CS540 Introduction to Artificial Intelligence
Lecture 2

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

June 3, 2020

Test

Quiz

$$\frac{2}{3}\frac{1}{n}\sum_{i=1}^{n} x_i$$

- Pick the number that is the closest to two-thirds of the average of the numbers other people picked.
- A: 0
- B: 1 ← 0.85
- C: 2
- D: 3
- E: 4

Socrative Room

CS540C

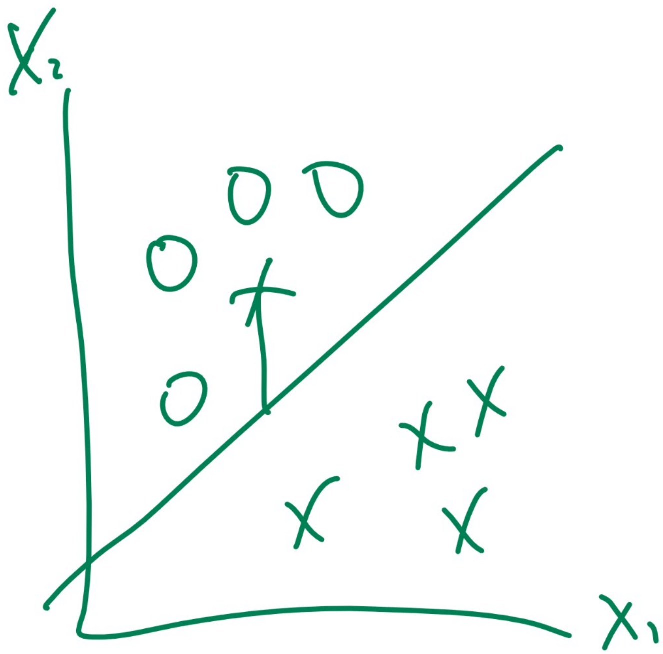
______@wisc.edu

↖ ID

Send all question to "Questions" on public.

Loss Function Diagram

Motivation



min # mistakes.

LTU Perceptron

min loss / cost

Zero-One Loss Function

Motivation

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

1 mistake

0 mistake

$$\hat{f} = \arg\min_{f} \sum_{i=1}^{n} \mathbb{1}_{\{f(x_i) \neq y_i\}}$$

- $\arg\min_{f}$ 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:

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

prediction   label

training

Loss Functions Equivalence

Quiz

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

Q2

$0, 1$

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

	$(f(x_i) - y_i)^2$	$f(x_i)$	y_i	D	$1\{f(x_i) \neq y_i\}$
1	0	0	0	X	0
0	1	0	1	1	1
0	1	1	0	1	1
1	0	1	1	0	0

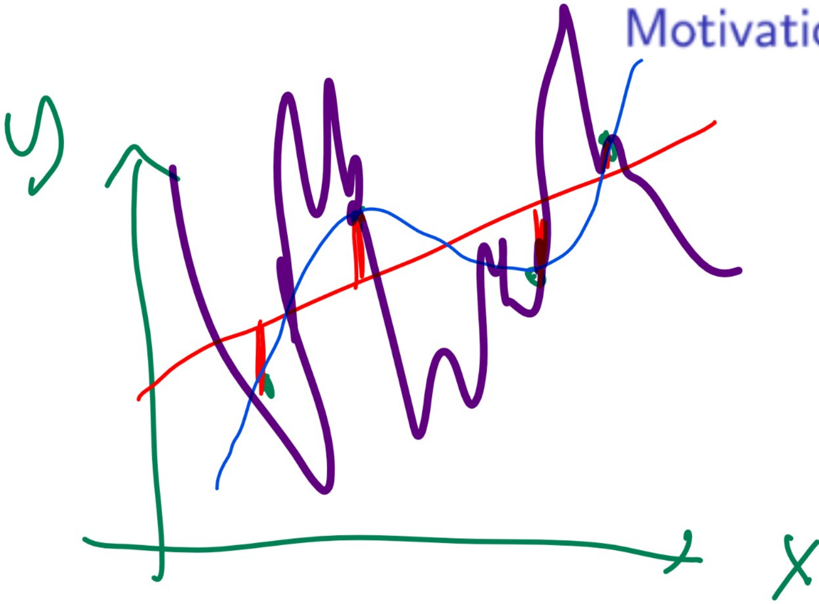
- A: $\sum \mathbb{1}_{\{f(x_i) \neq y_i\}}$
- ~~B:~~ $\sum \mathbb{1}_{\{f(x_i) = y_i\}}$
- C: $\sum |f(x_i) - y_i|$
- ~~D:~~ $\sum \max\{0, 1 - f(x_i)\, y_i\}$

0,1 → +1, −1

- E: $\sum \max\{0, 1 - (2 \cdot f(x_i) - 1)(2 \cdot y_i - 1)\}$

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.

Activation Function

Motivation

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

bias

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

weights

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

plane

surface

- g is called the activation function.

Linear Threshold Unit

Motivation

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

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

$w^T x + b$

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

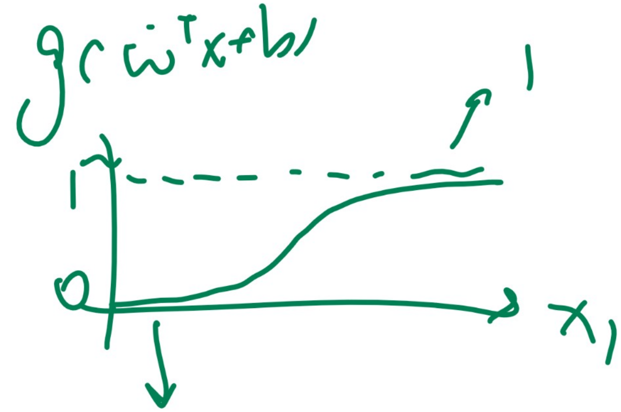
in LR

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

$$C = (y_i - f(x_i))^2$$

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

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

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

Logistic Regression Objective

Motivation

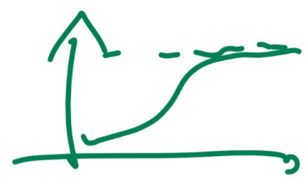
- The logistic regression problem can be summarized as the following.

P1

0, 1

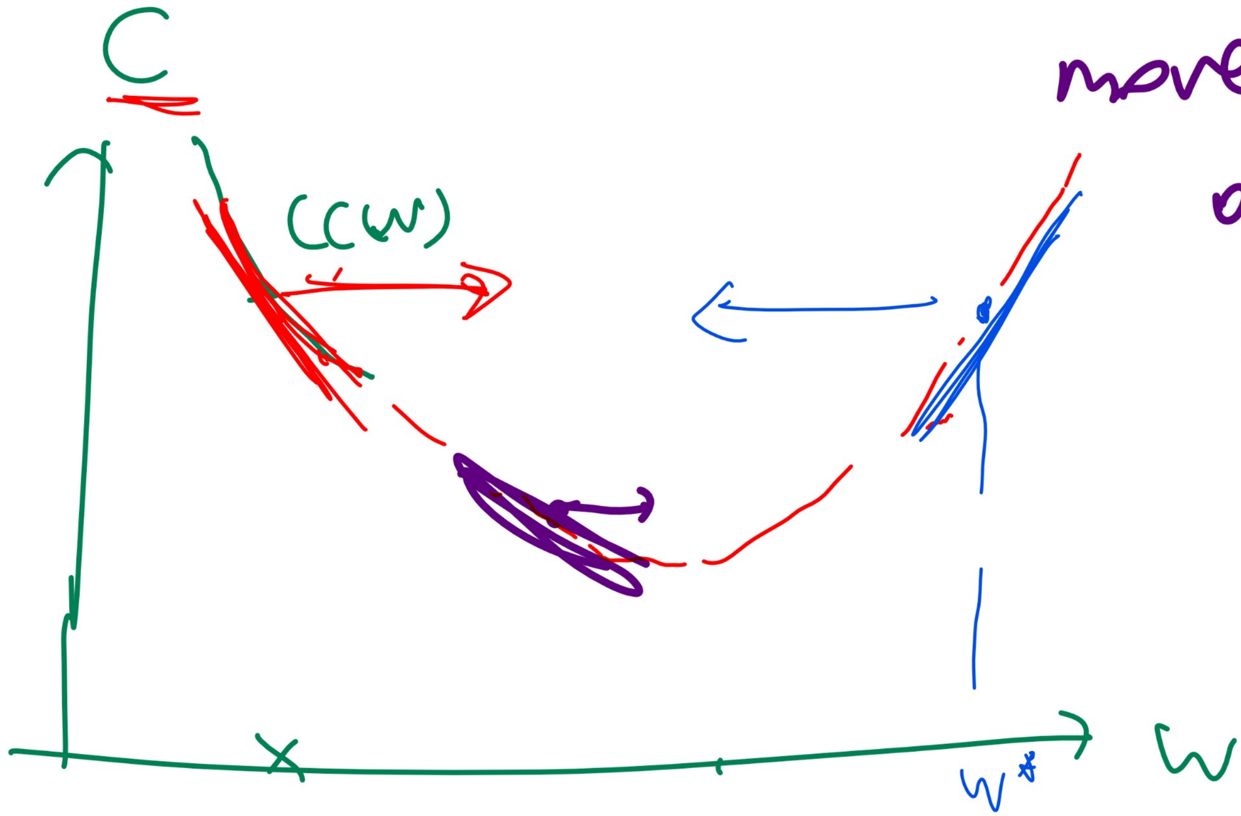
$$\left(\hat{w}, \hat{b}\right) = \arg\min_{w,b} - \sum_{i=1}^{n} \left(y_i \log\left(a_i\right) + (1 - y_i) \log\left(1 - a_i\right)\right)$$

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



Optimization Diagram

Motivation



move in opposite direction of derivative

Learning Rate Demo

Motivation

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.

$$a_i = g\,(x^T w + b)$$

Logistic Gradient Derivation 1

Definition

$n \rightarrow$ # training

min $C = -\sum_{i=1}^{n} y_i \log a_i + (1-y_i)\log(1-a_i)$

$a_i = \frac{1}{1+e^{-(w^T x_i + b)}} = (1+e^{-(w^T x_i + b)})^{-1}$

$x_1 \quad x_m$

$w_1, w_2 \ldots w_m$

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

$$= -\sum_{i=1}^{n}\left(\frac{y_i}{a_i} - \frac{1-y_i}{1-a_i}\right)\cdot \boxed{\frac{+e^{-(w^T x_i + b)}}{(1+e^{-(w^T x_i + b)})^2}}\, x_{ij}$$

Logistic Gradient Derivation 2

Definition

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

$$a_i = \frac{1}{1+e^{-(w^T x_i + b)}} \qquad 1 - a_i = \frac{e^{-(w^T x + b)}}{1+e^{-(w^T x + b)}}$$

$$= \boxed{\sum_{i=1}^{n} (a_i - y_i)\, x_j}$$

Gradient Descent Step

Definition

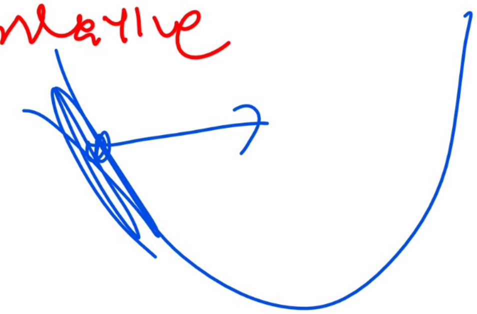
- For logistic regression, use chain rule twice.

P1

$$w = w - \alpha \boxed{\sum_{i=1}^{n} (a_i - y_i)\, x_i} \rightarrow \text{derivative}$$

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

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



- 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.

$$w = w - \alpha \boxed{(a_i - y_i)\, x_i}$$

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

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

geometric Trick

$$:= \begin{cases} 1 & \text{if } w^T x + b \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Gradient Descent

Quiz

Q3

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

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

$$\frac{1}{1+e^{-w^T x_i + b}} \Rightarrow \frac{e^{-w^T x_i + b}}{1+e^{-w^T x_i + b}} x_i$$

$$a_i(1-a_i)$$

$$\frac{\partial C}{\partial w} = \frac{\partial C}{\partial a} \cdot \frac{\partial a}{\partial w}$$

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

$$\begin{bmatrix} \frac{\partial C}{\partial w_1} \\ \frac{\partial C}{\partial w_2} \\ \frac{\partial C}{\partial w_3} \end{bmatrix}$$

- A: $w = w - \alpha \sum \left(a_i - y_i\right)$
- B: $w = w - \alpha \sum \left(a_i - y_i\right) x_i$ ← cross entropy loss.
- 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: $w = w - \alpha \sum \left(a_i - y_i\right) \boxed{a_i \left(1 - a_i\right) x_i}$

Gradient Descent, Answer

Quiz

Gradient Descent, Another One

Quiz

Q4

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

linear regression

least squares

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

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

- A: $w = w - \alpha \boxed{\sum (a_i - y_i)}$ → $\frac{\partial C}{\partial w}$
- B: $w = w - \alpha \sum (a_i - y_i)\, x_i$
- 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: $w = w - \alpha \sum (a_i - y_i)\, a_i\,(1 - a_i)\, x_i$

M1 → M12

top 10

P1 → P6

top 5

Gradient Descent, Another One, Answer

Quiz

$\cancel{g^{-1}(a) = w^T x + b}$

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

$$w_1 x_1 + w_2 x_2 + \dots + b$$

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

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

$$\begin{bmatrix} \frac{\partial C}{\partial w_1} \\ \frac{\partial C}{\partial w_2} \end{bmatrix}$$

$$\frac{\partial C}{\partial w_j} = \sum_{i=1}^{n}(a_i - y_i)x_{ij}$$

$$\frac{\partial C}{\partial w} = \nabla_w C = \sum_{i=1}^{n}(a_i - y_i)x_i$$

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): $g\left(\boxed{\cdot}\right) = \boxed{\cdot}\mathbb{1}_{\left\{\boxed{\cdot}\geqslant 0\right\}}$
- All these functions lead to objective functions that are convex and differentiable (almost everywhere). Gradient descent can be used.