CS540 Introduction to Artificial Intelligence
Lecture 5

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

June 6, 2019

Correction for Lecture 3 Slides

Review

- The gradient descent step formula in Lecture 3 Slides should have $a_i - y_i$ instead of $y_i - a_i$.

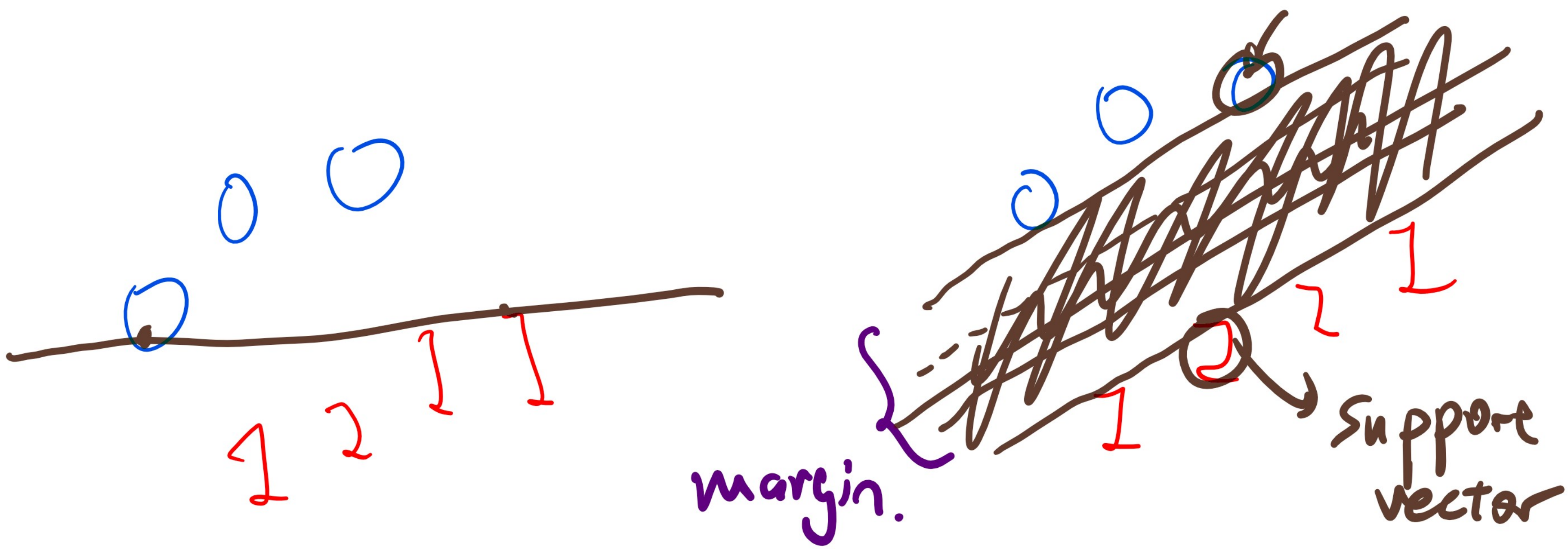
$$C = \frac{1}{2}\sum_{i=1}^{n}(a_i - y_i)^2 = \frac{1}{2}\sum_{i=1}^{n}(y_i - a_i)^2$$

$$\frac{\partial C}{\partial a_i} = (y_i - a_i)\cdot(-1) = a_i - y_i$$

- The slides are updated.

Maximum Margin Diagram

Motivation



Margin and Support Vectors

Motivation

- The perceptron algorithm finds any line (w, b) that separates the two classes.

$$\hat{y}_i = \mathbb{1}_{\{w^T x_i + b \geqslant 0\}}$$

- The margin is the maximum width (thickness) of the line before hitting any data point.
- The instances that the thick line hits are called support vectors.
- The model that finds the line that separates the two classes with the widest margin is call support vector machine (SVM).

Support Vector Machine

Description

- The problem is equivalent to minimizing the norm of the weights subject to the constraint that every instance is classified correctly (with the margin).
- Use subgradient descent to find the weights and the bias.

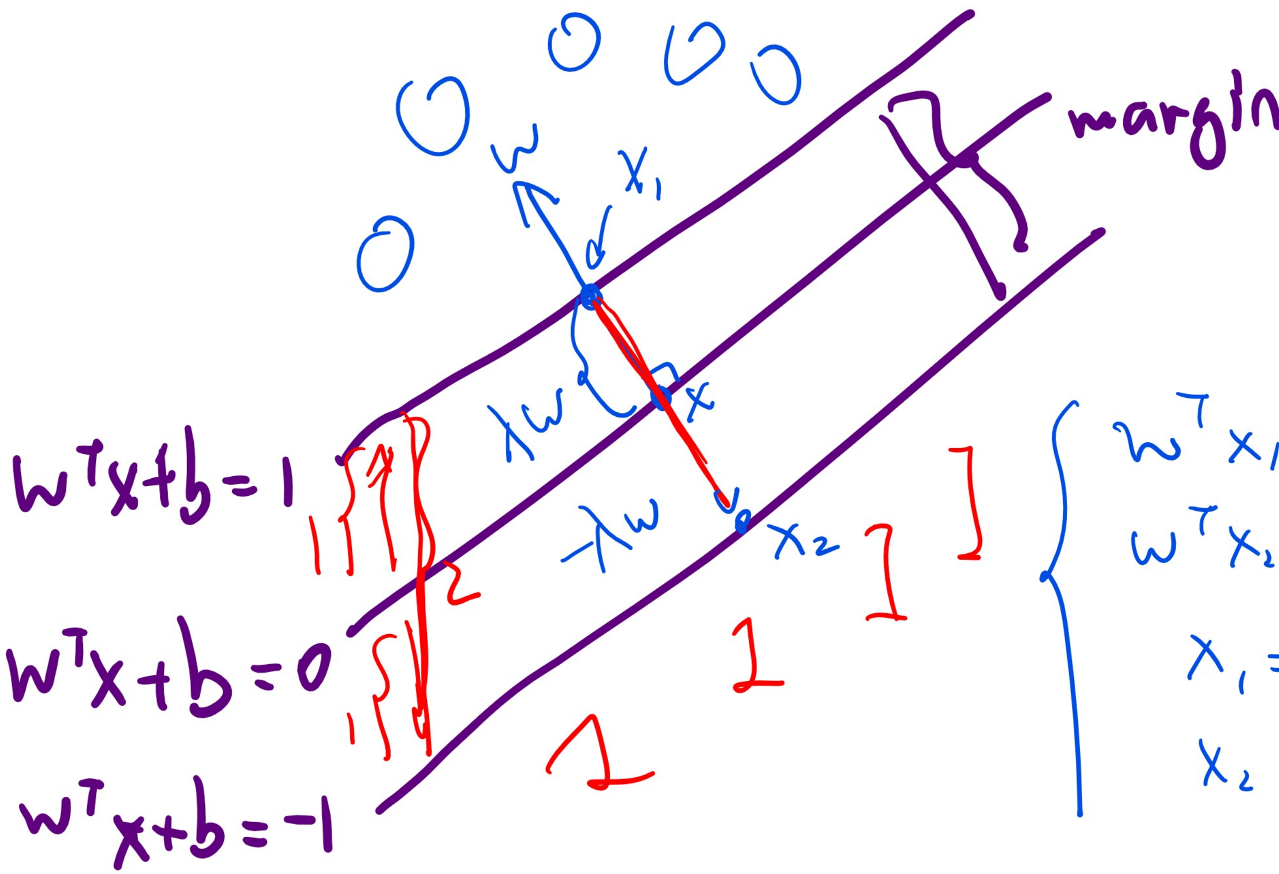
Finding the Margin

Definition

- Define two planes: plus plane $w^T x + b = 1$ and minus plane $w^T x + b = -1$.
- The distance between the two planes is $\dfrac{2}{\sqrt{w^T w}}$.
- If all of the instances with $y_i = 1$ are above the plus plane and all of the instances with $y_i = 0$ are below the minus plane, then the margin is $\dfrac{2}{\sqrt{w^T w}}$.

Constrained Optimization Derivation

Definition



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

$$\left\{ \begin{aligned} w^T x_1 + b &= 1 \\ w^T x_2 + b &= -1 \\ x_1 &= x + \lambda w \\ x_2 &= x - \lambda w \end{aligned} \right\} \Uparrow$$

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

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 $y_i = 1$.

margin

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

- The two constrains can be combined.

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

$y = 0 \Rightarrow -1$

$y = 1 \Rightarrow 1$

Hard Margin SVM

Definition

$$\max f \iff \min \frac{1}{f} \overset{f \geqslant 0}{\iff} \min \frac{1}{f^2} \iff \min \frac{2}{f^2}$$

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

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

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

SVM Weights

Quiz (Graded)

2

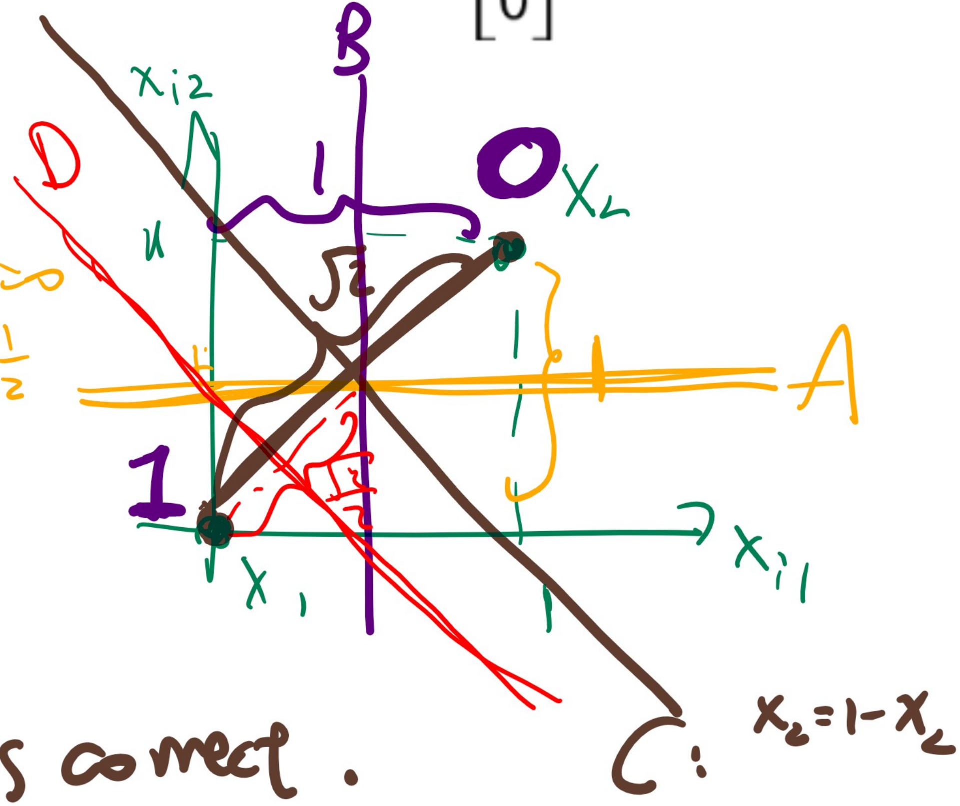
- Fall 2005 Final Q15 and Fall 2006 Final Q15
- Find the weights w_1, w_2 for the SVM classifier

SVM $\boxed{\mathbb{1}_{\{w_1 x_{i1} + w_2 x_{i2} + 1 \geqslant 0\}}}$ 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

$-2x_2 + 1 \geq 0$

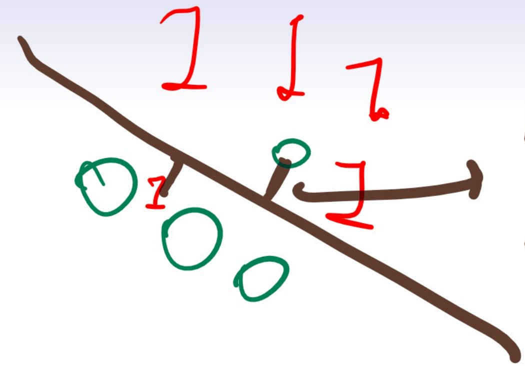
$x_2 = \frac{1}{2}$



check $\mathbb{1}\{-x_1 - x_2 + 1 \geqslant 0\}$ $\hat{y}$ is correct.

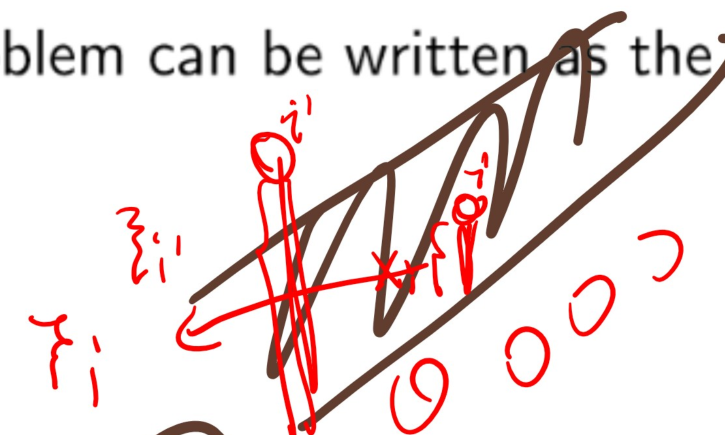
Soft Margin

Definition



mistakes + cost

- To allow for mistakes classifying a few instances, slack variables are introduced.
- The cost of violating the margin is given by some constant $\frac{1}{\lambda}$.
- Using slack variables ξ_i, the problem can be written as the following.



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

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

Soft Margin SVM

Definition

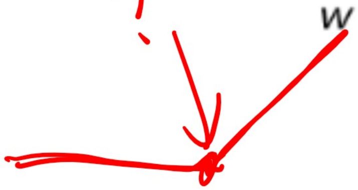
$$\lambda \cdot \left(\min_{w} \frac{1}{2} w^T w + \frac{1}{\lambda} \frac{1}{n} \sum_{i=1}^{n} \right) \xi_i$$

x_i

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

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

$$\min_{w} \frac{\lambda}{2} w^T w + \frac{1}{n} \sum_{i=1}^{n} \max \Big\{ 0, \overbrace{1 - (2y_i - 1)\left(w^T x_i + b\right)}^{\xi_1} \Big\}$$



when $\xi_1 < 0 \Rightarrow 0$

$\xi_1 > 0 \Rightarrow \xi_1$

Subgradient Descent

Definition

L2 regularization

min hinge loss cost function

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

- 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

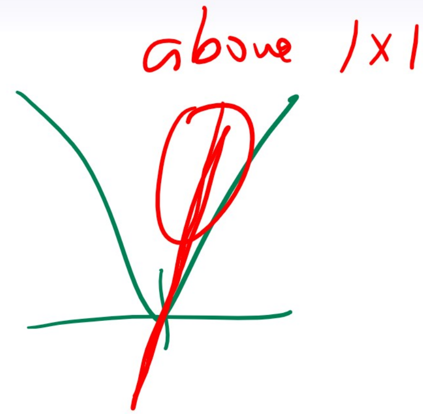
- The subderivative at a point of a convex function in one dimension is the set of slopes of the lines that are tangent to the function at that point.
- The subgradient is the version for higher dimensions.
- The subgradient $\partial f(x)$ is formally defined as the following set.

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

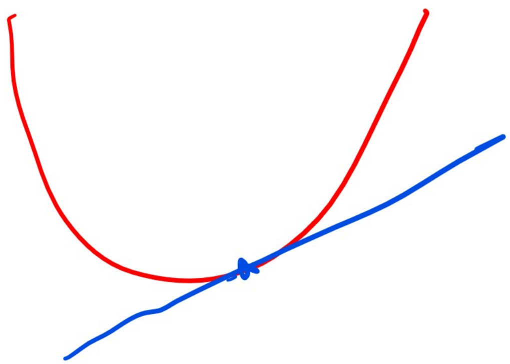
Subgradient, Part I

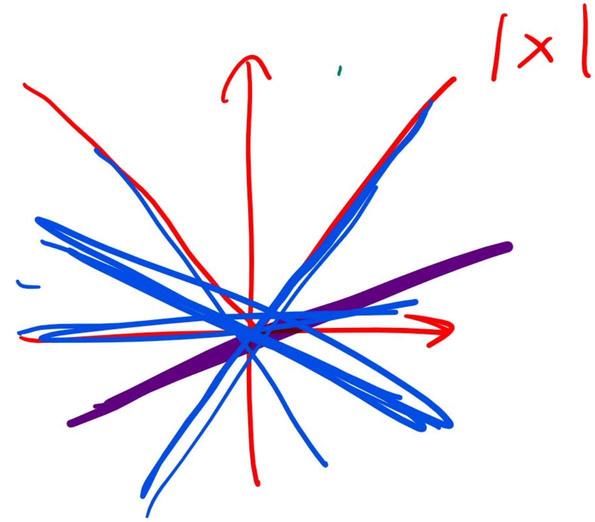
Quiz (Participation)

$$\partial_x |x| = [-1, 1]$$



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



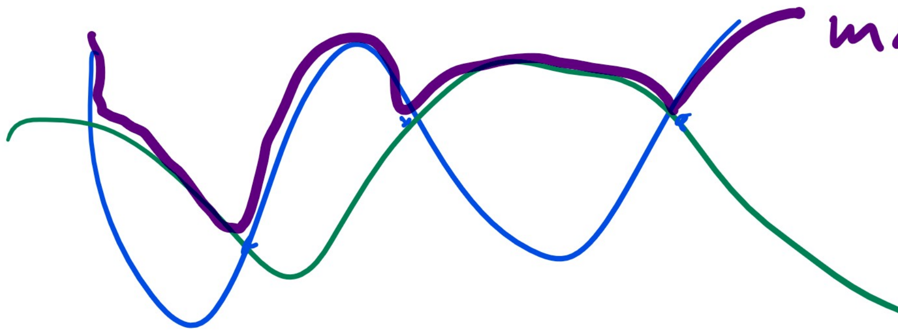


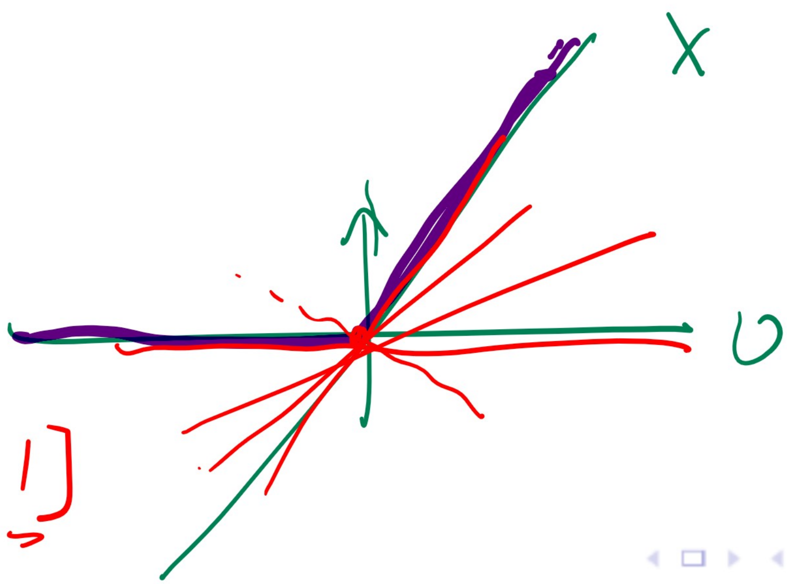
Subgradient, Part II

Quiz (Graded)

4

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





$\partial_x \max\{x, 0\} = [0, 1]$

Subgradient Descent Step

Definition

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

pick a fix subgradient.

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

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

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

Class Notation and Bias Term

Definition

- Usually, for SVM, the bias term is not included and updated. Also, the classes are -1 and +1 instead of 0 and 1. Let the labels be $z_i \in \{-1, +1\}$ instead of $y_i \in \{0, 1\}$. The gradient steps are usually written the following way.

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

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

$+ b \cdot \underbrace{1}_{x_j}$

have $b \iff$ add feature constant 1,

Regularization Parameter

Definition

λ looks like regularization term

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

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

- The parameter λ is slightly different from the one from the previous slides. λ is usually called the regularization parameter because it reduces the magnitude of w the same way as the parameter λ in L2 regularization.

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.

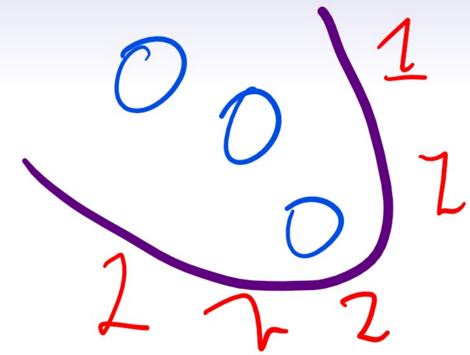
$$w_j \sim \text{ Unif } [0, 1]$$

- Update the weights using subgradient descent for a fixed number of iterations.

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

Kernel Trick

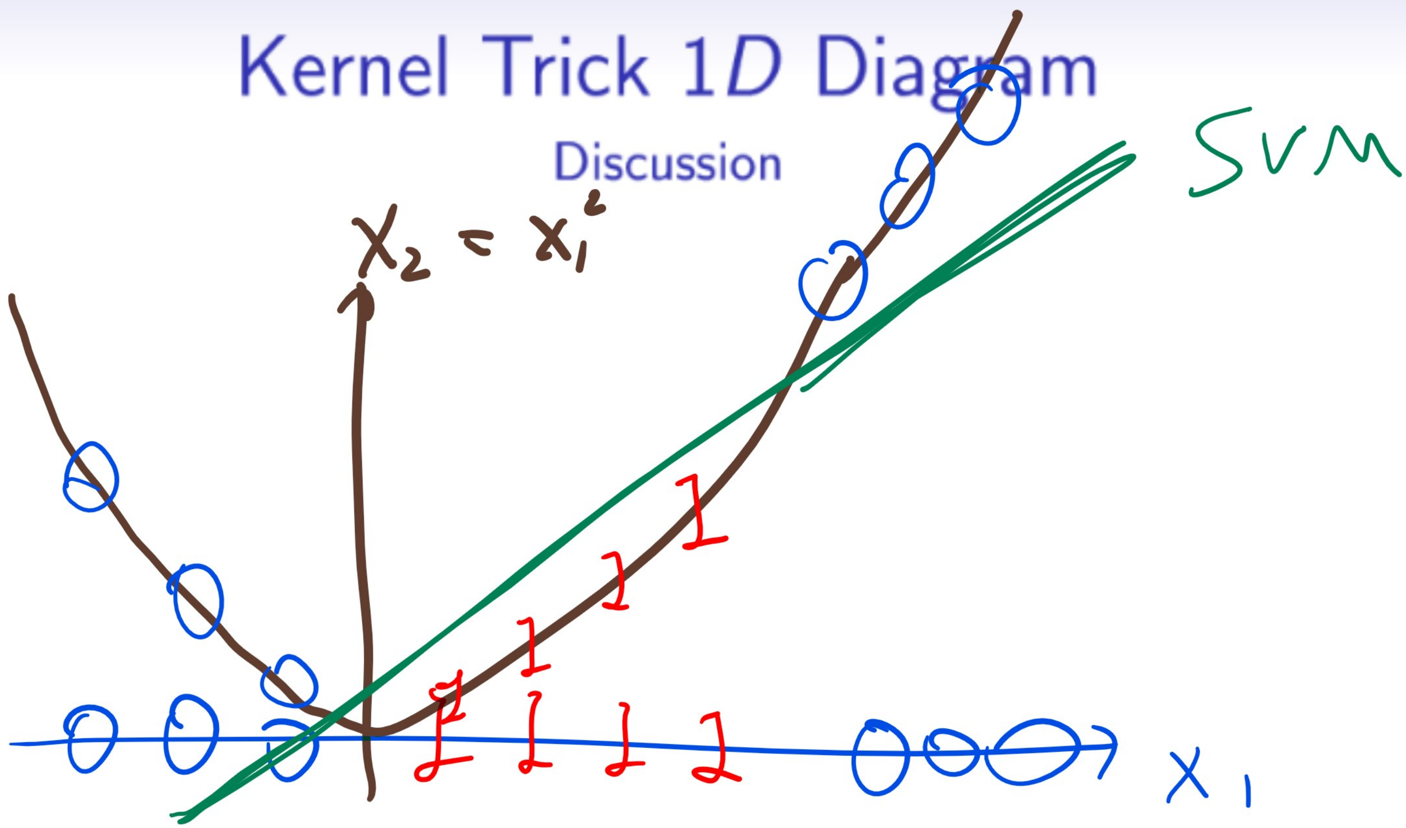
Discussion



- If the classes are not linearly separable, more features can be created.
- For example, a 1 dimensional x can be mapped to $\phi(x) = (x, x^2)$.
- Another example is to map a 2 dimensional (x_1, x_2) to $\phi(x = (x_1, x_2)) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$.

Kernel Trick 1D Diagram

Discussion



Kernelized SVM

Discussion

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

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

Kernel Matrix

Discussion

- The kernel is usually represented by a $n \times n$ matrix K called the Gram matrix.

data point j

$$K_{ij} = \phi(x_i)^T \phi(x_j)$$

data point instance i

Examples of Kernel Matrix

Discussion

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

$$K_{ij} = \left(x_i^T x_j\right)^2$$

- Another example is the quadratic kernel $K_{ij} = \left(x_i^T x_j + 1\right)^2$. It can be factored to have the following feature representations.

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

Kernel Matrix Characterization

Discussion

- A matrix K is kernel (Gram) matrix if and only if it is symmetric positive semidefinite.

Popular Kernels

Discussion

- Other popular kernels include the following.

1. Linear kernel: $K_{ij} = x_i^T x_j$ ← SVM
2. Polynomial kernel: $K_{ij} = \left(x_i^T x_j + 1\right)^d$
3. Radial Basis Function (Gaussian) kernel:
$$K_{ij} = \exp\left(-\frac{1}{\sigma^2}\left(x_i - x_j\right)^T\left(x_i - x_j\right)\right)$$

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

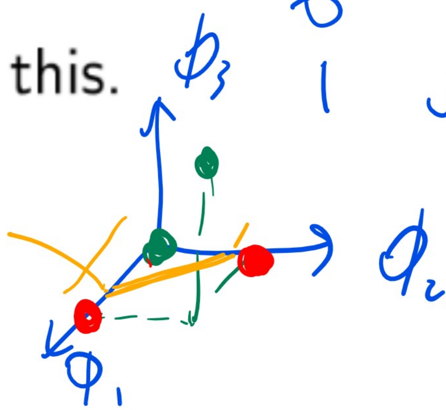
Kernel Trick for XOR

Quiz (Graded)

Q6

- March 2018 Final Q17
- SVM with quadratic kernel $\phi(x) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right)$ can correctly classify the training set for XOR.
- A: True.
- B: False.
- C: Do not choose this.
- D: Do not choose this.
- E: Do not choose this.

ϕ_1	ϕ_2	ϕ_3	x_1	x_2	y
0	0	1	0	1	1
1	0	0	1	0	1
0	0	0	0	0	0
1	$\sqrt{2}$	1	1	1	0



Kernel Matrix

Quiz (Graded)

Q29

- Fall 2009 Final Q2
- What is the feature vector $\phi(x)$ induced by the kernel $K_{ij} = \exp(x_i + x_j) + \sqrt{x_i x_j} + 3$?
- 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

$$K_{ij} = \phi(x_i)^T \phi(x_j)$$

$$D: \left(\sqrt{\exp(x_i)}, \sqrt{x_i}, \sqrt{3}\right)^T$$

$$\left(\sqrt{\exp(x_j)}, \sqrt{x_j}, \sqrt{3}\right)$$

$$\Downarrow$$

$$\overset{?}{=} K_{ij}$$

$$\exp(x_i + x_j) + \sqrt{x_i x_j} + \cancel{\sqrt{3}}\sqrt{3}$$

3