CS540 Introduction to Artificial Intelligence

Lecture 2

Young Wu

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

May 23, 2022

Two-thirds of the Average Game

Quiz

Make Up Lectures

Admin

Supervised Learning

Motivation

Loss Function Diagram

Motivation

Zero-One Loss Function

Motivation

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

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

- 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

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.

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

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

- 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.

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

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

Logistic Regression Objective

Motivation

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

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

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

Optimization Diagram

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.

Gradient Descent Step

Definition

- For logistic regression, use chain rule twice.

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

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

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

- 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.

$$w = w - \alpha \left(a_i - y_i \right) x_i$$

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

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

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

Gradient Descent

Quiz

Gradient Descent, Answer

Quiz

Gradient Descent, Answer Too

Quiz

Gradient Descent

Quiz

Gradient Descent, Another One, Answer

Quiz

Gradient Descent, Another One Too

Quiz

Gradient Descent, Another One Too, Answer

Quiz

Convexity Diagram

Discussion

Questions?

Admin