CS540 Introduction to Artificial Intelligence
Lecture 1

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

May 23, 2019

Grading

Admin

- Quizzes: best 10 of 11 weeks, 2 points each.
- Math homework: use them to replace quiz grades.
- Programming homework: best 10 of 11 weeks, 4 points each.
- Exams: one midterm and one final, 20 points each.

Quizzes

Admin

- Download Socrative, the room number is CS540S1 or CS540S2.
- Default login for Socrative is your wisc email ID.
- If someone else tries to hack your account, please email or post on Piazza.
- 1 point for Participation questions.
- 1 point for Graded questions.
- Points rounded up to one of $\{0, 0.5, 1, 1.5, 2\}$.
- Quiz questions can show up any time during the lecture.

Test

Quiz (Graded)

- A: Don't choose this
- B: Don't choose this
- C: Choose this
- D: Don't choose this
- E: Don't choose this

Quizzes on Canvas

Admin

- If there is any problem with your device or the app, you can submit your answers on paper or Canvas before the end of the lecture.

Guess Average Game

Quiz (Participation)

- Write down an integer between 0 and 100 that is the closest to two thirds (2/3) of the average of everyone's (including yours) integers.
- A: 0 – 20
- B: 21 – 40
- C: 41 – 60
- D: 61 – 80
- E: 81 – 100

Guess Average Game, Again

Quiz (Participation)

- Write down an integer between 0 and 100 that is the closest to two thirds (2/3) of the average of everyone's (including yours) integers.
- A: 0 – 10
- B: 11 – 20
- C: 21 – 30
- D: 31 – 60
- E: 61 – 100

Math Homework

Admin

- Due in 1 week Sunday (Monday morning is okay).
- Grade yourself: one of $\{1, 1.5, 2\}$
- 1 means you attempted something but you know it's completely incorrect.
- 1.5 means you attempted something but you know it's not completely correct.
- 2 means you think everything is correct and you give me permission to share it with other students as a sample solution.
- Put 2.5 if you already got 2 for the Quiz and just want to me to share your (hopefully) correct solutions with other students.

Programming Homework

Admin

- Due in 1 week Sunday (if you don't want spoilers).
- Can submit any time before Sunday in 3 weeks (we will post our solutions in Java, Python, or Matlab after the 1 week due date).
- You can fix your code and output and resubmit after the due dates to replace the previous grade.
- 2 points for output (auto-graded).
- 2 points for code (only check for correctness and plagiarism).
- You can submit output without code to get 2 if you use (steal) code from other people.
- If you are caught submitting someone else's code or output, you cannot resubmit.

Quiz (Participation)

Favorite Programming Language

- What is your favorite programming language (choose one)?
- A: Java
- B: Python
- C: Matlab
- D: Other
- E: None

Midterm and Final

Admin

- Two alternative dates, attend either one. The second one is harder.
- 40 Multiple Choice questions: around half will be math and statistics related questions, the other half will be algorithm related questions.

(Not recommended) Ways to Get B+

Admin

- Not attending any lecture and not doing any math homework.
- Not learning any math and statistic for exams.
- Not attending one of the exams.
- Not doing any programming: use the code from other people every week.

Only Way to Get A

Admin

- Do everything.

Textbook

Admin

- SS is available for free online.
- If you are planning to take $760, 761, 861$ in the future, it is highly recommended that you read the first few chapters of this book.
- Otherwise, you can skip all the error bound, VC dimension related materials.

Admin

Admin

- Math and Stat Review posted under W1.
- Complete slides (with diagrams and quiz questions etc) will be posted Thursday or Friday.
- Homework will be posted on Friday (due in 9 days, not 2 days).
- Exact due dates are on Canvas: programming homework can be submitted two weeks late (except for the last two homework (one week late)).

Questions

Admin

- Questions?

Supervised Learning

Motivation

- Supervised learning:

Data	Features (Input)	Output	-
Sample	$\{(x_{i1}, ..., x_{im})\}_{i=1}^{n}$	$\{y_i\}_{i=1}^{n}$	find "best" $\hat{f}$
-	observable	known	-
New	$(x_1', ..., x_m')$	y'	guess $\hat{y} = \hat{f}(x')$
-	observable	unknown	-

Training and Test Sets

Motivation

- Supervised learning:

Data	Features (Input)	Output	-
Training	$\{(x_{i1}, ..., x_{im})\}_{i=1}^{n'}$	$\{y_i\}_{i=1}^{n'}$	find "good" $\hat{f}$
-	observable	known	-
Validation	$\{(x_{i1}, ..., x_{im})\}_{i=n'}^{n}$	$\{y_i\}_{i=n'}^{n}$	find "best" $\hat{f}$
-	observable	known	-
Test	$(x'_1, ..., x'_m)$	y'	guess $\hat{y} = \hat{f}(x')$
-	observable	unknown	-

Loss Function

Motivation

- An objective function is needed to select the "best" $\hat{f}$. An example is the squared distance between the predicted and the actual y value:

$$\hat{f} = \arg\min_{f} \boxed{\frac{1}{2}\sum_{i=1}^{n}\left(f\left(x_i\right) - y_i\right)^2}$$

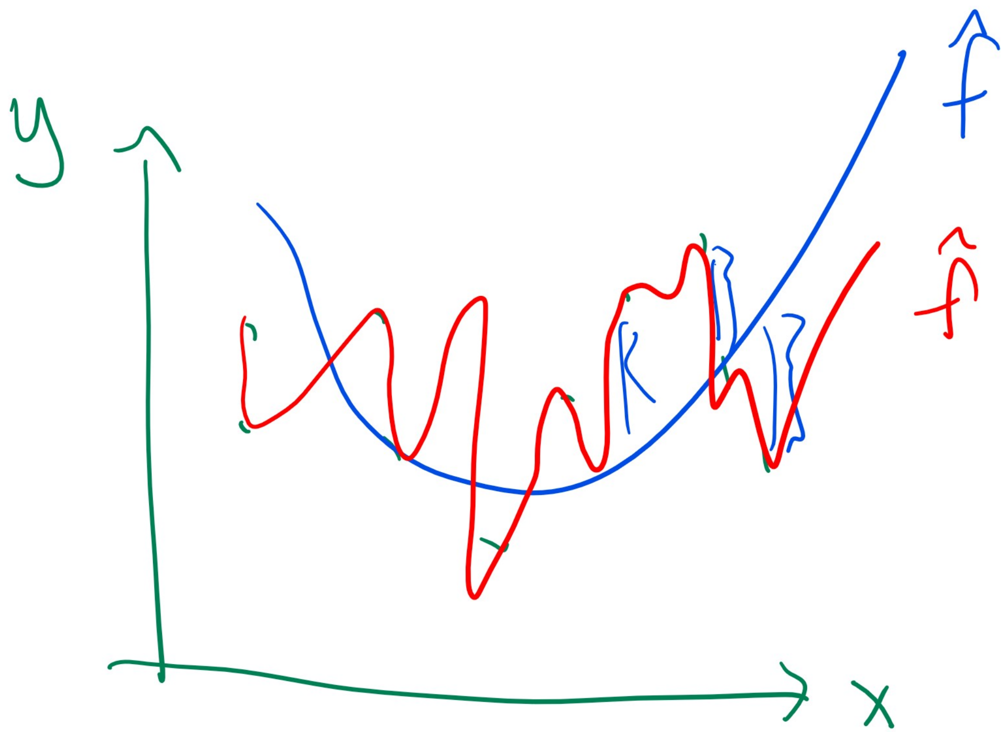
actual

prediction

- The objective function is called the cost function (or the loss function), and the objective is to minimize the cost.
- A training data point x_i is also called an instance.

Function Space Diagram

Motivation



Hypothesis Space

Motivation

- There are too many functions to choose from.
- There should be a smaller set of functions to choose $\hat{f}$ from.

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \frac{1}{2} \sum_{i=1}^{n} \left(f\left(x_i\right) - y_i \right)^2$$

- The set $\mathcal{H}$ is called the hypothesis space.

Linear Regression

Motivation

- For example, $\mathcal{H}$ can be the set of linear functions. Then the problem can be rewritten in terms of coefficients (parameters).

$$\left(\hat{w}_1, ..., \hat{w}_m, \hat{b}\right) = \arg \min_{w_1,...,w_m,b} \frac{1}{2}\sum_{i=1}^{n}(a_i - y_i)^2$$

find $\hat{f}$

where $a_i = w_1 x_{i1} + w_2 x_{i2} + ... + w_m x_{im} + b$

- $\{w_1, ..., w_m\}$ are called weights. b is called bias.
- The problem is called (least squares) linear regression.

Activation Function

Motivation

- Suppose $\mathcal{H}$ is the set of functions that are compositions between another function g and linear functions.

$$\left(\hat{w}_0, \hat{w}_1, \ldots, \hat{w}_m, \hat{b}\right) = \arg \min_{w_1, \ldots, w_m, b} \frac{1}{2} \sum_{i=1}^{n} (a_i - y_i)^2$$

$$\text{where } a_i = g\left(w_1 x_{i1} + w_2 x_{i2} + \ldots + w_m x_{im} + b\right)$$

$g(x_{i1}, x_{i2} \ldots)$

- g is called the activation function.

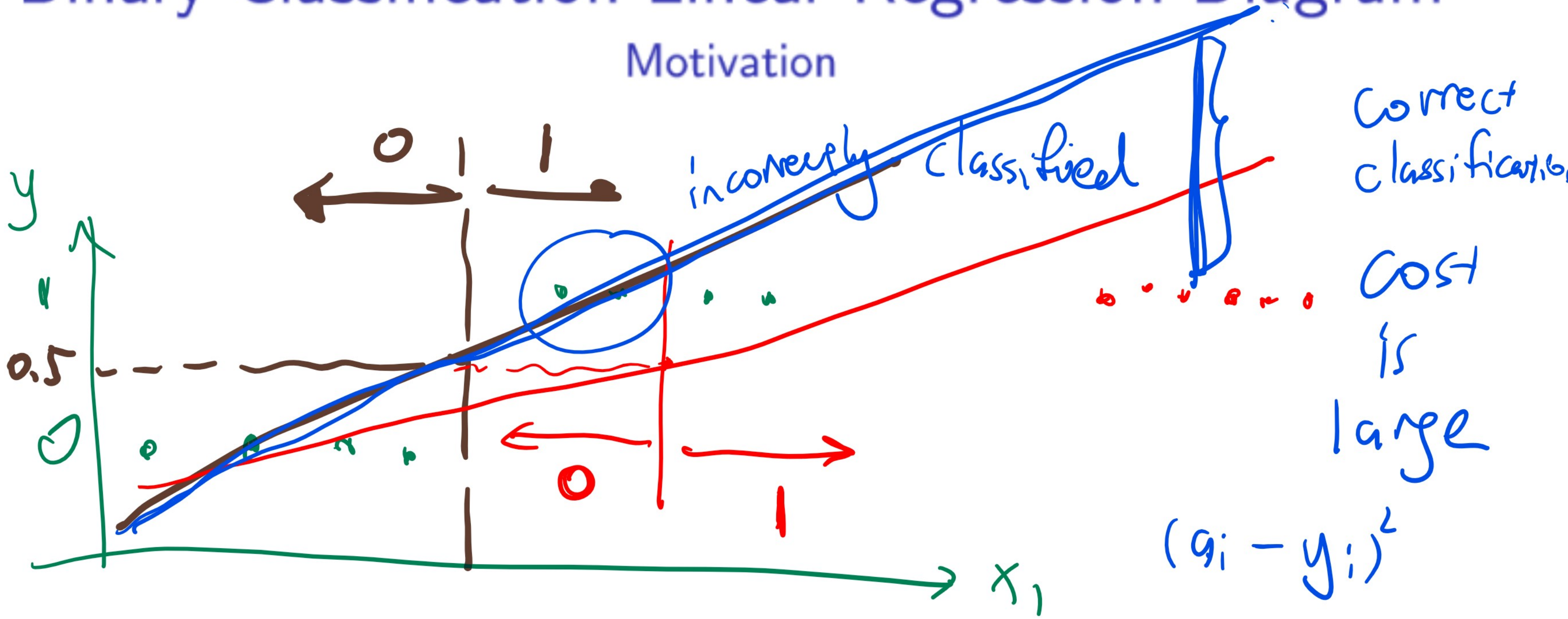
Binary Classification

Motivation

- If the problem is binary classification, y is either 0 or 1, and linear regression is not a great choice.
- This is because if the prediction is either too large or too small, the prediction is correct, but the cost is large.

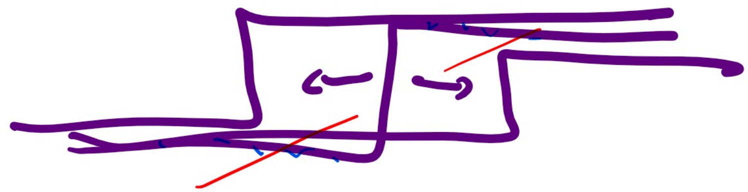
Binary Classification Linear Regression Diagram

Motivation



Linear Threshold Unit

Motivation



- One simple choice is to use the step function as the activation function:

$w_i x_{i1} + \sim \sim$

$$g\left(\boxed{\cdot}\right) = \mathbb{1}_{\{\boxed{\cdot} \geqslant 0\}} = \begin{cases} 1 & \text{if } \boxed{\cdot} \geqslant 0 \\ 0 & \text{if } \boxed{\cdot} < 0 \end{cases}$$

- This activation function is called linear threshold unit (LTU).

$$\mathbb{1}_{\{x \geq 0\}} = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}$$

same Q 5

Objective Function for LTU

Quiz (Graded)

up to constant

- Which ones (multiple) of the following functions are equivalent to the squared error for binary classification?

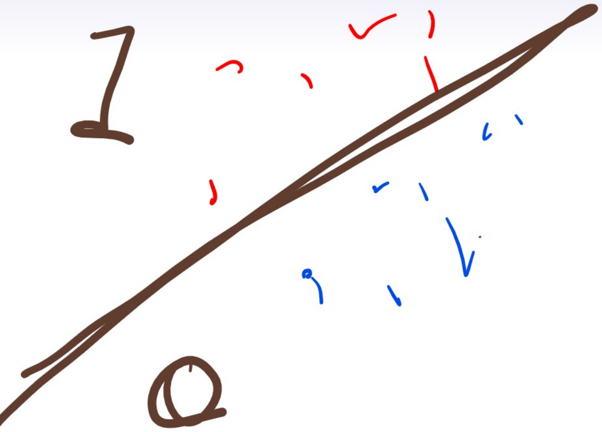
$$\begin{cases} 1 & \text{if } a_i = y_i \\ 0 & \text{if } a_i \neq y_i \end{cases}$$

$$C = \frac{1}{2}\sum_{i=1}^{n} (a_i - y_i)^2, \; y_i \in \{0, 1\}$$

y_i	a_i	C
0	0	0
1	1	0
1	0	1
0	1	1

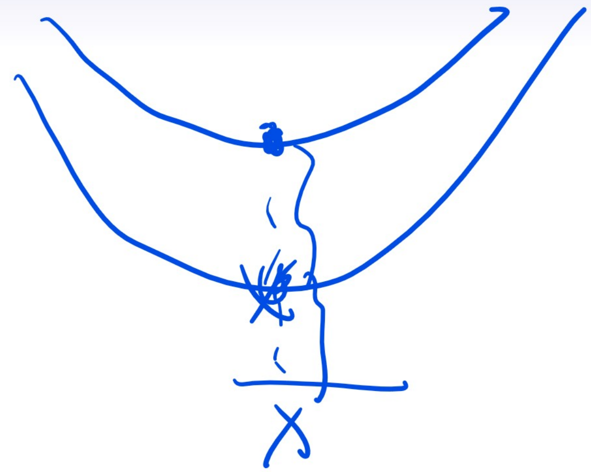
- A: $\sum \mathbb{1}_{\{a_i = y_i\}}$
- ✓ B: $\sum \mathbb{1}_{\{a_i \neq y_i\}}$
- ✓ C: $\sum |a_i - y_i|$
- D: $\sum \max\{0, 1 - a_i y_i\}$
- ✓ E: $\sum \max\{0, 1 - (2 \cdot a_i - 1)(2 \cdot y_i - 1)\}$

C (E)	y_i	a_i	C
0	0	0	0
0	1	1	0
2	1	0	1
2	0	1	1



Perceptron Algorithm

Description

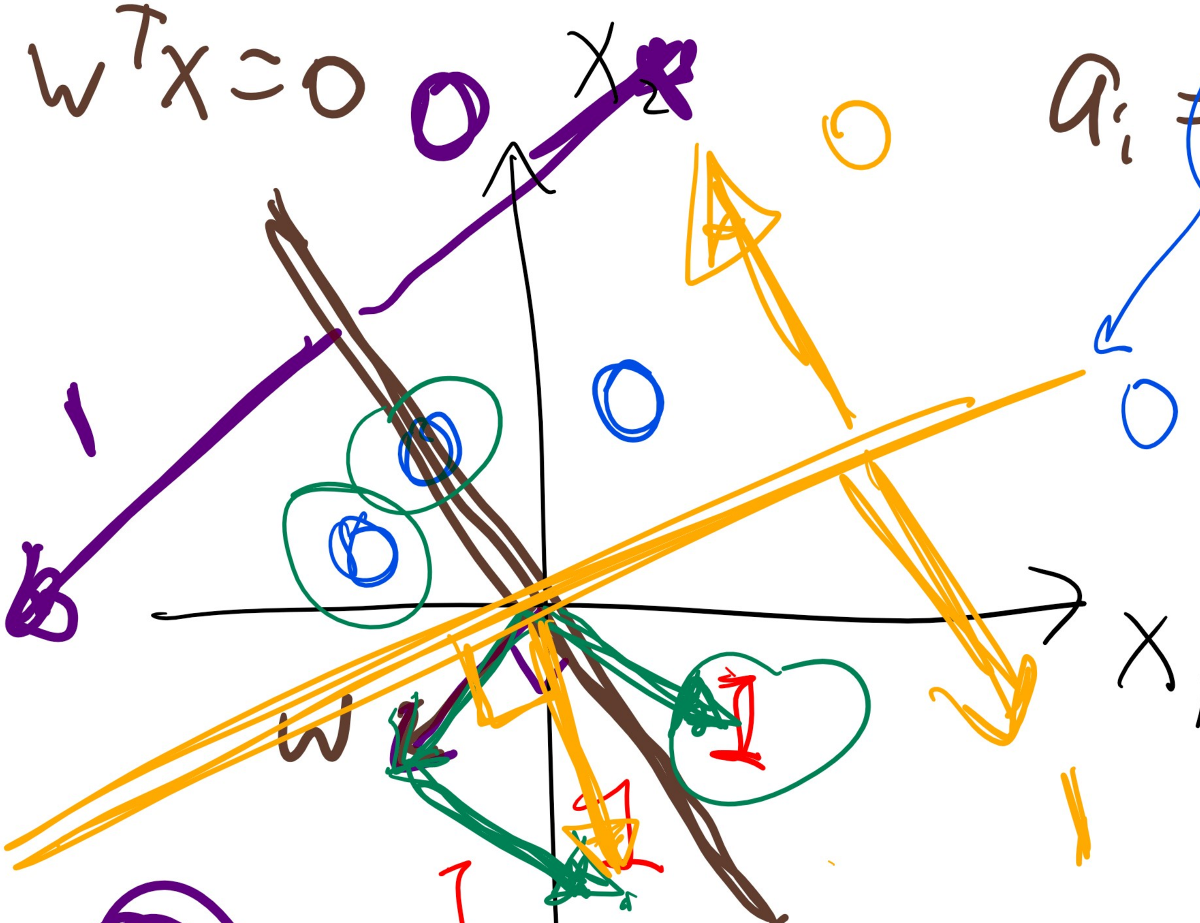


- Initialize random weights.
- Evaluate the activation function at one instance x_i to get $\hat{y}_i$.
- If the prediction $\hat{y}_i$ is 0 and actual y_i is 1, increase the weights by x_i.
- If the prediction $\hat{y}_i$ is 1 and actual y_i is 0, decrease the weights by x_i.
- Repeat for all data points and until convergent.

Perceptron Algorithm Diagram, 0 Example

Description

$w^T x = 0$



$w^T x \geq 0$

$w + x$

-2

$-1 \quad 1 \quad -1 \quad 1$

$a_i = \mathbb{1}\{w_1 x_1 + w_2 x_2 \geq 0\}$

no bias

≥ 0

$>0 \quad >0 \quad <0 \quad <0$

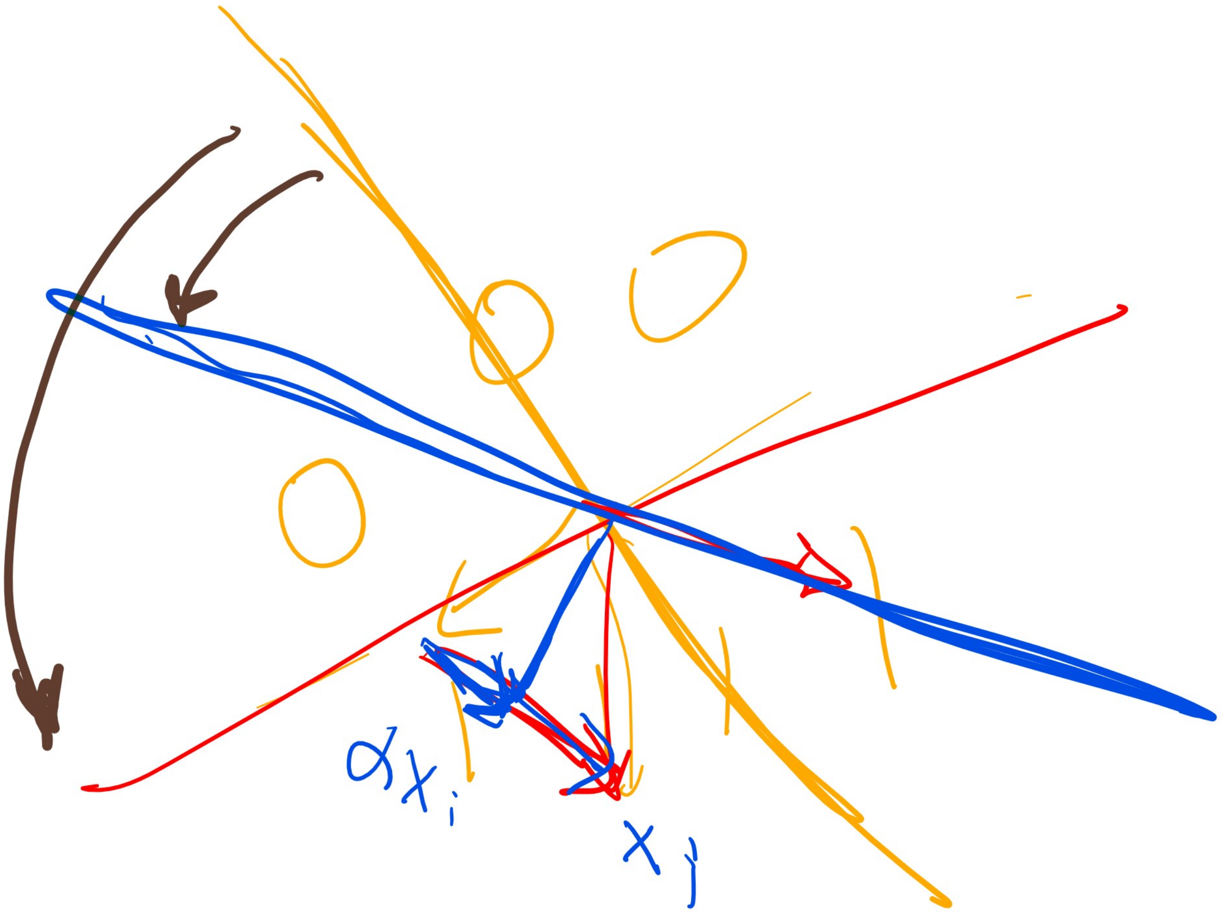
$\leq 0 \quad \leq 0 \quad \leq 0 \quad \leq 0$

$w_1 x_1 + w_2 x_2$

$w = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$ $\quad w^T x = (w_1, w_2)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

Perceptron Algorithm Diagram, 1 Example

Description



Perceptron Algorithm, Part 1

Algorithm

- Inputs: instances: $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$
- Outputs: weights and biases: $w_1, ..., w_m$, and b
- Initialize the weights,

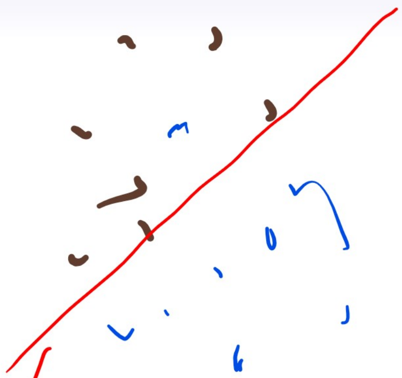
$$w_1, ..., w_m, b \sim \text{ Unif } [0, 1]$$

- Evaluate the activation function at a single data point x_i,

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

Perceptron Algorithm, Part 2

Algorithm



a_i	y_i	
0	0	
1	1	
1	0	$-x_i$
0	1	$+x_i$

- Update weights using the following rule,

$$w = w - \alpha(a_i - y_i)x_i$$
$$b = b - \alpha(a_i - y_i)$$

learning rate

- Repeat the process for every $x_i, i = 1, 2, ..., n$
- Repeat until $a_i = y_i$ for every $i = 1, 2, ..., n$

$0 < \alpha \leq 1$

if (x, y) is not linearly separable $\Longrightarrow$ not converge! (stop)

Learning Rate

Discussion

- The learning rate α controls how fast the weights are updated.
- They can be constant for each update or they can change (usually decrease) for each update.
- For perceptron learning, it is typically set to 1.

Q7

Perceptron Algorithm

Quiz (Graded)

$a_i = \mathbb{1}\{w_1 x_1 + w_2 x_2 + w_3 x_3 + b \geq 0\}$

1

0.2

- 2017 May Final Exam Q3
- Let the learning rate be $\alpha = 0.2$. Currently $w = [0.2 \quad 0.7 \quad 0.9]^T$, $b = -0.7$, and $x_i = [0 \quad 0 \quad 1]^T$ and $y_i = 0$. What is the updated weights $\begin{bmatrix} w \\ b \end{bmatrix}$?
- A: $[0 \quad 0.5 \quad 0.9 \quad -0.7]^T$
- B: $[0.2 \quad 0.7 \quad 1.1 \quad -0.5]^T$
- ✓ C: $[0.2 \quad 0.7 \quad 0.7 \quad -0.9]^T$
- D: $[0.4 \quad 0.9 \quad 0.9 \quad -0.7]^T$
- E: none of the above

$$\boxed{\begin{aligned} w &= w - \alpha(a_i - y_i)\, x_i \\ b &= b - \alpha(a_i - y_i) \end{aligned}}$$

$-0.7 - 0.2(1 - 0)$

-0.9