Deep Learning

Vazgen Mikayelyan

YSU, Krisp

September 23, 2020

Outline

1. Gradient Descent

2. Linear and Logistic Regressions

3. Softmax Classifier

Gradient Descent

Let $f : \mathbb{R}^k \to \mathbb{R}$ be a convex function and we want to find its global minimum.

Gradient Descent

Let $f : \mathbb{R}^k \to \mathbb{R}$ be a convex function and we want to find its global minimum. This optimization algorithm is based on the fact that the fastest decreasing direction of the function is the opposite direction of gradient:

$$x_{n+1} = x_n - \alpha \nabla f\left(x_n\right)$$

and $x_0 \in \mathbb{R}^k$ is a arbitrary point.

Outline

Linear Regression

Let $(x_i, y_i)_{i=1}^{n}$, $x_i \in \mathbb{R}^k$, $y_i \in \mathbb{R}$ be our training data.

Linear Regression

Let $(x_i, y_i)_{i=1}^{n}$, $x_i \in \mathbb{R}^k$, $y_i \in \mathbb{R}$ be our training data. Consider the function

$$f(x) = f\left(x^1, x^2, \ldots, x^k\right) = w^1 x^1 + w^2 x^2 + \ldots + w^k x^k + b = w^T x + b.$$

Linear Regression

Let $(x_i, y_i)_{i=1}^n$, $x_i \in \mathbb{R}^k$, $y_i \in \mathbb{R}$ be our training data. Consider the function

$$f(x) = f\left(x^1, x^2, \ldots, x^k\right) = w^1 x^1 + w^2 x^2 + \ldots + w^k x^k + b = w^T x + b.$$

Our aim is to find parameters $b, w^1, w^2, \ldots, w^k$ such that

$$f(x_i) \approx y_i, i = 1, \ldots, n.$$

Linear Regression

Let $(x_i, y_i)_{i=1}^{n}$, $x_i \in \mathbb{R}^k$, $y_i \in \mathbb{R}$ be our training data. Consider the function

$$f(x) = f\left(x^1, x^2, \ldots, x^k\right) = w^1 x^1 + w^2 x^2 + \ldots + w^k x^k + b = w^T x + b.$$

Our aim is to find parameters $b, w^1, w^2, \ldots, w^k$ such that

$$f(x_i) \approx y_i, i = 1, \ldots, n.$$

We choose L^2 distance as our loss function:

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

Questions

1. Should we minimize the loss function using gradient descent?

Questions

1. Should we minimize the loss function using gradient descent?
2. Can you represent this model as a neural network?

Questions

1. Should we minimize the loss function using gradient descent?
2. Can you represent this model as a neural network?
3. Can we solve a classification problem using the model described above?

Logistic Regression

Let $(x_i, y_i)_{i=1}^n$, $x_i \in \mathbb{R}^k$, $y_i \in \{0, 1\}$ be our training data.

Logistic Regression

Let $(x_i, y_i)_{i=1}^n$, $x_i \in \mathbb{R}^k$, $y_i \in \{0, 1\}$ be our training data. Consider the function

$$f(x) = f\left(x^1, x^2, \ldots, x^k\right) = \sigma\left(w^1 x^1 + w^2 x^2 + \ldots + w^k x^k + b\right)$$
$$= \sigma(w^T x + b).$$

Logistic Regression

Let $(x_i, y_i)_{i=1}^n$, $x_i \in \mathbb{R}^k$, $y_i \in \{0,1\}$ be our training data. Consider the function

$$f(x) = f\left(x^1, x^2, \ldots, x^k\right) = \sigma\left(w^1 x^1 + w^2 x^2 + \ldots + w^k x^k + b\right)$$
$$= \sigma(w^T x + b).$$

Our aim is to find parameters $b, w^1, w^2, \ldots, w^k$ such that

$$f(x_i) \approx y_i, i = 1, \ldots, n.$$

Logistic Regression

Let $(x_i, y_i)_{i=1}^n$, $x_i \in \mathbb{R}^k$, $y_i \in \{0, 1\}$ be our training data. Consider the function

$$f(x) = f\left(x^1, x^2, \ldots, x^k\right) = \sigma\left(w^1 x^1 + w^2 x^2 + \ldots + w^k x^k + b\right)$$

$$= \sigma(w^T x + b).$$

Our aim is to find parameters $b, w^1, w^2, \ldots, w^k$ such that

$$f(x_i) \approx y_i, i = 1, \ldots, n.$$

We choose cross entropy distance as our loss function:

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

Questions

1. Can you represent this model as a neural network?

Questions

1. Can you represent this model as a neural network?
2. Why do we use the function sigmoid in this case?

Questions

1. Can you represent this model as a neural network?
2. Why do we use the function sigmoid in this case?
3. Why don't we use L^2 distance in this case?

Questions

1. Can you represent this model as a neural network?
2. Why do we use the function sigmoid in this case?
3. Why don't we use L^2 distance in this case?
4. Can we do logistic regression when number of classes is greater than 2?

L1 and L2 Regularizations

L1 and L2 Regularizations

In linear regression instead of L^2 loss we use one from this two:

$$\frac{1}{n}\sum_{i=1}^{n}\left(f\left(x_i\right)-y_i\right)^2+\lambda\sum_{j=1}^{k}\left|w_i\right|,$$

$$\frac{1}{n}\sum_{i=1}^{n}\left(f\left(x_i\right)-y_i\right)^2+\lambda\sum_{j=1}^{k}w_i^2.$$

Outline

Softmax Classifier

Let $(x_i, y_i)_{i=1}^n$, $x_i \in \mathbb{R}^k$, $y_i \in \mathbb{R}^m$ (one-hot vectors) be our training data.

Softmax Classifier

Let $(x_i, y_i)_{i=1}^n$, $x_i \in \mathbb{R}^k$, $y_i \in \mathbb{R}^m$ (one-hot vectors) be our training data. Consider the function

$$f(x) = \left(\frac{e^{w_1^T x + b_1}}{\sum\limits_{i=1}^m e^{w_i^T x + b_i}}, \ldots, \frac{e^{w_m^T x + b_m}}{\sum\limits_{i=1}^m e^{w_i^T x + b_i}} \right)$$

Softmax Classifier

Let $(x_i, y_i)_{i=1}^n$, $x_i \in \mathbb{R}^k$, $y_i \in \mathbb{R}^m$ (one-hot vectors) be our training data. Consider the function

$$f(x) = \left(\frac{e^{w_1^T x + b_1}}{\sum_{i=1}^{m} e^{w_i^T x + b_i}}, \ldots, \frac{e^{w_m^T x + b_m}}{\sum_{i=1}^{m} e^{w_i^T x + b_i}} \right)$$

Our aim is to find parameters $(b_i, w_i)_{i=1}^m$ such that

$$f(x_i) \approx y_i, i = 1, \ldots, n.$$

Softmax Classifier

Let $(x_i, y_i)_{i=1}^{n}$, $x_i \in \mathbb{R}^k$, $y_i \in \mathbb{R}^m$ (one-hot vectors) be our training data. Consider the function

$$f(x) = \left(\frac{e^{w_1^T x + b_1}}{\sum_{i=1}^{m} e^{w_i^T x + b_i}}, \ldots, \frac{e^{w_m^T x + b_m}}{\sum_{i=1}^{m} e^{w_i^T x + b_i}} \right)$$

Our aim is to find parameters $(b_i, w_i)_{i=1}^{m}$ such that

$$f(x_i) \approx y_i, i = 1, \ldots, n.$$

We choose cross entropy distance as our loss function:

$$\frac{1}{n} \sum_{i=1}^{n} \left(-y_i^T \log f(x_i) \right).$$

Questions

1. Can you represent this model as a neural network?

Questions

1. Can you represent this model as a neural network?
2. Can we use the function sigmoid in this case?

Questions

1. Can you represent this model as a neural network?
2. Can we use the function sigmoid in this case?
3. What to do in the case of multi-label classification?