CS540 Introduction to Artificial Intelligence

Lecture 2

Young Wu

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

June 28, 2022

Two-thirds of the Average Game

Quiz

AI

Socrative room

CS540C

- Pick an integer between 0 and 100 (including 0 and 100) that is the closest to two-thirds of the average of the numbers other people picked.

Quizzes, Math Homework, Discussions

Admin

- Due dates: Monday, late submission withint a week or so without penalty (regrade requests).
- Share solutions ($M2$ etc): before due date (one or two days late is okay).
- Share solutions ($X1$ etc): a week before the exam.
- Group discussions: no due dates.

0.5

Office Hours, Discussion Sessions

Admin

recorded
on Zoom

- Answer M, P homework questions on Saturday evenings?
- A : Yes, I will attend.
- B : Yes, but I will not attend.
- C : No.

Supervised Learning

Motivation

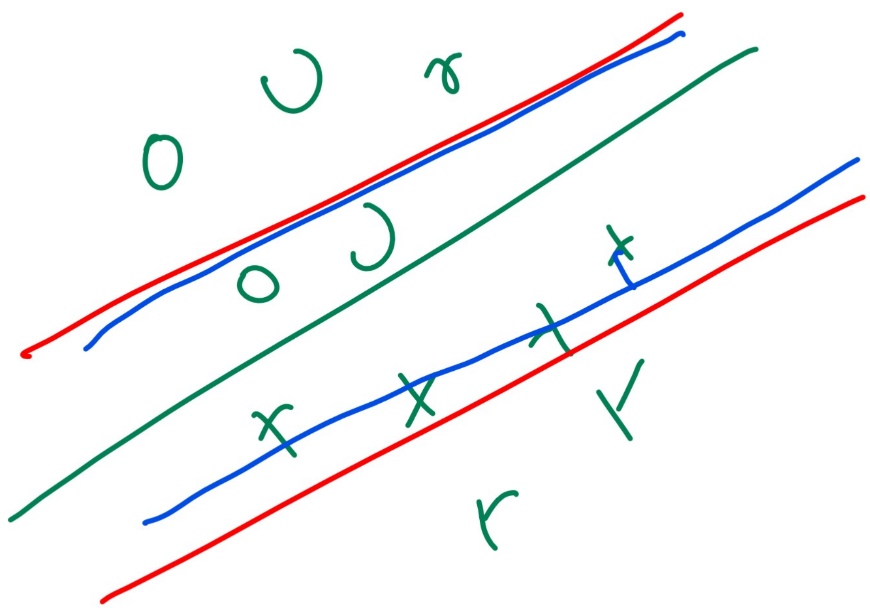
LTa

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

Loss Function Diagram

Motivation

count # mistakes.



Zero-One Loss Function

Motivation

- An objective function is needed to select the "best" $\hat{f}$. An example is the zero-one loss.

$$\hat{f} = \underset{f}{\operatorname{argmin}} \underbrace{\sum_{i=1}^{n} \mathbb{1}_{\{f(x_i) \neq y_i\}}}$$

$$\begin{cases} 1 & \text{if } f(x_i) \neq y_i \\ 0 & \text{if } f(x_i) = y_i \end{cases}$$

- $\boxed{\operatorname{argmin}_f}$ objective (f) outputs the function that minimizes the objective.
- The objective function is called the cost function (or the loss function), and the objective is to minimize the cost.

Squared Loss Function

Motivation

- Zero-one loss counts the number of mistakes made by the classifier. The best classifier is the one that makes the fewest mistakes.
- Another example is the squared distance between the predicted and the actual y value:

$$\hat{f} = \operatorname*{argmin}_{f} \frac{1}{2} \sum_{i=1}^{n} \left(f\left(x_i\right) - y_i\right)^2$$

Loss Functions Equivalence

Quiz

- Which one of the following functions is not equivalent to the squared error for binary classification?

Q3

$$C = \sum_{i=1}^{n} \left(f\left(x_i\right) - y_i\right)^2, f\left(x_i\right) \in \{0, 1\}, y_i \in \{0, 1\}$$

[0, 1]

- $\cancel{A}$: $\sum \mathbb{1}_{\{f(x_i) \neq y_i\}}$
- B : $\sum \mathbb{1}_{\{f(x_i) = y_i\}}$
- $\cancel{C}$: $\sum \left|f\left(x_i\right) - y_i\right|$
- $\boxed{D}$: $\sum \max\left\{0, 1 - f\left(x_i\right) y_i\right\}$
- $\cancel{E}$: $\sum \frac{1}{2} \max\left\{0, 1 - \left(2 \cdot f\left(x_i\right) - 1\right)\left(2 \cdot y_i - 1\right)\right\}$

choice

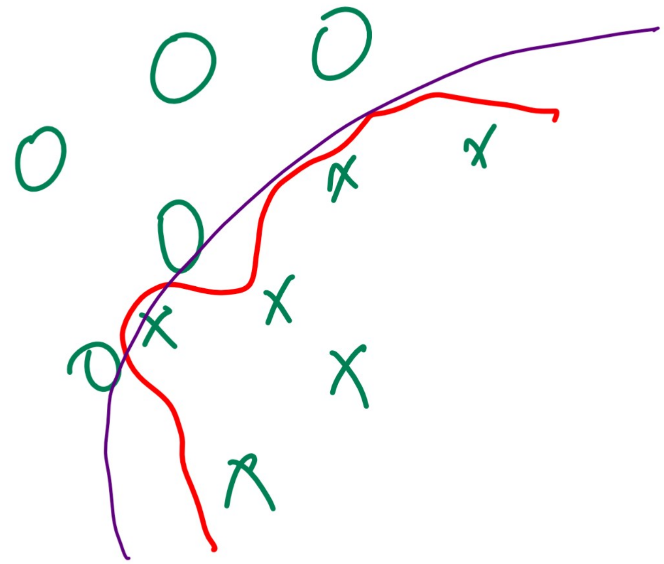
D	E	C	y_i	$f(x_i)$	sq	A	B
1	0	0	0	0	0	0	1
	1	1	0	1	1	1	0
	1	1	1	0	1	1	0
	0	0	1	1	0	0	1

Loss Functions Equivalence, Answer

Quiz

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.

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

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

Activation Function

Motivation

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

$$\left(\hat{w}, \hat{b}\right) = \underset{w,b}{\operatorname{argmin}} \frac{1}{2} \sum_{i=1}^{n} \left(a_i - y_i\right)^2$$

where $a_i = \boxed{g\left(w^T x + b\right)}$

- g is called the activation function.

Linear Threshold Unit

Motivation

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

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

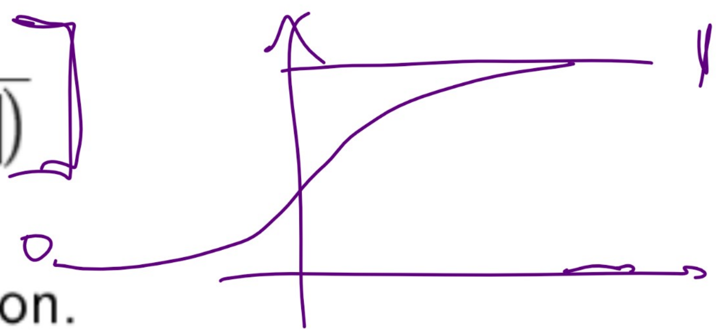
- This activation function is called linear threshold unit (LTU).

Sigmoid Activation Function

Motivation

- When the activation function g is the sigmoid function, the problem is called logistic regression.

$$g\left(\boxed{\cdot}\right) = \frac{1}{1 + \exp\left(-\boxed{\cdot}\right)}$$



- This g is also called the logistic function.

Sigmoid Function Diagram

Motivation

Cross-Entropy Loss Function

Motivation

- The cost function used for logistic regression is usually the log cost function.

$$C(f) = -\sum_{i=1}^{n} \left(y_i \log\left(f\left(x_i\right)\right) + \left(1 - y_i\right)\log\left(1 - f\left(x_i\right)\right)\right)$$

amount

Count ⊕

mistakes

- It is also called the cross-entropy loss function.

Logistic Regression Objective

Motivation

- The logistic regression problem can be summarized as the following.

$$\left(\hat{w}, \hat{b}\right) = \underset{w,b}{\operatorname{argmin}} - \boxed{\sum_{i=1}^{n} \left(y_i \log\left(a_i\right) + \left(1 - y_i\right) \log\left(1 - a_i\right)\right)}$$

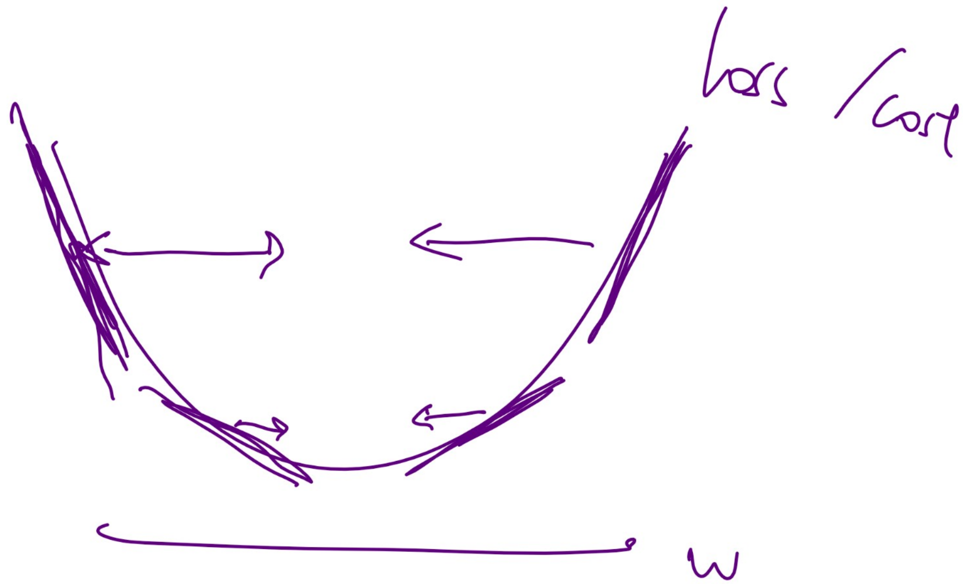
where $a_i = \dfrac{1}{1 + \exp\left(-z_i\right)}$ and $z_i = w^T x_i + b$

$\hat{y}_i$

y_i	a_i	
0	0	0
1	1	0
0	1	∞
1	0	∞

Optimization Diagram

Motivation



Opposite direction
of derivative
(gradient)

Logistic Regression

Description

- Initialize random weights.
- Evaluate the activation function.
- Compute the gradient of the cost function with respect to each weight and bias.
- Update the weights and biases using gradient descent.
- Repeat until convergent.

Gradient Descent Step

Definition

- For logistic regression, use chain rule twice.

$$w = w - \alpha \underbrace{\sum_{i=1}^{n} (a_i - y_i) x_i}$$

$$b = b - \alpha \underbrace{\sum_{i=1}^{n} (a_i - y_i)}$$

$$a_i = g\left(w^T x_i + b\right), g\left(\boxed{\cdot}\right) = \frac{1}{1 + \exp\left(-\boxed{\cdot}\right)}$$

CE

$y_i \log a_i + (1-y_i) \log(1-a_i)$

- α is the learning rate. It is the step size for each step of gradient descent.

Perceptron Algorithm

Definition

- Update weights using the following rule.

$$w = w - \alpha (a_i - y_i) x_i$$
$$b = b - \alpha (a_i - y_i)$$
$$a_i = \mathbb{1}_{\{w^T x_i + b \geqslant 0\}}$$

geometric

NOT gradient descent

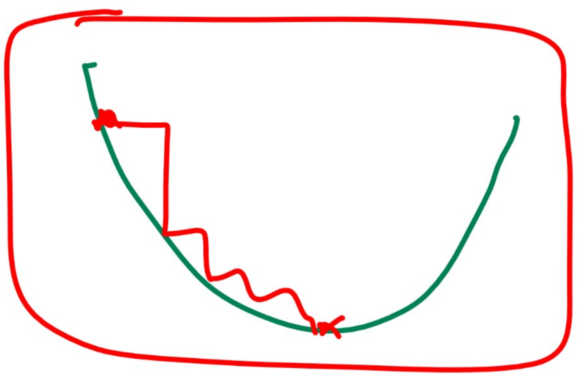
Learning Rate Diagram

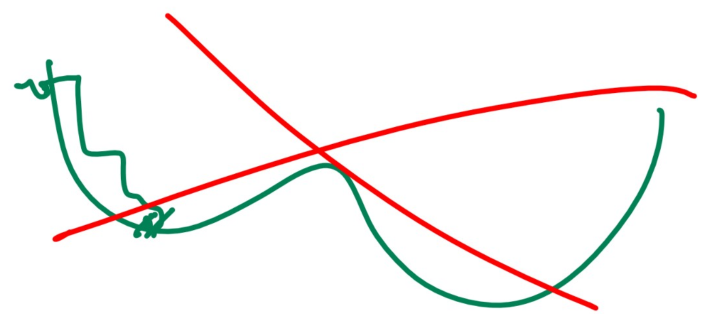
Definition

Other Non-linear Activation Function

Discussion

- Activation function: $g\left(\boxed{\cdot}\right) = \tanh\left(\boxed{\cdot}\right) = \dfrac{e^{\boxed{\cdot}} - e^{-\boxed{\cdot}}}{e^{\boxed{\cdot}} + e^{-\boxed{\cdot}}}$
- Activation function: $g\left(\boxed{\cdot}\right) = \arctan\left(\boxed{\cdot}\right)$
- Activation function (rectified linear unit): $g\left(\boxed{\cdot}\right) = \boxed{\cdot}\,\mathbb{1}_{\{\boxed{\cdot} \geqslant 0\}}$
- All these functions lead to objective functions that are convex and differentiable (almost everywhere). Gradient descent can be used.





Gradient Descent

Quiz

- What is the gradient descent step for w if the objective (cost) function is the squared error?

$w_1 x_{i1} + w_2 x_{i2} + w_3 x_{i3} + \ldots$

$$C = \frac{1}{2} \sum_{i=1}^{n} (a_i - y_i)^2, a_i = g\left(w^T x_i + b\right), g(z) = \frac{1}{1 + e^{-z}}$$

$$\frac{\partial C}{\partial w_j} = \sum_{i=1}^{n} \frac{\partial C}{\partial a_i} \frac{\partial a_i}{\partial w_j} = \sum_{i=1}^{n} (a_i - y_i) \cdot \underbrace{a_i (1 - a_i)}_{g'(z)} \cdot x_{ij}$$

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

Gradient Descent, Answer

Quiz

$$\frac{\partial}{\partial z}\underbrace{\frac{1}{1+e^{-z}}}_{a_i} = \frac{+e^{-z}}{(1+e^{-z})^2} = \frac{e^{-z}}{1+e^{-z}}\cdot\frac{1}{1+e^{-z}}$$

$$= \left(1-\frac{1}{1+e^{-z}}\right)\left(\frac{1}{1+e^{-z}}\right)$$

$$= (1-a_i)\,a_i$$

Gradient Descent, Answer Too

Quiz

Gradient Descent

Quiz

- What is the gradient descent step for w if the objective (cost) function is the squared error?

$$C = \frac{1}{2} \sum_{i=1}^{n} \left(a_i - y_i\right)^2, a_i = g\left(w^T x_i + b\right), g'\left(z\right) = g\left(z\right) \cdot \left(1 - g\left(z\right)\right)$$

- $A : w = w - \alpha \sum \left(a_i - y_i\right)$
- $B : w = w - \alpha \sum \left(a_i - y_i\right) x_i$
- $C : w = w - \alpha \sum \left(a_i - y_i\right) a_i x_i$
- $D : w = w - \alpha \sum \left(a_i - y_i\right) \left(1 - a_i\right) x_i$
- $E : w = w - \alpha \sum \left(a_i - y_i\right) a_i \left(1 - a_i\right) x_i$

Gradient Descent, Another One, Answer

Quiz

Gradient Descent, Another One Too

Quiz

- What is the gradient descent step for w if the activation function is the identity function?

Q4

$$C = \frac{1}{2}\sum_{i=1}^{n}(a_i - y_i)^2, a_i = w^T x_i + b$$

least square
linear regression.

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

- $A : w = w - \alpha \sum (a_i - y_i)$
- $B : w = w - \alpha \sum (a_i - y_i)\, x_i$

cross-entropy loss
+ logistic activation

- $C : w = w - \alpha \sum (a_i - y_i)\, a_i x_i$
- $D : w = w - \alpha \sum (a_i - y_i)\,(1 - a_i)\, x_i$
- $E : w = w - \alpha \sum (a_i - y_i)\, a_i\,(1 - a_i)\, x_i$

squared loss
+ logistic activation

Gradient Descent, Another One Too, Answer

Quiz

$$C = \underbrace{\frac{1}{2}\sum (a_i - y_i)^2} \qquad a_i = w^T x_i + b$$

$$\frac{\partial C}{\partial w_j} = \frac{\partial}{\partial w_j} \sum_{i=1}^{n} \Big(\underbrace{\boxed{\sum_{j=1}^{m} w_j x_j + b}}_{a_i} - y_i \Big)^2$$

$a_i \rightarrow w_1 x_{i1} + w_2 x_{i2} + w_3 x_{i3}$

$$= \sum_{i=1}^{n} \frac{\partial C}{\partial a_i} \frac{\partial a_i}{\partial w_j}$$

$$= \sum_{i=1}^{n} \cancel{\frac{1}{2}} \cdot \cancel{2} (a_i - y_i) x_{ij}$$

ith item

jth feature.

$$\nabla_w C = \begin{bmatrix} \frac{\partial C}{\partial w_1} \\ \frac{\partial C}{\partial w_2} \\ \frac{\partial C}{\partial w_3} \end{bmatrix} = \sum_{i=1}^{n} (a_i - y_i) \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{i3} \end{pmatrix} = \sum_{i=1}^{n} (a_i - y_i) x_i$$

Convexity Diagram

Discussion