CS540 Introduction to Artificial Intelligence

Lecture 5

Young Wu

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

June 6, 2022

Guess the Percentage

Admin

Q1

- Guess what percentage of the students (who are here) started $P1$?
- A : 0 to 20 percent.
- B : 20 to 40 percent.
- C : 40 to 60 percent.
- D : 60 to 80 percent.
- E : 80 to 100 percent.

Room : CS540E

The Percentage

Admin

Q2

- Did you start $P1$?
- A :
- B : Yes. 62%
- C :
- D : No.
- E :

Guess the Percentage

Admin

Q3

- Guess what percentage of the students (who are here) submitted $P1$?
- A : 0 to 20 percent.
- B : 20 to 40 percent.
- C : 40 to 60 percent.
- D : 60 to 80 percent.
- E : 80 to 100 percent.

Sharing Solutions

Admin

1. Use LaTeX (Word, Maple, MyScript etc).

   sqrt((a_1^2) / (2 pi)) is difficult to read compared to $\sqrt{\frac{a_1^2}{2\pi}}$.
2. Handwritten on tablet or on paper and photo or scan (Office Lens).
3. Other suggestions?

- Remember to make it a public Piazza Note (not a Question).
- I will either "good note" the post or leave a comment: if I leave a comment (please update your answers, reply to my comment, and remember to make the reply "unresolved" so I can see).

Shared Solution List and Feedback

Admin

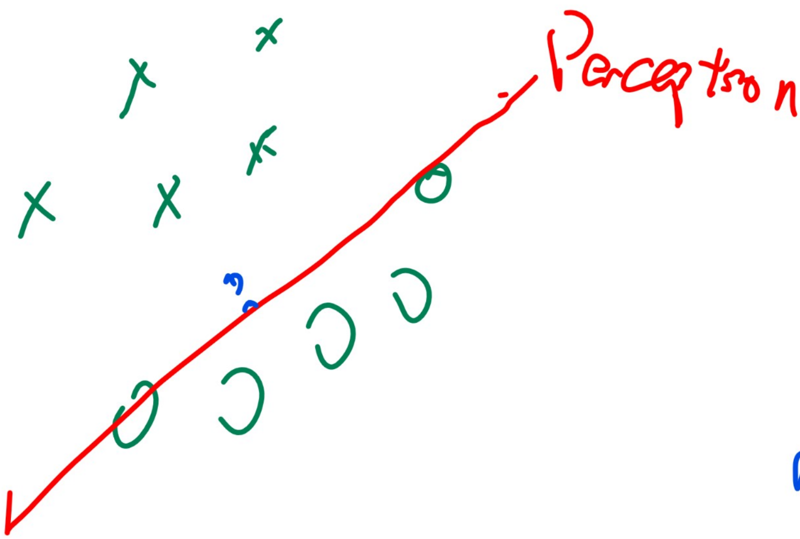
Q4

- The shared solutions are listed in the Main post on Piazza.
- Thank you for the feedback! I posted the responses to those in the Feedback post on Piazza.
- Question: can you see the labels in the $3D$ diagrams?
- A : Yes.
- B : Cannot see label.
- C : Cannot see anything.

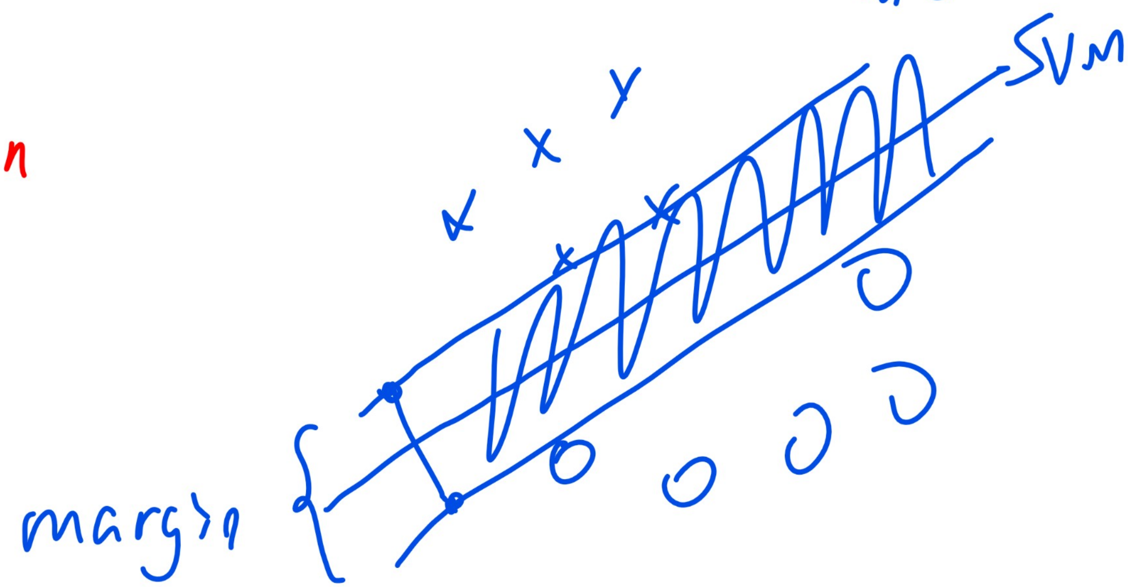
Maximum Margin Diagram

Motivation

LTU



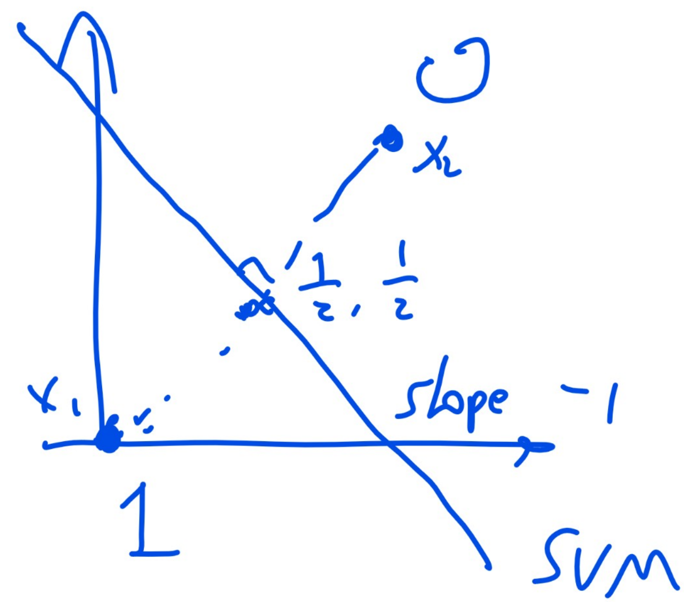
SVM



SVM Weights

Quiz

- Find the weights w_1, w_2 for the SVM classifier $\mathbb{1}_{\{w_1x_{i1}+w_2x_{i2}+\underbrace{1}_{b}\geqslant 0\}}$ given the training data $x_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $x_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ with $y_1 = 1, y_2 = 0$.
- $A : w_1 = 0, w_2 = -2$
- $B : w_1 = -2, w_2 = 0$
- $C : w_1 = -1, w_2 = -1$
- $D : w_1 = -2, w_2 = -2$



SVM Weights Diagram

Quiz

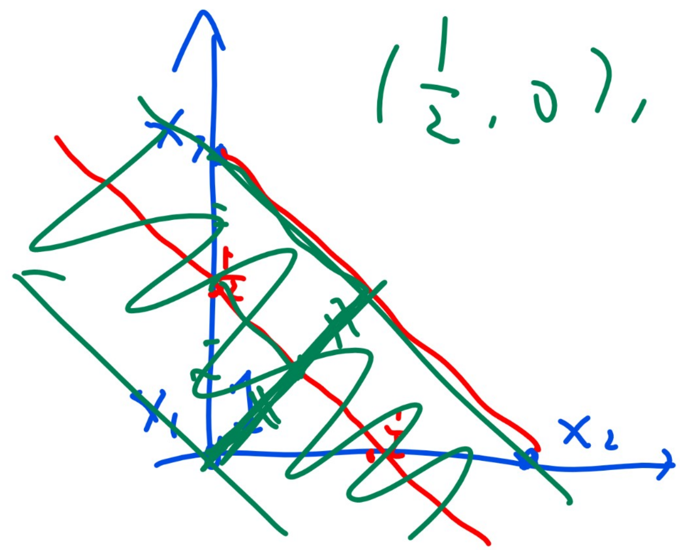
SVM Weights

Quiz

Q5

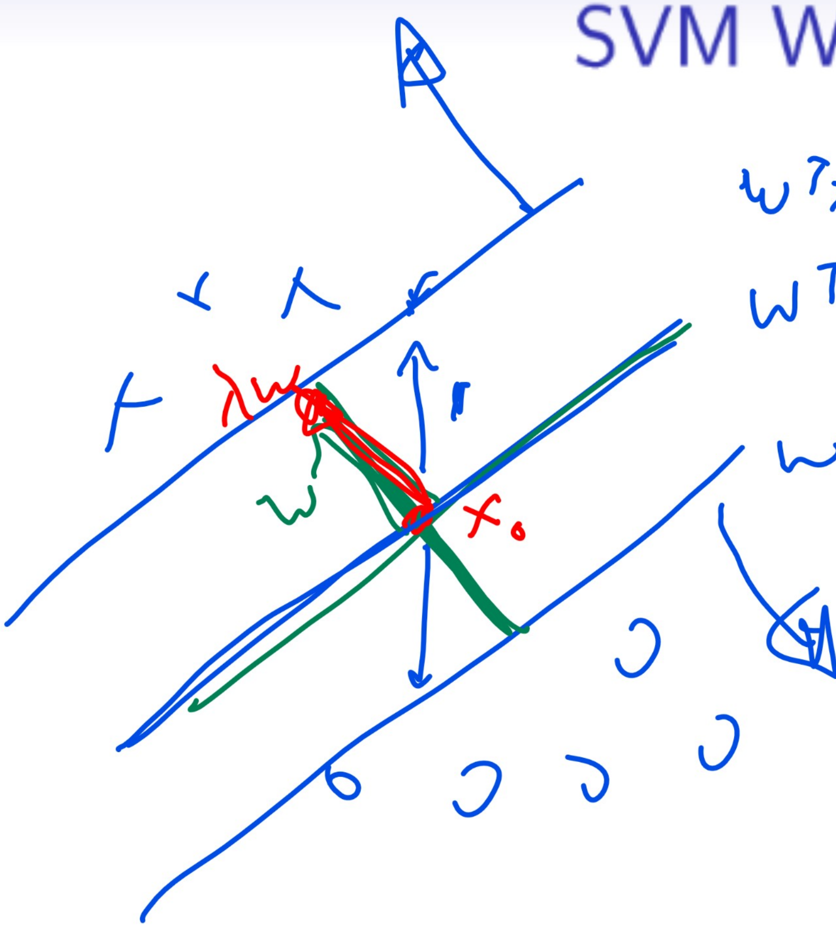
- Fall 2005 Final $Q15$ and Fall 2006 Final $Q15$
- Find the weights w_1, w_2 for the SVM classifier $\mathbb{1}_{\{w_1 x_{i1} + w_2 x_{i2} + 1 \geqslant 0\}}$ given the training data $x_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, x_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, x_3 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ with $y_1 = 1, y_2 = y_3 = 0$.
- $A : w_1 = -1.5, w_2 = -1.5$
- $B : w_1 = -2, w_2 = -1.5$
- $C : w_1 = -1.5, w_2 = -2$
- $D : w_1 = -2, w_2 = -2$
- $E : w_1 = -4, w_2 = -4$

$\left(\frac{1}{2}, 0\right), \left(0, \frac{1}{2}\right)$



SVM Weights Diagram

Quiz



$w^Tx+b+1=0$ ← $x_0+\lambda w$

$w^Tx+b=0$

$w^Tx+b-1>0$

$w^T(x_0-\lambda w)+b+1=0$

$w^T(x_0+\lambda w)+b-1=0$

$2\lambda w^Tw = 2$

$\lambda = \frac{1}{w^Tw}$

$= \frac{1}{\|w\|^2}$

thickness $= \|\lambda w\| = \lambda\|w\|$

$= \boxed{\frac{1}{\|w\|^2}}$

Constrained Optimization Diagram

Definition

half-margin

$$\arg\max \frac{1}{\|w\|^2} \quad \text{s.t.} \begin{cases} w^T x_i + b + 1 \geq 0 & \text{if } y_i = 0 \\ w^T x_i + b - 1 \leq 0 & \text{if } y_i = 1 \end{cases}$$

$$\Downarrow$$

$$\arg\min \frac{1}{2}\|\cancel{w}\|^2 \quad \text{s.t.} \quad \boxed{(2y_i - 1)(w^T x_i + b) \leq 1}$$

$w^T w$

$$\begin{cases} y_i = 1 \Rightarrow 1 \\ y_i = 0 \Rightarrow -1 \end{cases}$$

Hard Margin SVM.

allow mistake

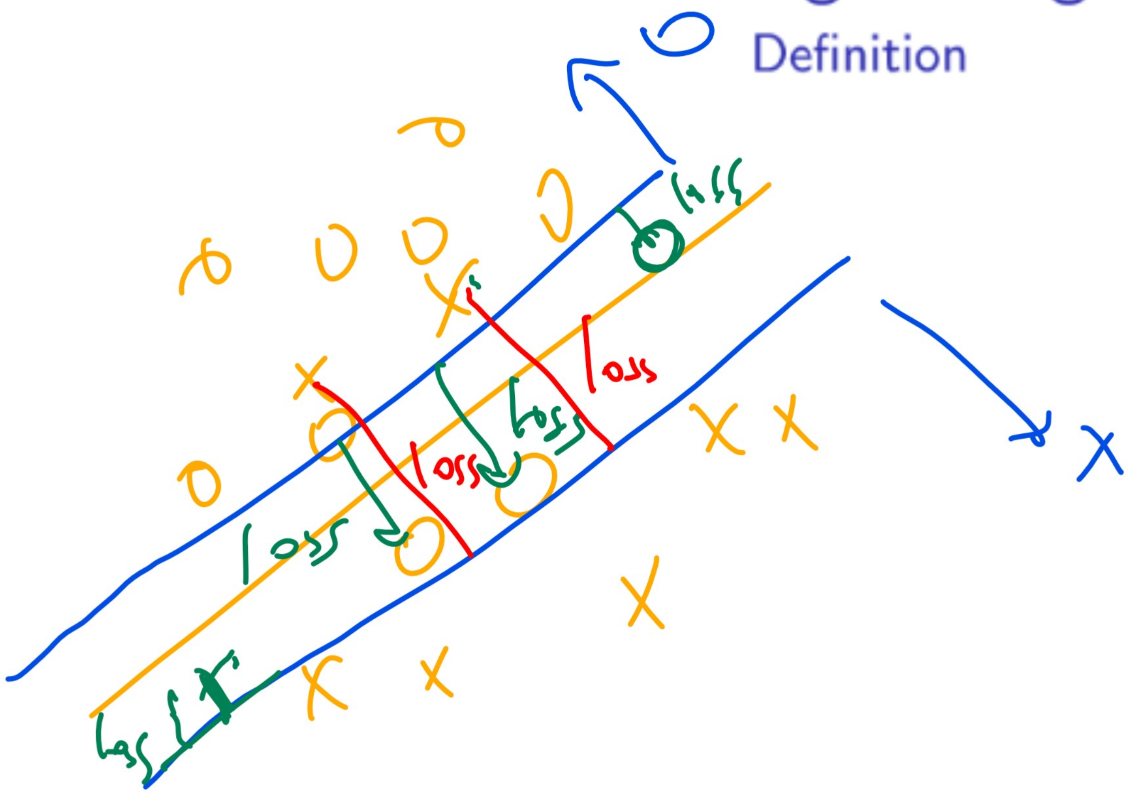
↳ there is a cost (loss)

Constrained Optimization Derivation

Definition

Soft Margin Diagram

Definition



add diagram in W3.

Soft Margin Derivation

Definition

penalty param

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

$$\text{s.t.} \quad (2y_i - 1)(w^T x_i + b) \leq 1 - \xi_i$$

$$\xi_i \geq 0$$

x_i

SVM Formulations

Definition

$y_i = 0, 1$

- Hard margin:

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

$-1 \quad 1$

- Soft margin:

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

ξ_i

Soft Margin

Quiz

- 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?

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

$$-1 \ (14 + 3) \geqslant 1 - \xi_i$$

$$\boxed{\begin{gathered} \xi_i \geqslant 0 \\ \xi_i \geqslant 18 \end{gathered}}$$

$$\boxed{\xi_i = 18}$$

Soft Margin 2

Quiz

Q6

- 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?
- $A : -12$
- $B : -10$
- $C : 0$
- $D : 10$
- $E : 12$

$$(2y_i - 1)(x_i^T w + b) \le 1 - \xi_i$$

$$\boxed{\xi_i \ge 0}$$

$$-1(-14+3) \le 1 - \xi_i$$

$$\boxed{\xi_i \ge -10}$$

Subgradient Descent

Definition

$$w = w - \frac{\partial C}{\partial w}$$

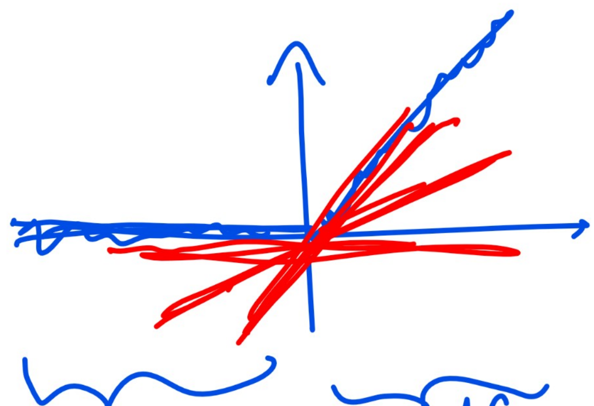
$$\min_{w} \frac{\lambda}{2} w^T w + \frac{1}{n} \sum_{i=1}^{n} \max\left\{0, 1 - (2y_i - 1)\left(w^T x_i + b\right)\right\} = C$$

- 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 a gradient.

Subgradient 1

Quiz

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



$$\partial \max\{x, 0\}\big|_0 = [0, 1]$$

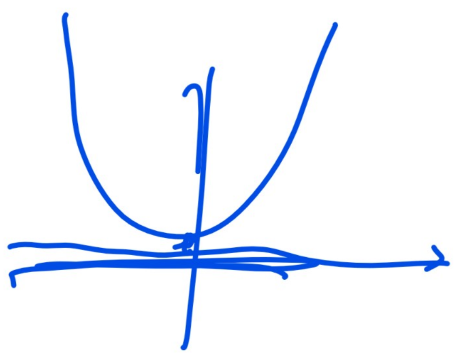
Subgradient 2

Quiz

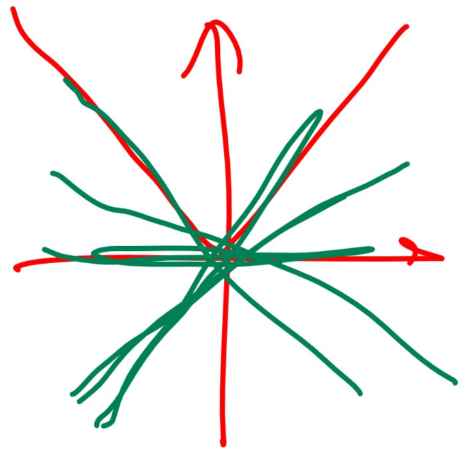
Q7

x^2

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



$|x|$



Subgradient Descent Step

Definition

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

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

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

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

Regularization Parameter

$w - \frac{\partial c}{\partial w} - \lambda w$

Definition

subgrad.

$$w = w - \alpha \sum_{i=1}^{n} z_i \mathbb{1}_{\{z_i w^T x_i \geq 1\}} x_i - \lambda w$$

$+\frac{1}{2}\lambda \|w\|^2$

$$z_i = 2y_i - 1, i = 1, 2, ..., n$$

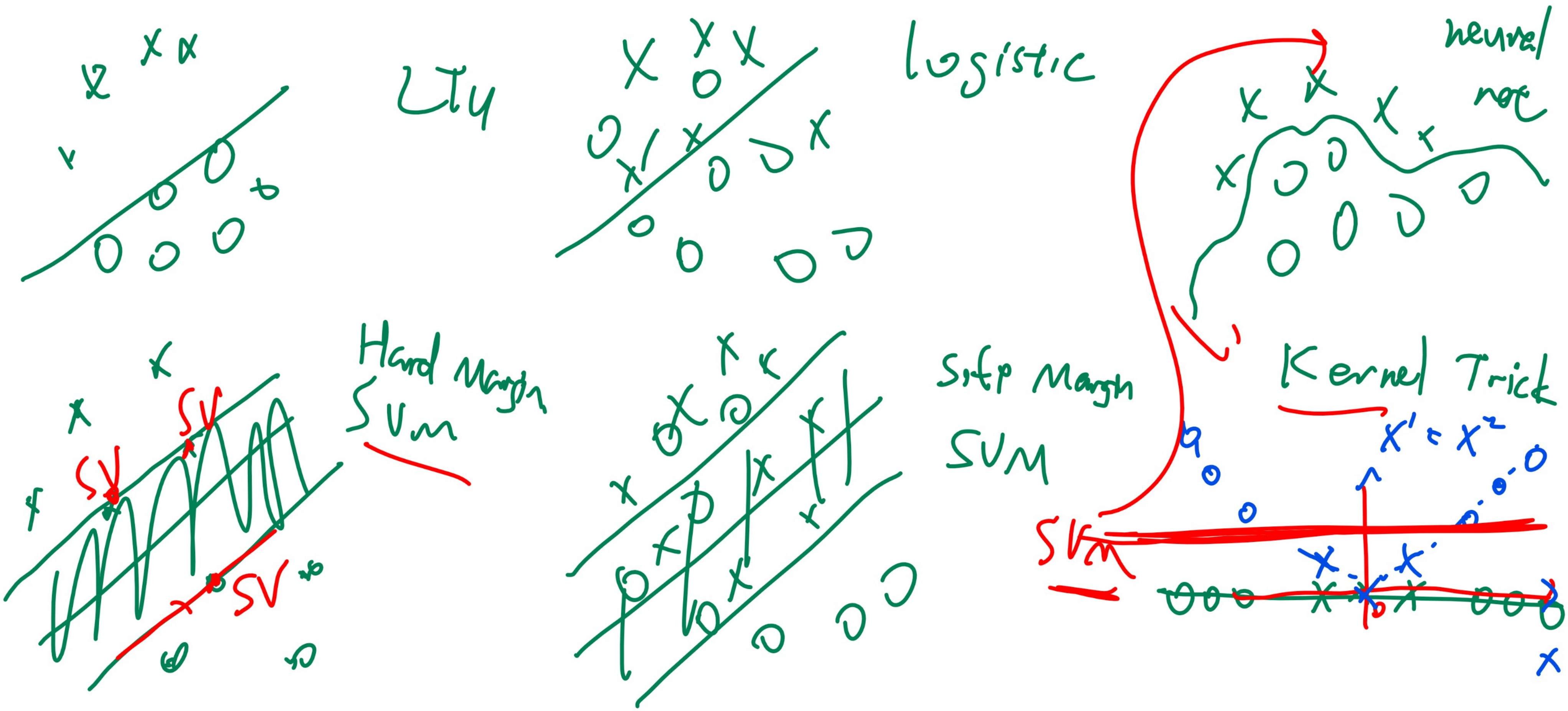
Svm

- λ is usually called the regularization parameter because it reduces the magnitude of w the same way as the parameter λ in $L2$ regularization.
- The stochastic subgradient descent algorithm for SVM is called PEGASOS: Primal Estimated sub-GrAdient SOlver for Svm.

Kernel Trick 1D Diagram

Motivation

back @ 6:50



Kernelized SVM

Definition

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

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

Kernel Trick for XOR

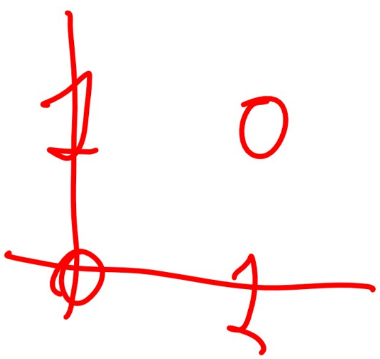
Quiz

x_1 , x_2 add $x_1^2 + x_2^2$

- SVM with quadratic kernel $\varphi(x) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right)$ can correctly classify the following training set?

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

x_1'	x_2'	x_3'	y
0	0	0	0
0	0	1	1
1	0	0	1
1	$\sqrt{2}$	1	0



Kernel Trick for XOR

Quiz

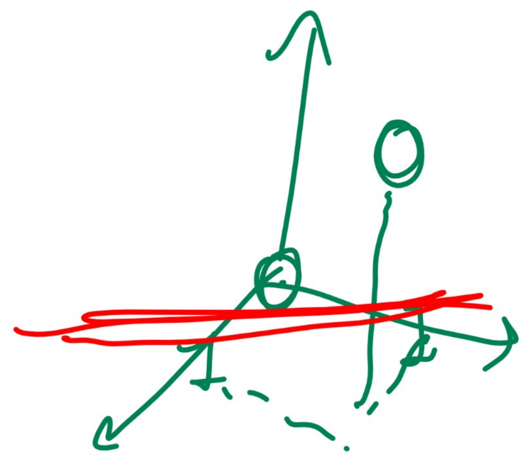
Q8

- SVM with kernel $\varphi(x) = (x_1, x_1 x_2, x_2)$ can correctly classify the following training set.

x_3

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

0
0
0
1



- A : True.
- B : False.

Kernel Matrix

Definition

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

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

$$K = \begin{bmatrix} \phi(x_1)^T \phi(x_1) \\ \phi(x_2)^T \phi(x_1) \\ \phi(x_3^T) \phi(x_1) \end{bmatrix}$$

~~# n x # m~~

n x # n

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.

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

symmetric positive semi-definite,

- 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.

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

Examples of Kernel Matrix Derivation

Definition

Popular Kernels

Discussion

- Other popular kernels include the following.

1. Linear kernel: $K_{ii'} = x_i^T x_{i'}$ → sum
2. Polynomial kernel: $K_{ii'} = \boxed{\left(x_i^T x_{i'} + 1\right)^d}$
3. Radial Basis Function (Gaussian) kernel:

$$\boxed{K_{ii'} = \exp\left(-\frac{1}{\sigma^2}\left(x_i - x_{i'}\right)^T\left(x_i - x_{i'}\right)\right)} \rightarrow \phi^T \phi$$

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

Kernel Matrix

Quiz

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

$= \phi(x_i)^T \phi(x_i')$

$= \underbrace{e^{x_i}}_{x_i} \underbrace{e^{x_i'}}_{x_i'} + \underbrace{\sqrt{x_i}}_{x_i} \underbrace{\sqrt{x_i'}}_{x_i'} + \underbrace{\sqrt{3}}_{x_i} \underbrace{\sqrt{3}}_{x_i'}$

$= [e^{x_i}, \sqrt{x_i}, \sqrt{3}] \begin{bmatrix} e^{x_i'} \\ \sqrt{x_i'} \\ \sqrt{3} \end{bmatrix}$

x

$\phi(x) = \begin{pmatrix} e^x \\ \sqrt{x} \\ \sqrt{3} \end{pmatrix}$

Kernel Matrix Math

Quiz

Kernel Matrix 2

Quiz

Q9

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

$\Rightarrow \underline{\phi(x_i)^T} \; \underline{\phi(x_i')}$

$\underbrace{2e^{x_i}}_{x_i} \; \underbrace{2e^{x_i'}}_{x_i'} + \underbrace{\sqrt{2}x_i}\;\underbrace{\sqrt{2}x_i'}$

$\underline{\left[2e^{x_i}, \sqrt{2}x_i\right]} \; \underline{\begin{bmatrix} 2e^{x_i'} \\ \sqrt{2}x_i' \end{bmatrix}}$

Kernel Matrix Math 2

Quiz