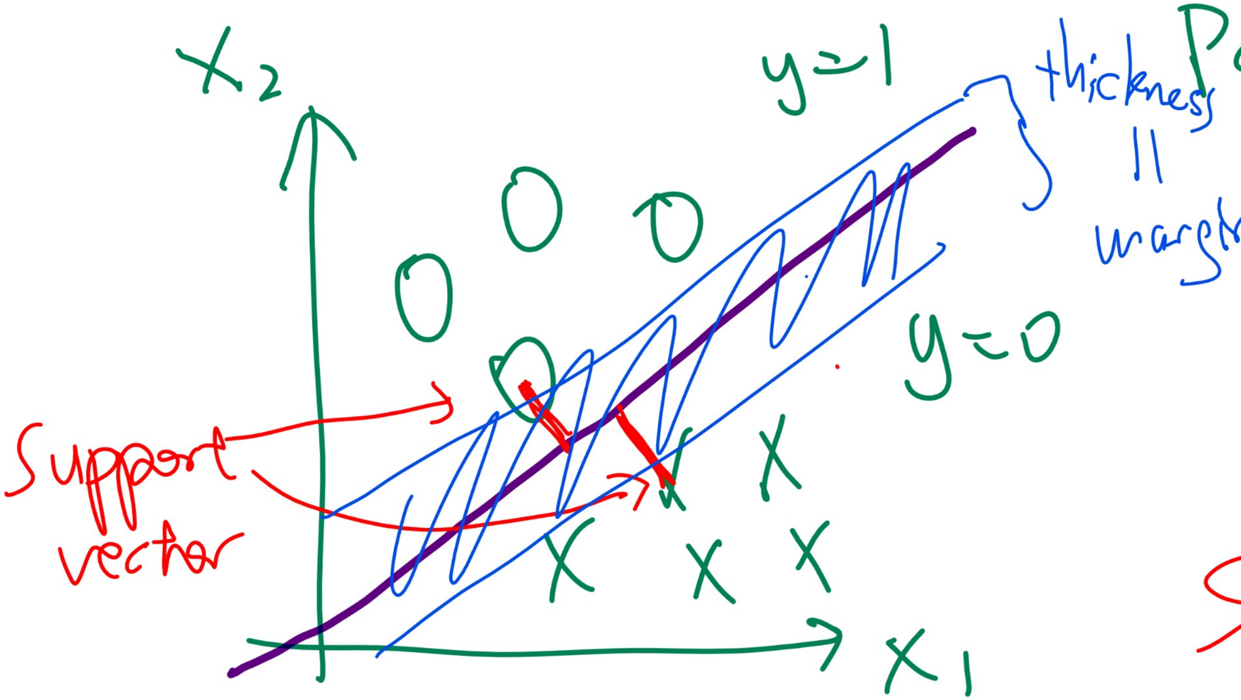
Schedule

Admin

- Week 2 Examples and Quiz questions on Week 4
- Week 3 SVM and DTree

Maximum Margin Diagram

Motivation



Perceptron
Logistic

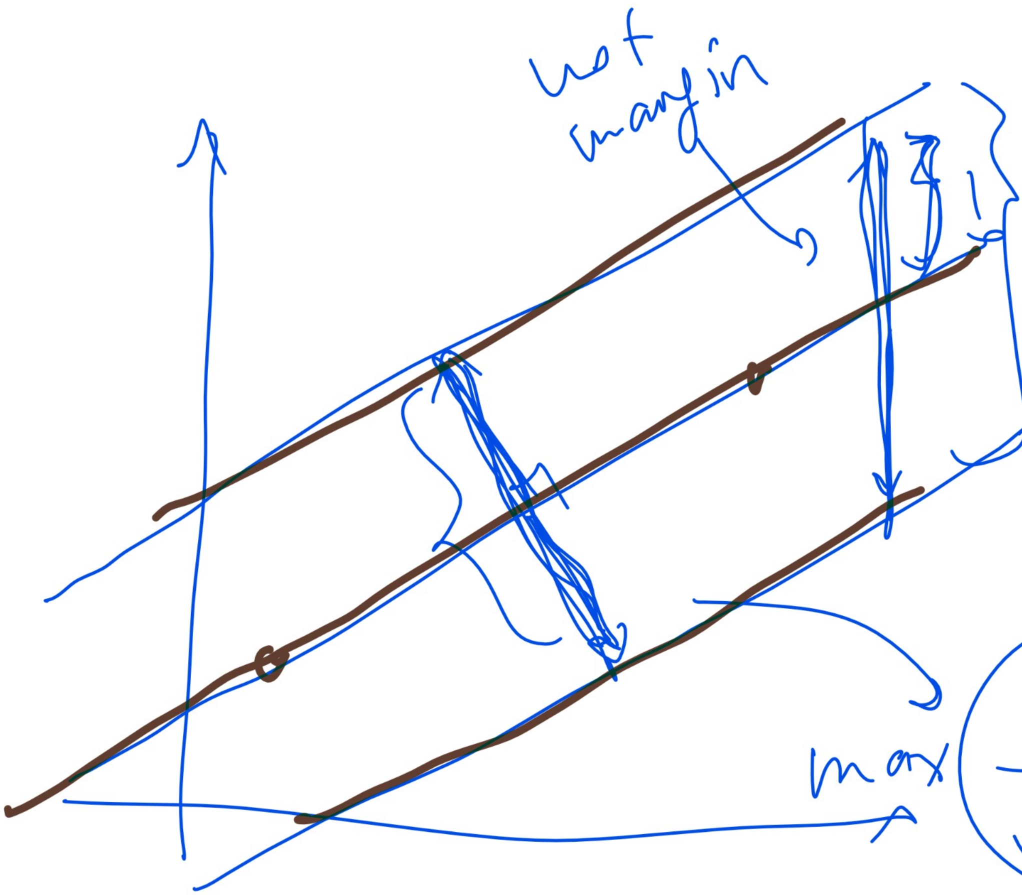
min Loss/Cost

SVM

max margin

Constrained Optimization Derivation

Definition



$w^T x + b = +1$

$w^T x + b = 0$

$w^T x + b = -1$

$\max \dfrac{2}{\sqrt{w^T w}}$

$y = 1$ points above $+1$ plane line

$y = 0$ points below -1 plane

Constrained Optimization

Definition

- The goal is to maximize the margin subject to the constraint that the plus plane and the minus plane separates the instances with $y_i = 0$ and $y_i = 1$.

$$\max_w \frac{2}{\sqrt{w^T w}} \text{ such that } \begin{cases} \left(w^T x_i + b\right) \leqslant -1 & \text{if } y_i = 0 \\ \left(w^T x_i + b\right) \geqslant 1 & \text{if } y_i = 1 \end{cases}, i = 1, 2, ..., n$$

- The two constrains can be combined.

$$\max_w \frac{2}{\sqrt{w^T w}} \text{ such that } (2y_i - 1)\left(w^T x_i + b\right) \geqslant 1, i = 1, 2, ..., n$$

Hard Margin SVM

Definition

$$\max_{w} \frac{2}{\sqrt{w^T w}} \text{ such that } (2y_i - 1)\left(w^T x_i + b\right) \geqslant 1, i = 1, 2, ..., n$$

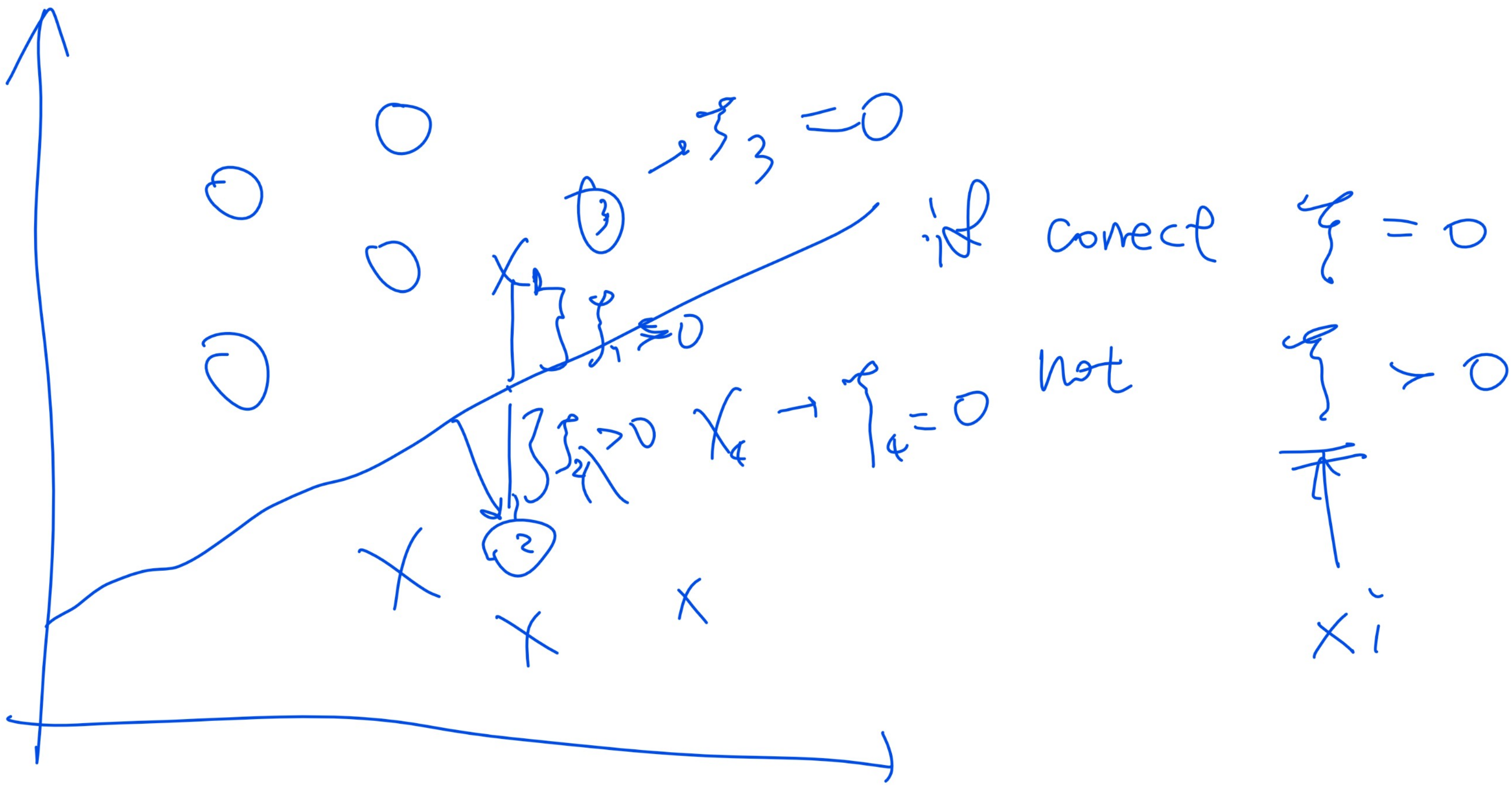
positive

- This is equivalent to the following minimization problem, called hard margin SVM.

$$\min_{w} \frac{1}{2} w^T w \text{ such that } (2y_i - 1)\left(w^T x_i + b\right) \geqslant 1, i = 1, 2, ..., n$$

Soft Margin Diagram

Definition



Soft Margin SVM

Definition

max ~~margin~~ min total amount of mistake

Loss

$$\min_w \frac{1}{2} w^T w + \frac{1}{\lambda} \frac{1}{n} \sum_{i=1}^{n} \xi_i$$

such that $(2y_i - 1)\left(w^T x_i + b\right) \geqslant 1 - \xi_i, \xi_i \geqslant 0, i = 1, 2, \ldots, n$

$\xi_i \geq 1 - (2y_i - 1)(w^T x_i + b)$, $\xi_i \geq 0$

- This is equivalent to the following minimization problem, called soft margin SVM.

ξ_i

$$\min_w \frac{\lambda}{2} w^T w + \frac{1}{n} \sum_{i=1}^{n} \max\left\{0, 1 - (2y_i - 1)\left(w^T x_i + b\right)\right\}$$

SVM Weights

Quiz

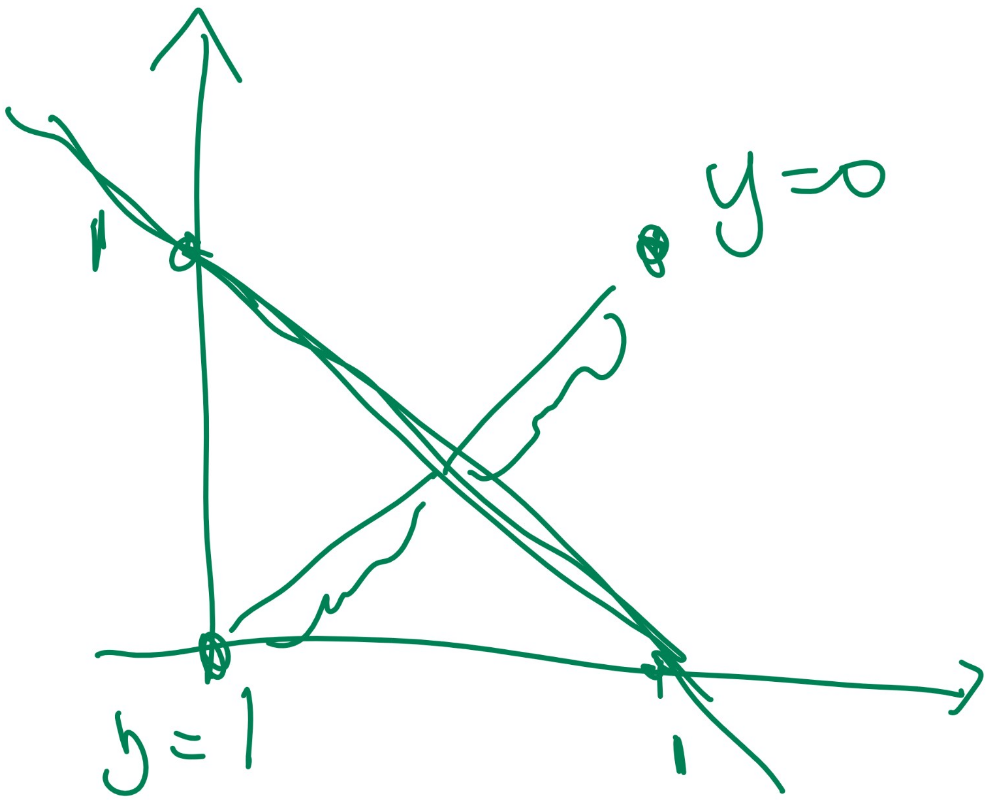
- Fall 2005 Final Q15 and Fall 2006 Final Q15
- Find the weights w_1, w_2 for the SVM classifier $\mathbb{1}_{\{w_1 x_{i1} + w_2 x_{i2} + 1 \geqslant 0\}}$ given the training data $x_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $x_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ with $y_1 = 1, y_2 = 0$
- A: $w_1 = 0, w_2 = -2$
- B: $w_1 = -2, w_2 = 0$
- C: $w_1 = -1, w_2 = -1$
- D: $w_1 = -2, w_2 = -2$
- E: none of the above

$\mathbb{1}\{\cancel{x_1} - 2x_2 + 1 \geq 0\}$

SVM

SVM Weights Diagram

Quiz



$w_1 + 1 = 0$

$w_2 + 1 = 0$

$1, 0$

$0, 1$

SVM Weights 2

Quiz

will be on midterm

Q2

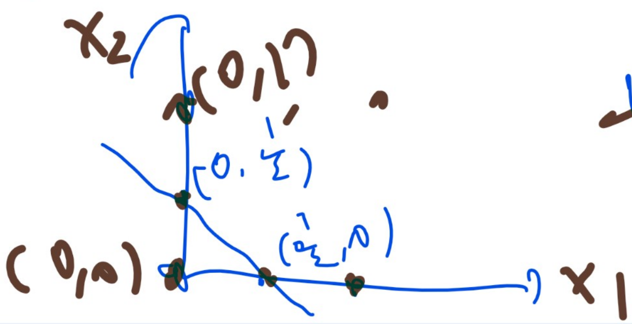
D

okay with A.

- Find the weights w_1, w_2 for the SVM classifier $\mathbb{1}_{\{w_1 x_{i1} + w_2 x_{i2} + 2 \geqslant 0\}}$ given the training data $y = \neg (x_1 \vee x_2)$, $x_1, x_2, y \in \{0, 1\}$.

x_1	x_2	y
0	0	1
0	1	0
1	0	0
1	1	0

- A: $w_1 = -3, w_2 = -3$
- B ~~A~~: $w_1 = -4, w_2 = -3$
- C ~~A~~: $w_1 = -3, w_2 = -4$
- D ~~A~~: $\boxed{w_1 = -4, w_2 = -4}$
- E ~~A~~: $w_1 = -8, w_2 = -8$

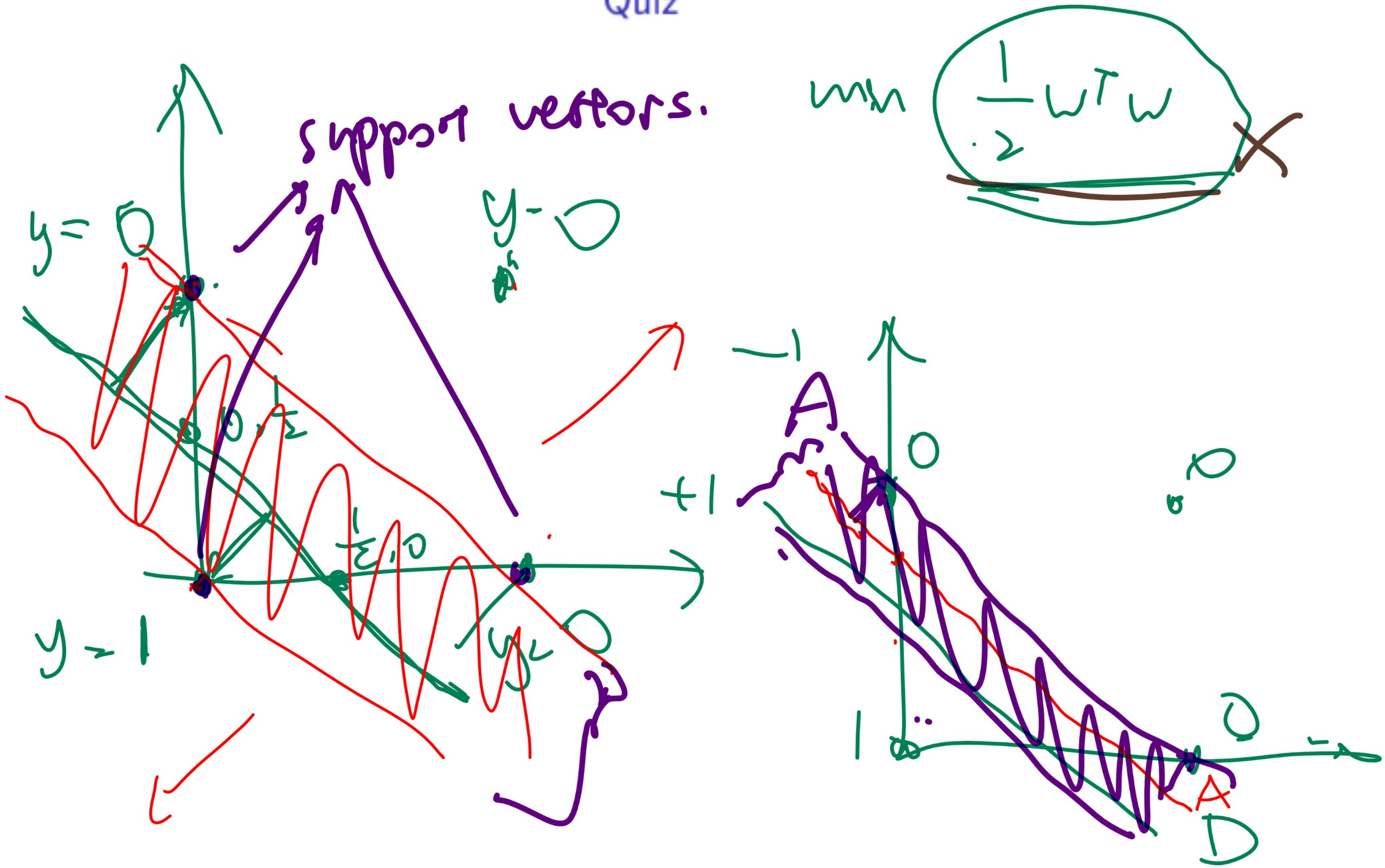


$\cancel{w_1 0} + \underbrace{w_2}_{=-4} \frac{1}{2} + 2 = 0$

$\underbrace{w_1}_{=-4} \frac{1}{2} + \cancel{w_2 0} + 2 = 0$

SVM Weights 2 Diagram

Quiz



Soft Margin

Quiz

- Fall 2011 Midterm Q8 and Fall 2009 Final Q1
- Let $w = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $b = 3$. For the point $x = \begin{bmatrix} 4 \\ 5 \end{bmatrix}$, $y = 0$, what is the smallest slack variable ξ for it to satisfy the margin constraint?

SVM

$$\boxed{(2y_i - 1)\left(w^T x_i + b\right) \geqslant 1 - \xi_i}, \xi_i \geqslant 0$$

0

min $\sum \xi_i$

$$-1\left((1,2)\begin{pmatrix}4\\5\end{pmatrix} + 3\right) \geqslant 1 - \xi_i$$

$$\xi_i \geqslant 1 + (14 + 3) = 18$$

Soft Margin 2

Quiz

$\xi_i \geq 1 - (-1)\left((1,2) \begin{pmatrix} -1 \\ -2 \end{pmatrix} + 3 \right) = 1 - 2 = -1$

Q3

-2

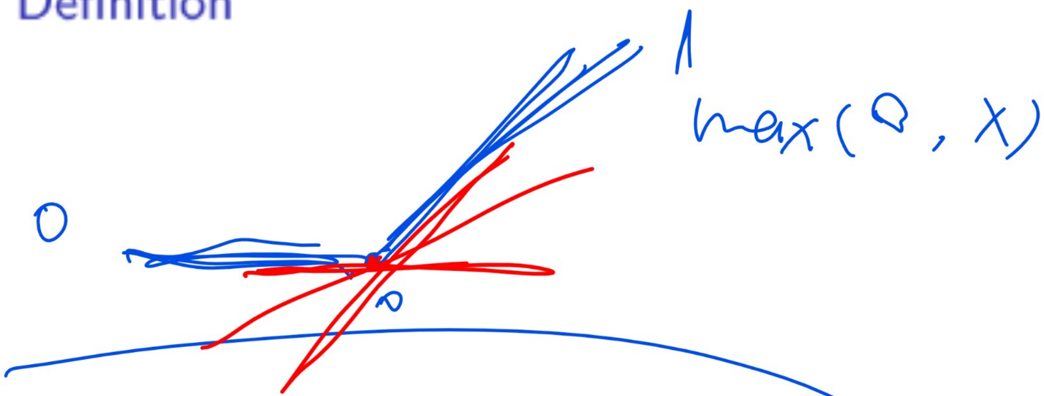
- Let $w = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ and $b = 3$. For the point $x = \begin{bmatrix} -1 \\ -2 \end{bmatrix}$, $y = 0$, what is the smallest slack variable ξ for it to satisfy the margin constraint?
- A: ~~-1~~
- B: 0
- C: 1
- D: 2
- E: 3

$$\begin{cases} \boxed{(2y_i - 1)(w^T x_i + b) \geq 1 - \xi_i} \\ \boxed{\xi_i \geq 0} \end{cases}$$

$$\begin{cases} \xi_i \geq -1 \\ \xi_i \geq 0 \end{cases}$$

Subgradient Descent

Definition



$$\min_{w} \frac{\lambda}{2} w^T w + \frac{1}{n} \sum_{i=1}^{n} \max\left\{0, 1 - (2y_i - 1)\left(w^T x_i + b\right)\right\}$$

- The gradient for the above expression is not defined at points with $1 - (2y_i - 1)\left(w^T x_i + b\right) = 0$.
- Subgradient can be used instead of gradient.

Subgradient

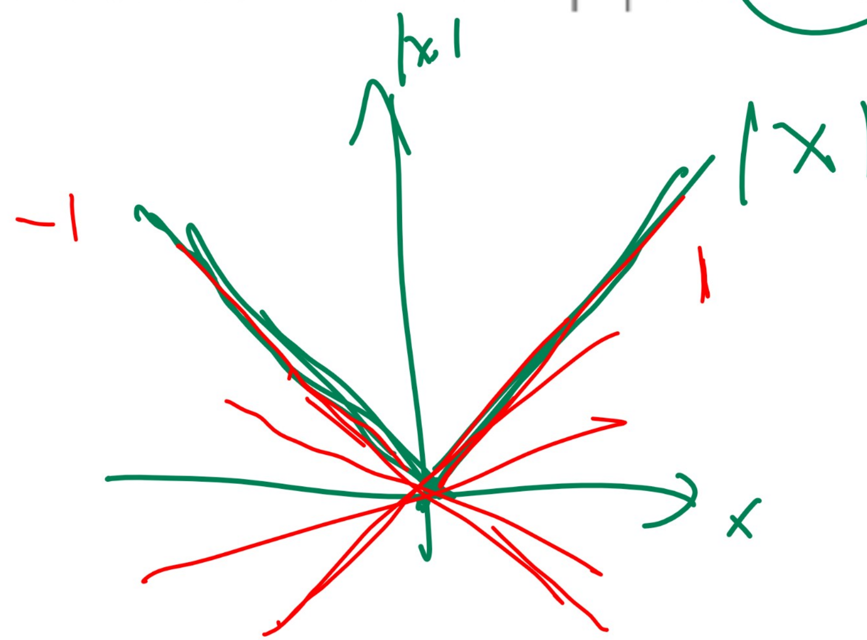
- The subderivative at a point of a convex function in one dimension is the set of slopes of the lines that are tangent to the function at that point.
- The subgradient is the version for higher dimensions.
- The subgradient $\partial f(x)$ is formally defined as the following set.

$$\partial f(x) = \left\{ v : f(x') \geqslant f(x) + v^T (x' - x) \ \forall \, x' \right\}$$

Subgradient 1

Quiz

- Which ones (multiple) are subderivatives of $|x|$ at $x = 0$?
- A: -1
- B: -0.5
- C: 0
- D: 0.5
- E: 1

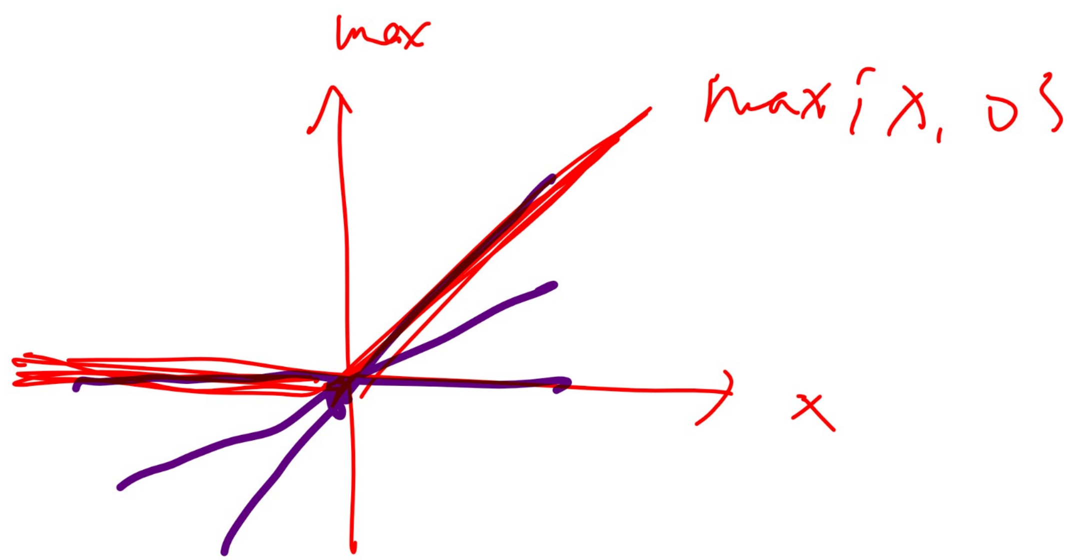


Subgradient 2

Quiz

Q 4

- Which ones (select one of them) are subderivatives of $\max\{x, 0\}$ at $x = 0$?
- A: -1
- B: -0.5
- C: 0
- D: 0.5
- E: 1



Subgradient Descent Step

Definition

- One possible set of subgradients with respect to w and b are the following.

max 0 =

$$\partial_w C \ni \lambda w - \underbrace{\sum_{i=1}^{n} (2y_i - 1)\, x_i \mathbb{1}_{\{(2y_i-1)(w^T x_i + b) \geqslant 1\}}}$$

$$\partial_b C \ni -\sum_{i=1}^{n} (2y_i - 1))\mathbb{1}_{\{(2y_i-1)(w^T x_i + b) \geqslant 1\}}$$

- The gradient descent step is the same as usual, using one of the subgradients in place of the gradient.

PEGASOS Algorithm

Algorithm

- Inputs: instances: $\{x_i\}_{i=1}^{n}$ and $\{z_i = 2y_i - 1\}_{i=1}^{n}$
- Outputs: weights: $\{w_j\}_{j=1}^{m}$
- Initialize the weights.

$$w_j \sim \text{ Unif } [0, 1]$$

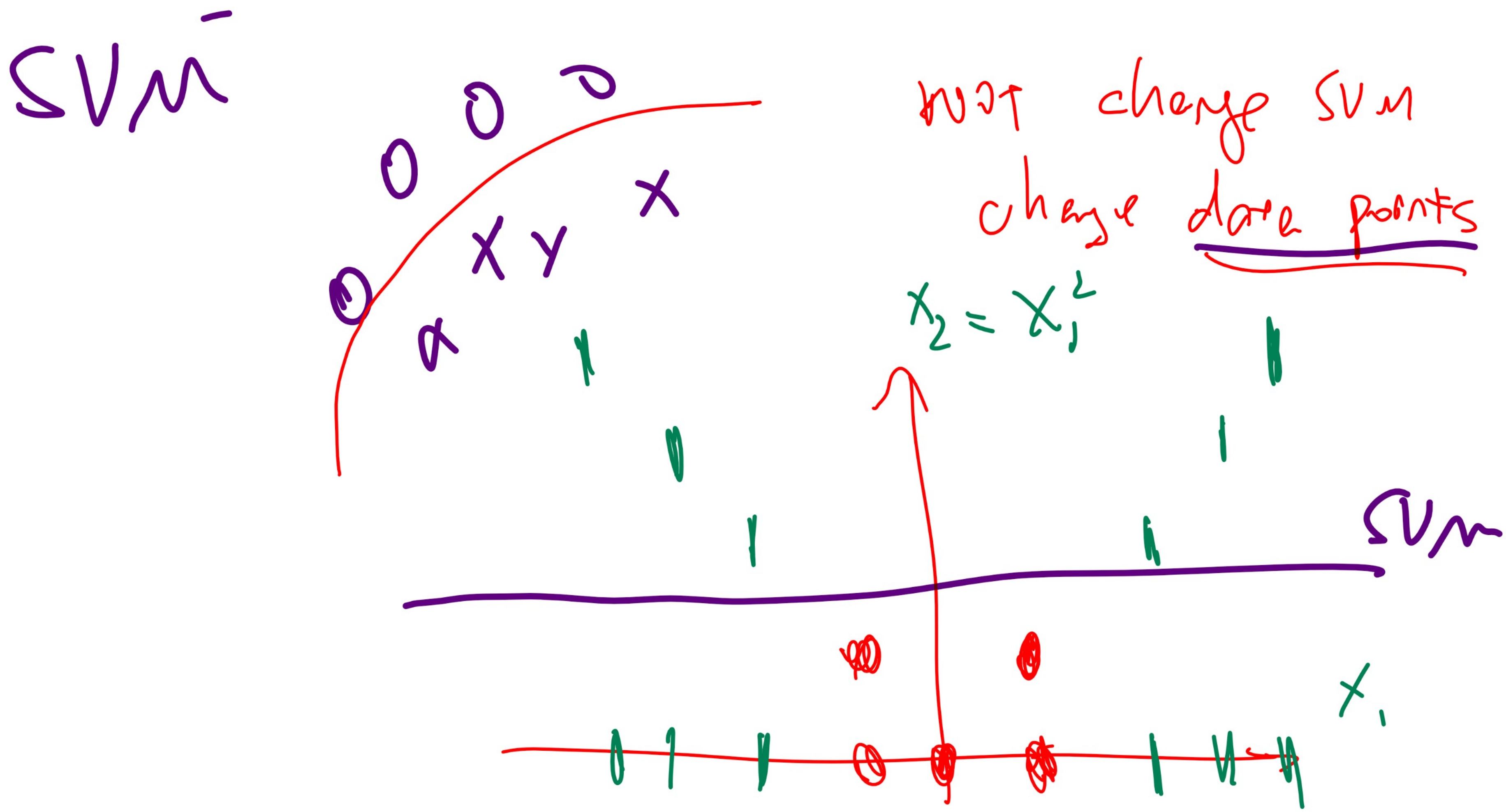
- Randomly permute (shuffle) the training set and performance subgradient descent for each instance i.

$$\boxed{w = (1 - \lambda)\, w - \alpha z_i \mathbb{1}_{\{z_i w^T x_i \geqslant 1\}} x_i}$$

- Repeat for a fixed number of iterations.

Kernel Trick 1D Diagram

Motivation



Kernelized SVM

Definition

- With a feature map φ, the SVM can be trained on new data points $\{(\varphi(x_1), y_1), (\varphi(x_2), y_2), \ldots, (\varphi(x_n), y_n)\}$.
- The weights w correspond to the new features $\varphi(x_i)$.
- Therefore, test instances are transformed to have the same new features.

$$\hat{y}_i = \mathbb{1}_{\{w^T \varphi(x_i) \geq 0\}}$$

Kernel Matrix

Definition

- The feature map is usually represented by a $n \times n$ matrix K called the Gram matrix (or kernel matrix).

$$K_{ii'} = \varphi(x_i)^T \varphi(x_{i'})$$

Examples of Kernel Matrix

Definition

- For example, if $\varphi(x) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right)$, then the kernel matrix can be simplified.

$$K_{ii'} = \left(x_i^T x_{i'}\right)^2$$

- Another example is the quadratic kernel $K_{ii'} = \left(x_i^T x_{i'} + 1\right)^2$. It can be factored to have the following feature representations.

2D → 6D

x_1 x_2

$$\varphi(x) = \left(x_1^2, x_2^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, 1\right)$$

Examples of Kernel Matrix Derivation

Definition

K an $n \times n$ matrix

data points

$$K_{ii'} = (x_i^T x_{i'} + 1)^2$$

$$= \left(\begin{pmatrix} x_{i1} \\ x_{i2} \end{pmatrix}^T \begin{pmatrix} x_{i'1} \\ x_{i'2} \end{pmatrix} + 1 \right)^2$$

$$= (x_{i1} x_{i'1} + x_{i2} x_{i'2} + 1)^2$$

$$= \left(x_{i1}^2 \boxed{x_{i'1}^2} + \sqrt{2} x_{i1} x_{i2} \boxed{\sqrt{2} x_{i'1} x_{i'2}} + x_{i2}^2 \boxed{x_{i'2}^2} + \sqrt{2} x_{i1} \boxed{\sqrt{2} x_{i'1}} + \sqrt{2} x_{i2} \boxed{\sqrt{2} x_{i'2}} + 1 \cdot \boxed{1} \right)$$

$$\phi(x_1, x_2) = (x_1^2, \sqrt{2} x_1 x_2, \sqrt{2} x_1, \sqrt{2} x_2, x_2^2, 1)$$

Popular Kernels

Discussion

- Other popular kernels include the following.

1. Linear kernel: $K_{ii'} = x_i^T x_{i'}$ SVM
2. Polynomial kernel: $K_{ii'} = \left(x_i^T x_{i'} + 1\right)^d$
3. Radial Basis Function (Gaussian) kernel:
$$K_{ii'} = \exp\left(-\frac{1}{\sigma^2}\left(x_i - x_{i'}\right)^T\left(x_i - x_{i'}\right)\right)$$

- Gaussian kernel has infinite dimensional feature representations. There are dual optimization techniques to find w and b for these kernels.

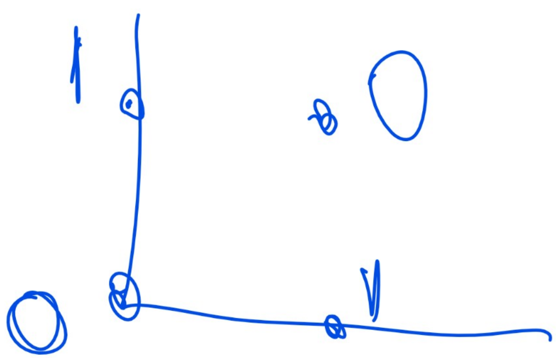
Kernel Trick for XOR

Quiz

- March 2018 Final Q17
- SVM with quadratic kernel $\varphi(x) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right)$ can correctly classify the training set for $y = x_1 \text{ XOR } x_2$.
- A: True.
- B: False.

$\phi(x)$			y
0	0	0	0
0	0	1	1
1	0	0	1
1	$\sqrt{2}$	1	0

x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	0

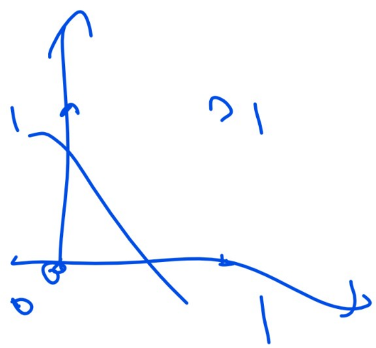


Kernel Trick for XOR 2

Quiz

Q5

- SVM with quadratic kernel $\varphi(x) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right)$ can correctly classify the training set for $y = x_1$ NAND x_2. NAND is just "not and".
- A: True.
- B: False.



x_1	x_2	y
0	0	1
0	1	1
1	0	1
1	1	0

Kernel Matrix

Quiz

- Fall 2009 Final Q2
- What is the feature vector $\varphi(x)$ induced by the kernel $K_{ii'} = \exp(x_i + x_{i'}) + \sqrt{x_i x_{i'}} + 3$?

map

- A: $(\exp(x), \sqrt{x}, 3)$
- B: $(\exp(x), \sqrt{x}, \sqrt{3})$
- C: $\left(\sqrt{\exp(x)}, \sqrt{x}, 3\right)$
- D: $\left(\sqrt{\exp(x)}, \sqrt{x}, \sqrt{3}\right)$
- E: None of the above

$$e^{x_i} \, e^{x_{i'}} + \sqrt{x_i} \cdot \sqrt{x_{i'}} + \sqrt{3} \cdot \sqrt{3}$$

$$(e^x, \sqrt{x}, \sqrt{3})$$

Kernel Matrix Math

Quiz

Kernel Matrix 2

Quiz

Back at 7:30

$e^{a+b} = e^a \cdot e^b$

Q6

- What is the feature vector $\varphi(x)$ induced by the kernel $K_{ii'} = 4\exp(x_i + x_{i'}) + 2x_i x_{i'}$?
- A: $(4\exp(x), 2\sqrt{x})$
- B: $(2\exp(x), \sqrt{2}\sqrt{x})$
- C: $(4\exp(x), 2x)$
- D: $(2\exp(x), \sqrt{2}x)$
- E: None of the above

$2e^{x_i}$ $2e^{x_{i'}}$

$+ \sqrt{2}x_i$ $\sqrt{2}x_{i'}$

$K_{ii'} = \phi(x_i)^T \phi(x_{i'})$

$\phi(x_i)$

Kernel Matrix Math 2

Quiz

Hat Game

Quiz (Participation)

Q7

- 5 kids are wearing either green or red hats in a party: they can see every other kid's hat but not their own.
- Dad said to everyone: at least one of you is wearing green hat.
- Dad asked everyone: do you know the color of your hat?
- Everyone said no.
- Dad asked again: do you know the color of your hat?
- Everyone said no.
- Dad asked again: do you know the color of your hat?
- Some kids (at least one) said yes.
- No one lied. How many kids are wearing green hats?
- A: 1... B: 2... C: 3... D: 4... E: 5

Hat Game Diagram

Discussion