CS540 Introduction to Artificial Intelligence
Lecture 5

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

June 9, 2020

Survey Question

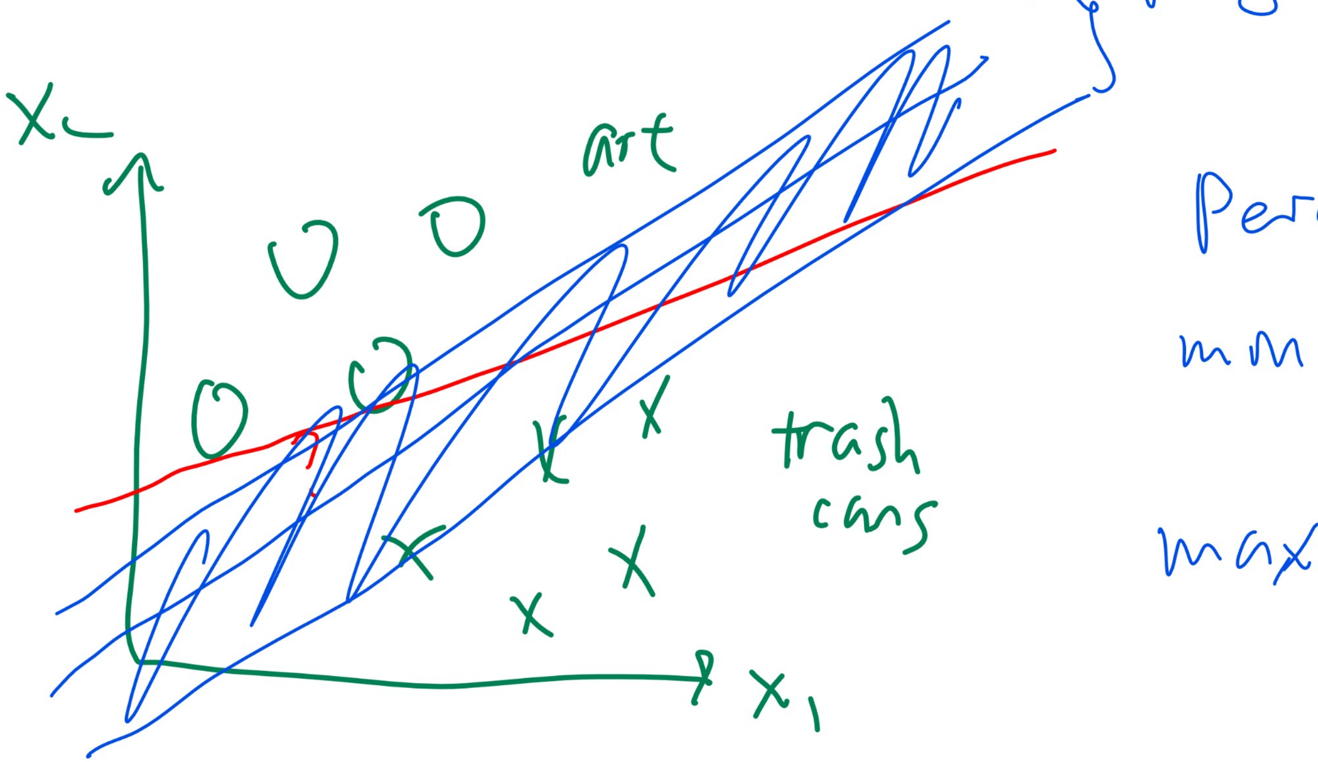
Admin

Socrative App Room: CS540C

- 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

Maximum Margin Diagram

Motivation



Perceptron

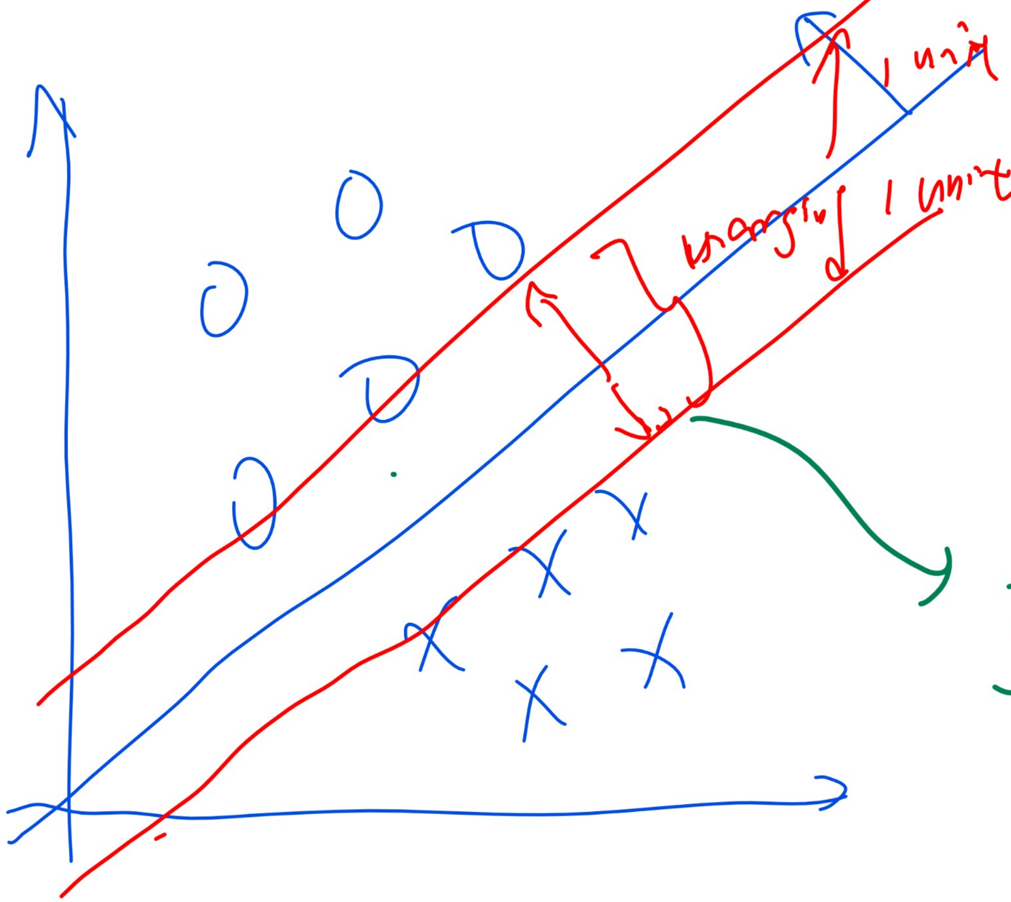
min mistake (loss)
(cost)

max margin

Constrained Optimization Derivation

Definition

$w^T x + b = 1$



decision boundary

$w^T x + b = 0$

$w^T x + b = -1$

$\frac{2}{\sqrt{w^T w}} = \text{margin},$

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 $y_i = 1$.

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

- 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

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

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

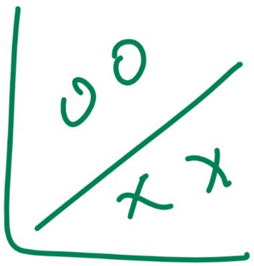
$$\min_{w} \frac{1}{2} w^T w \text{ such that } (2y_i - 1)\left(w^T x_i + b\right) \geqslant 1, i = 1, 2, \ldots, n$$

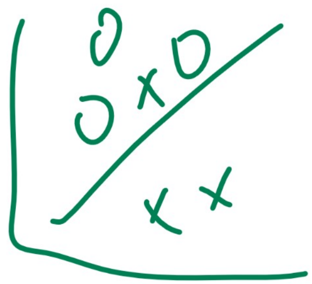
instance i
m features.

Soft Margin Diagram

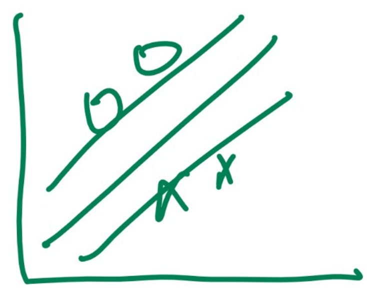
Definition

LTU → Logistic

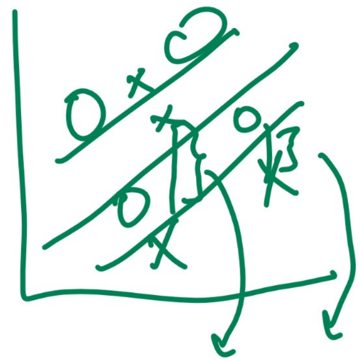




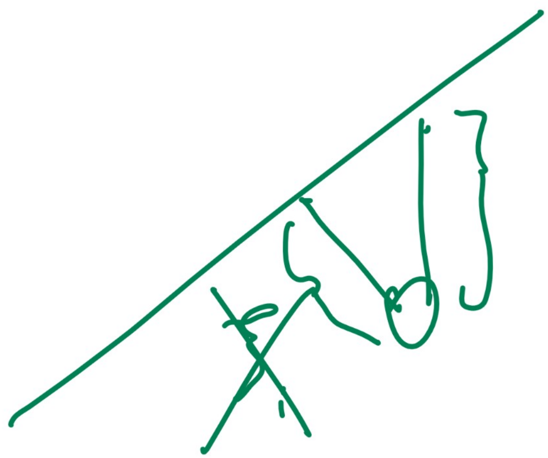
Hard-margin



Soft



x_i $\quad \xi_i > 0 \quad \xi_i = 0$



+1

kind of ξ_i

Soft Margin SVM

Definition

$\frac{1}{\text{margin}}$ average mistake relative importance

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

$$\text{such that } (2y_i - 1)\left(w^Tx_i + b\right) \geqslant 1 - \xi_i, \xi_i \geqslant 0, i = 1, 2, ..., n$$

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

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

SVM Weights

Quiz

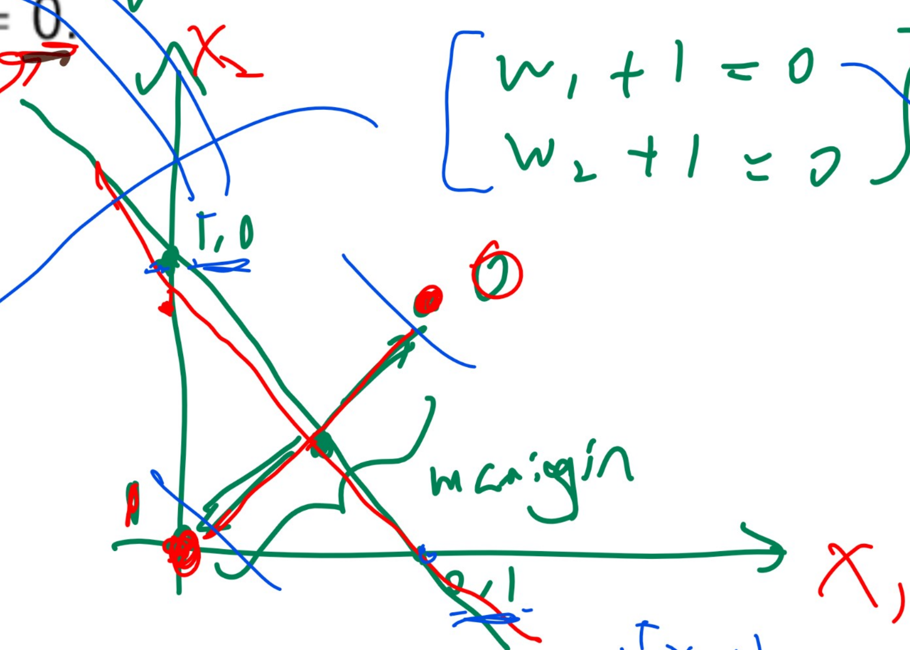
hard-margin.

- 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 $\boxed{x_1} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ and $x_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ with $\boxed{y_1} = 1, y_2 = 0$.

on midterm M4

$$\left[\begin{array}{l} w_1 + 1 = 0 \\ w_2 + 1 = 0 \end{array}\right]$$

- 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



$w^T x + b = 0$

SVM Weights Diagram

Quiz

SVM Weights 2

Quiz

Q2

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

training

min $w^T w$

- A: $w_1 = -3, w_2 = -3$
- B: $w_1 = -4, w_2 = -3$
- C: $w_1 = -3, w_2 = -4$
- D: $w_1 = -4, w_2 = -4$
- E: $w_1 = -8, w_2 = -8$

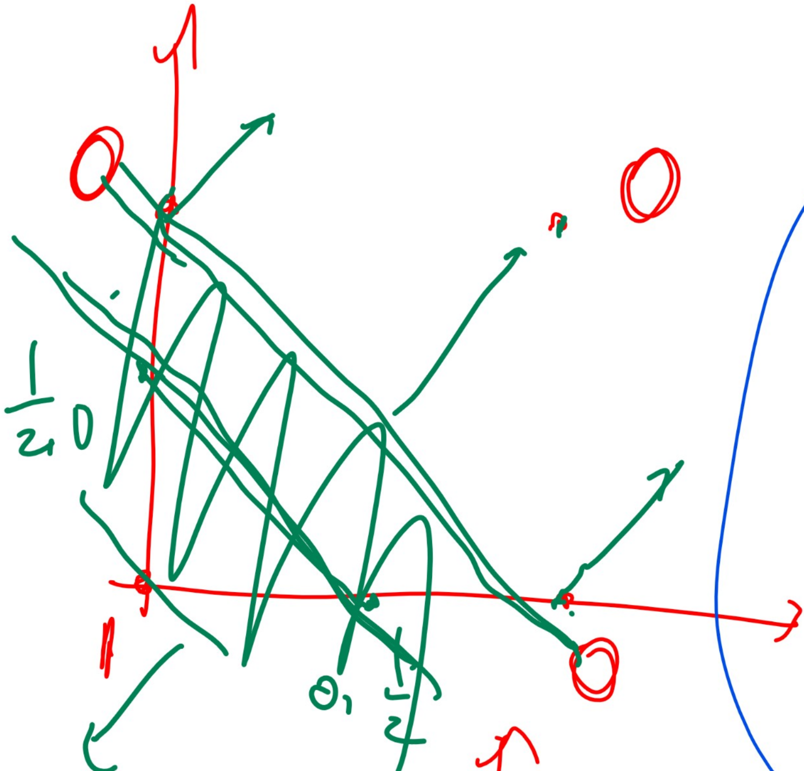
not constraints.

2

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

SVM Weights 2 Diagram

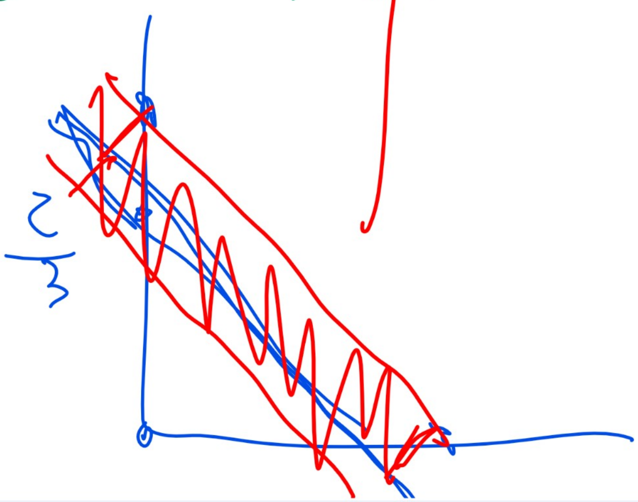
Quiz



$$\boxed{-3x_1 - 3x_2 + 2 = 0}$$

$$\begin{cases} w_1 \frac{1}{2} + w_2 0 + 2 = 0 \\ w_1 0 + w_2 \frac{1}{2} + 2 = 0 \end{cases}$$

$$\boxed{w_1 = -4, \; w_2 = -4}$$



$$x_2 = \frac{2}{3} - x_1$$

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?

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

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

$$-1\left((1,2)\begin{pmatrix} 4 \\ 5 \end{pmatrix} + 3\right) \geq 1 - \xi_i, \ \xi_i \geq 0$$

$$\boxed{\xi_i \geq 18, \ \xi_i \geq 0}$$

Soft Margin 2

Quiz

Q3

- 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

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

$\boxed{\xi_i \geq 0}$

$-1 \cdot -2 \geq 1 - \xi_i$

$\boxed{\xi_i \geq -1 \quad \xi_i \geq 0}$

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.

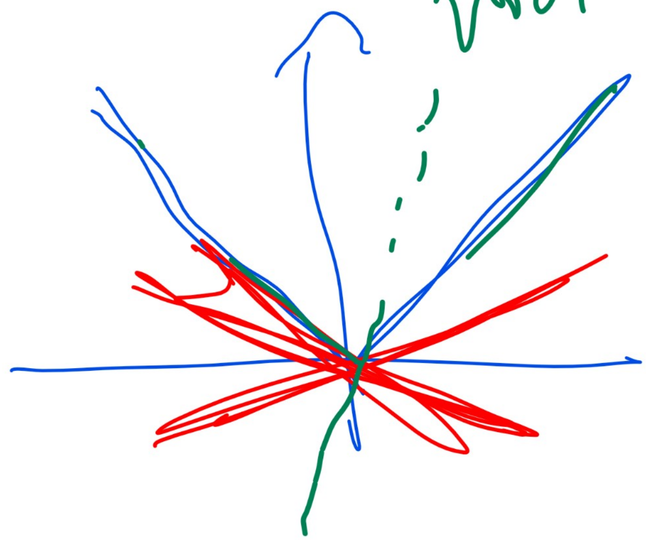
$$\partial f(x) = \left\{ v : f(x') \geqslant f(x) + v^T (x' - x)\ \forall x' \right\}$$

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

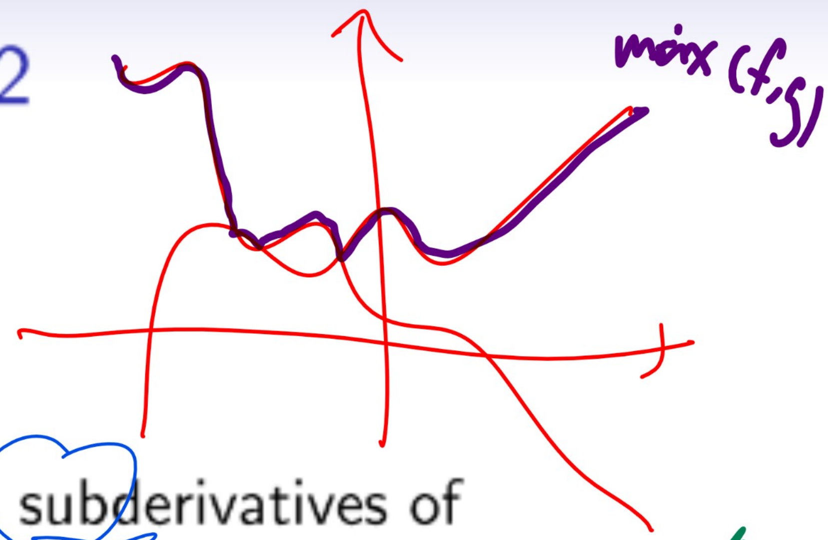
NOT tangent



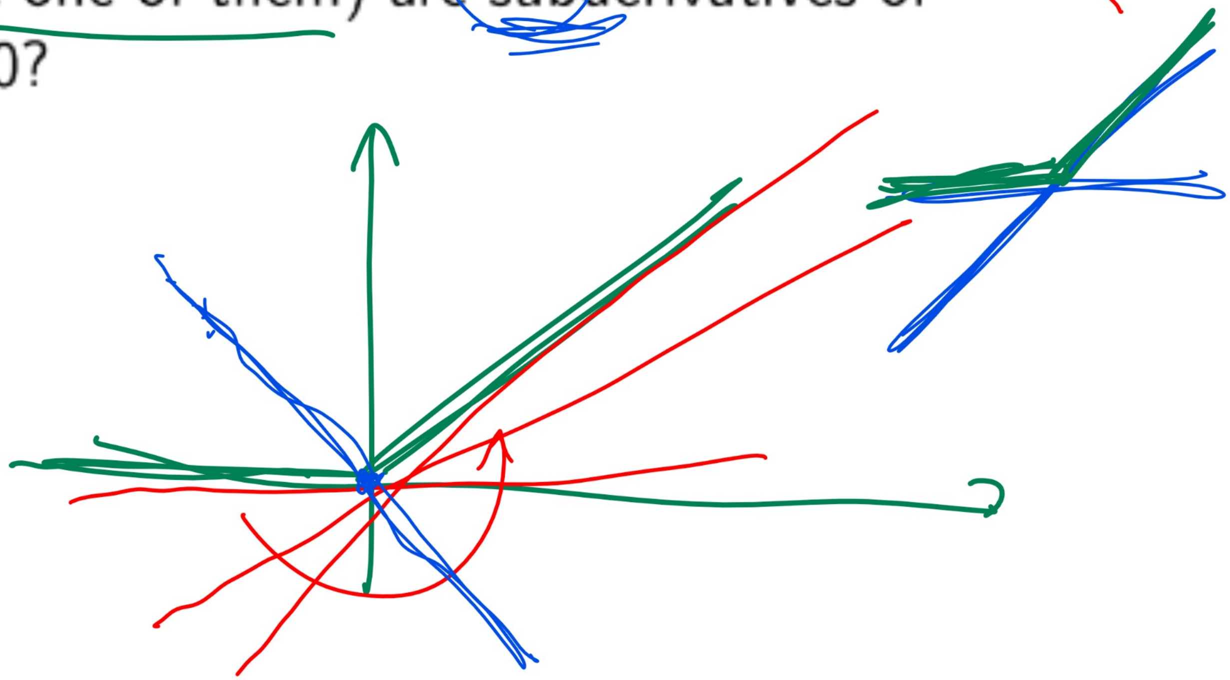
Subgradient 2

Quiz

Q4 (last)



- 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.

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

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

- 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.

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

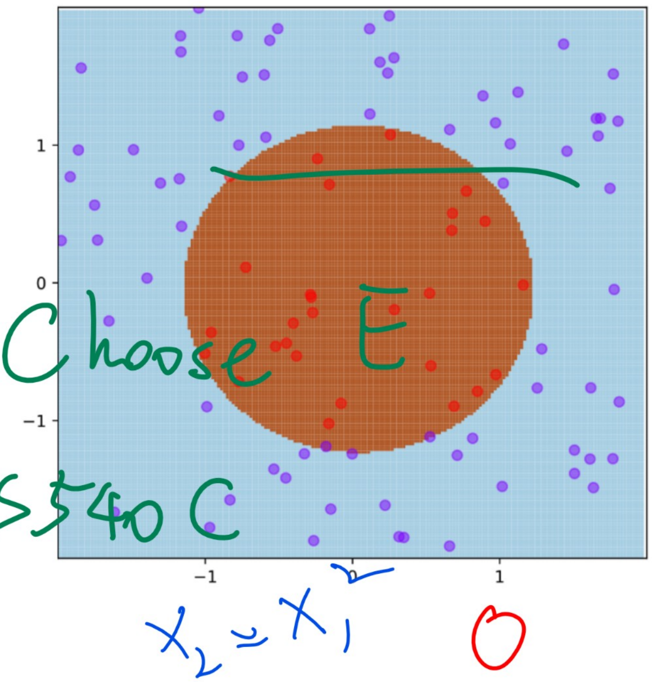
- Randomly permute (shuffle) the training set and performance subgradient descent for each instance i.

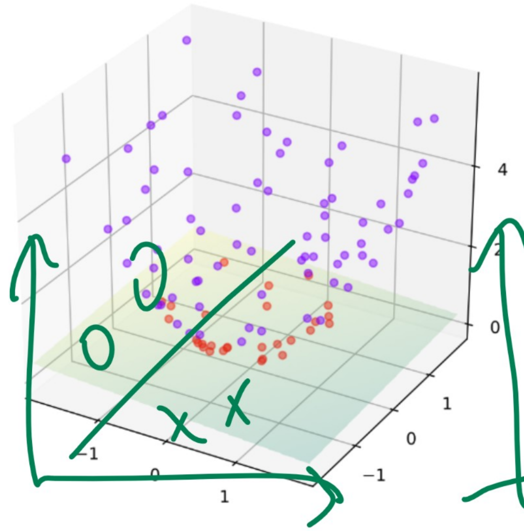
$$w = (1 - \lambda)\, w - \alpha z_i \mathbb{1}_{\{z_i w^T x_i \geq 1\}} x_i$$

- Repeat for a fixed number of iterations.

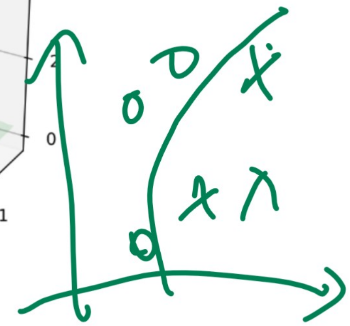
Q1 Choose E

Room: CS540 C

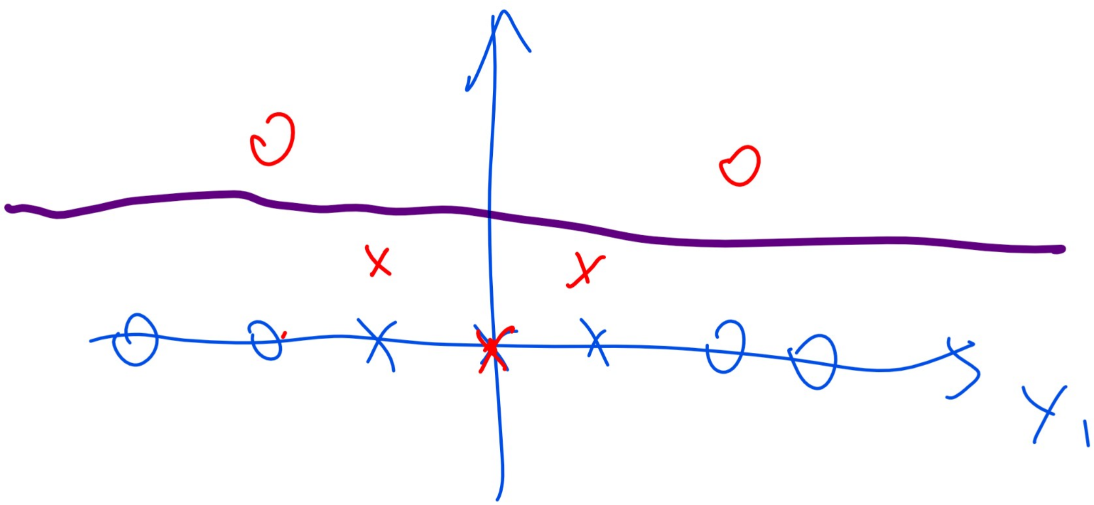




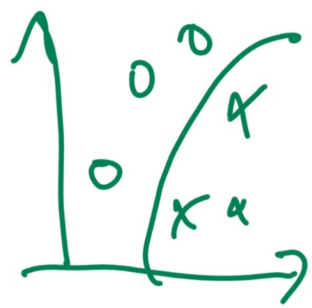
Perceptron



NN



SVM



Kernelized SVM

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.

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

Kernel Matrix

Definition

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

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

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.

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

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

✓

Examples of Kernel Matrix Derivation

x_{1}, x_{2} 2 different images

Definition

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

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

$$= \boxed{x_{i1}^2}\,\boxed{x_{i'1}^2} + \boxed{x_{i2}^2}\,\boxed{x_{i'2}^2} + \boxed{1}\cdot\boxed{1} + \boxed{\cancel{2}_{\sqrt{2}}\, x_{i1} x_{i2}}\,\boxed{\sqrt{2}\, x_{i'1} x_{i'2}}$$

$$+ \boxed{\cancel{2}_{\sqrt{2}}\, x_{i1}}\,\boxed{\sqrt{2}\, x_{i'1}} + \boxed{\cancel{2}_{\sqrt{2}}\, x_{i2}}\,\boxed{\sqrt{2}\, x_{i'2}}$$

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

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:

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

Kernel SVM

$n \times n$

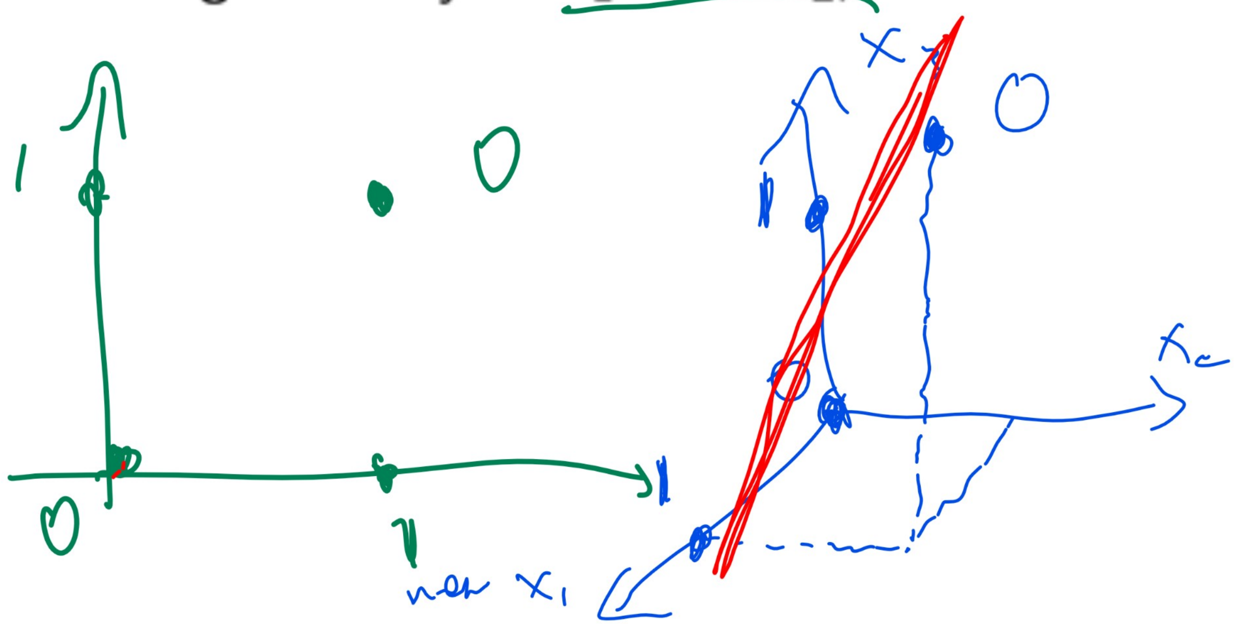
- Gaussian kernel has infinite dimensional feature representations. There are dual optimization techniques to find w and b for these kernels.

Kernel Trick for XOR

Quiz

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

- 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 $y = x_1$ XOR x_2.
- A: True.
- B: False.



Kernel Trick for XOR 2

Quiz

Q2

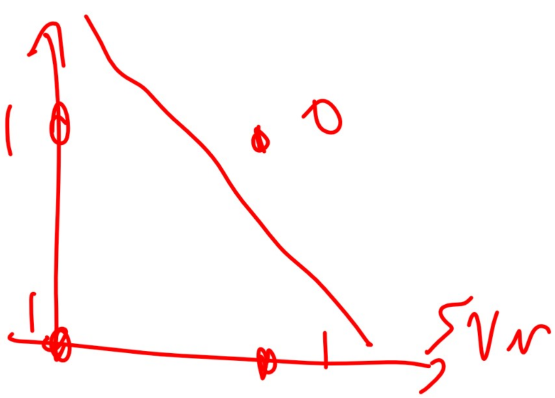
$\varphi'(x_1, x_2) = (x_1^3, x_1^2, x_1)$

feature mapping

- 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	NAND
0	0	1
0	1	1
1	0	1
1	1	0

$x_1 = x_1^2$	$x_2 = x_2^2$	
0	0	0
0	1	0
1	0	0
1	1	$\sqrt{2}$



Kernel Matrix

Quiz

- Fall 2009 Final Q2
- What is the feature vector $\varphi(x)$ induced by the kernel $K_{ii'} = \exp(x_i + x_{i'}) + \sqrt{x_i x_{i'}} + 3$? in 1D
- A: $\left(\exp(x), \sqrt{x}, 3\right)$
- B: $\left(\exp(x), \sqrt{x}, \sqrt{3}\right)$
- C: $\left(\sqrt{\exp(x)}, \sqrt{x}, 3\right)$
- D: $\left(\sqrt{\exp(x)}, \sqrt{x}, \sqrt{3}\right)$
- E: None of the above

$x_i = x_{i1}$ one feature

$\boxed{\exp(x_i)} \cdot \boxed{\exp(x_{i'})} + \boxed{\sqrt{x_i}} \cdot \boxed{\sqrt{x_{i'}}} + \boxed{\sqrt{3}} \cdot \boxed{\sqrt{3}}$

Kernel Matrix Math

Quiz

Kernel Matrix 2

Quiz

~~Q3~~ Q4

- 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

$\boxed{2\exp(x_i)} \cdot \boxed{2\exp(x_{i'})}$

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

Kernel Matrix Math 2

Quiz