Deep Learning

Vazgen Mikayelyan

YSU, Krisp

October 10, 2019

Outline

Simple Moving Average

Definition 1

Simple moving average of the given data is the arithmetic mean of the previous k data.

Simple Moving Average

Definition 1

Simple moving average of the given data is the arithmetic mean of the previous k data.

If you have the data $x_1, x_2, \dots,$ then its simple moving average will be the following

$$\mu_n = \frac{x_{n-k+1} + \ldots + x_n}{k}, n = k, k+1, \ldots .$$

Cumulative Moving Average

Definition 2

Cumulative moving average of the given data is the arithmetic mean of the all previous data up to the current time.

Cumulative Moving Average

Definition 2

Cumulative moving average of the given data is the arithmetic mean of the all previous data up to the current time.

If you have the data $x_1, x_2, \ldots$, then its cumulative moving average will be the following

$$\mu_n = \frac{x_1 + x_2 + \ldots + x_n}{n}, n = 1, 2, \ldots.$$

Cumulative Moving Average

Definition 2

Cumulative moving average of the given data is the arithmetic mean of the all previous data up to the current time.

If you have the data $x_1, x_2, \ldots$, then its cumulative moving average will be the following

$$\mu_n = \frac{x_1 + x_2 + \ldots + x_n}{n}, n = 1, 2, \ldots .$$

Note that

$$\mu_n = \frac{(x_1 + x_2 + \ldots + x_{n-1}) + x_n}{n} = \frac{(n-1)\,\mu_{n-1} + x_n}{n}$$

$$= \left(1 - \frac{1}{n}\right)\mu_{n-1} + \frac{1}{n} x_n.$$

Exponential Moving Average

If you have the data $x_1, x_2, \ldots$, then its exponential moving average will be the following

$$\mu_1 = x_1,$$

$$\mu_n = \alpha \mu_{n-1} + (1 - \alpha)\, x_n, \;\; n \geq 2$$

Exponential Moving Average

If you have the data $x_1, x_2, \ldots$, then its exponential moving average will be the following

$$\mu_1 = x_1,$$

$$\mu_n = \alpha\mu_{n-1} + (1-\alpha)\, x_n, \;\; n \geq 2$$

Note that

$$\mu_n = \alpha\mu_{n-1} + (1-\alpha)\, x_n = \alpha\left(\alpha\mu_{n-2} + (1-\alpha)\, x_{n-1}\right) + (1-\alpha)\, x_n$$

Exponential Moving Average

If you have the data $x_1, x_2, \ldots$, then its exponential moving average will be the following

$$\mu_1 = x_1,$$

$$\mu_n = \alpha\mu_{n-1} + (1-\alpha)\,x_n,\;\; n \geq 2$$

Note that

$$\mu_n = \alpha\mu_{n-1} + (1-\alpha)\,x_n = \alpha\,(\alpha\mu_{n-2} + (1-\alpha)\,x_{n-1}) + (1-\alpha)\,x_n$$

$$= \alpha^2\mu_{n-2} + (1-\alpha)\,\alpha x_{n-1} + (1-\alpha)\,x_n$$

Exponential Moving Average

If you have the data $x_1, x_2, \ldots$, then its exponential moving average will be the following

$$\mu_1 = x_1,$$

$$\mu_n = \alpha\mu_{n-1} + (1-\alpha)\,x_n,\ \ n \geq 2$$

Note that

$$\mu_n = \alpha\mu_{n-1} + (1-\alpha)\,x_n = \alpha\left(\alpha\mu_{n-2} + (1-\alpha)\,x_{n-1}\right) + (1-\alpha)\,x_n$$

$$= \alpha^2\mu_{n-2} + (1-\alpha)\,\alpha x_{n-1} + (1-\alpha)\,x_n$$

$$= \alpha^3\mu_{n-3} + (1-\alpha)\,\alpha^2 x_{n-2} + (1-\alpha)\,\alpha x_{n-1} + (1-\alpha)\,x_n$$

Exponential Moving Average

If you have the data $x_1, x_2, \ldots$, then its exponential moving average will be the following

$$\mu_1 = x_1,$$

$$\mu_n = \alpha \mu_{n-1} + (1-\alpha)\, x_n, \;\; n \geq 2$$

Note that

$$\mu_n = \alpha \mu_{n-1} + (1-\alpha)\, x_n = \alpha \left(\alpha \mu_{n-2} + (1-\alpha)\, x_{n-1}\right) + (1-\alpha)\, x_n$$

$$= \alpha^2 \mu_{n-2} + (1-\alpha)\, \alpha x_{n-1} + (1-\alpha)\, x_n$$

$$= \alpha^3 \mu_{n-3} + (1-\alpha)\, \alpha^2 x_{n-2} + (1-\alpha)\, \alpha x_{n-1} + (1-\alpha)\, x_n$$

$$= \alpha^{n-1} \mu_1 + (1-\alpha)\, \alpha^{n-2} x_2 + \ldots + (1-\alpha)\, \alpha x_{n-1} + (1-\alpha)\, x_n.$$

Exponential Moving Average

If you have the data $x_1, x_2, \ldots$, then its exponential moving average will be the following

$$\mu_1 = x_1,$$

$$\mu_n = \alpha\mu_{n-1} + (1-\alpha)\,x_n,\;\; n \geq 2$$

Note that

$$\mu_n = \alpha\mu_{n-1} + (1-\alpha)\,x_n = \alpha\left(\alpha\mu_{n-2} + (1-\alpha)\,x_{n-1}\right) + (1-\alpha)\,x_n$$

$$= \alpha^2\mu_{n-2} + (1-\alpha)\,\alpha x_{n-1} + (1-\alpha)\,x_n$$

$$= \alpha^3\mu_{n-3} + (1-\alpha)\,\alpha^2 x_{n-2} + (1-\alpha)\,\alpha x_{n-1} + (1-\alpha)\,x_n$$

$$= \alpha^{n-1}\mu_1 + (1-\alpha)\,\alpha^{n-2} x_2 + \ldots + (1-\alpha)\,\alpha x_{n-1} + (1-\alpha)\,x_n.$$

It easy to see that sum of the coefficients is equal to 1.

Outline

Batch Normalization

- Problem:
  - The distribution of each layer's input changes during training.

Batch Normalization

- Problem:
  - The distribution of each layer's input changes during training.
- Solution:
  - Fix the distribution of inputs into subnetwork.

Batch Normalization

- Problem:
  - The distribution of each layer's input changes during training.
- Solution:
  - Fix the distribution of inputs into subnetwork.
- Effects:
  - Improve accuracy.
  - Faster learning.
  - Availability of high learning rates.

Batch Normalization

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1\ldots m}\}$;
Parameters to be learned: γ, β

Output: $\{y_i = \mathrm{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma_{\mathcal{B}}^2 \leftarrow \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma \widehat{x}_i + \beta \equiv \mathrm{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

Questions

1. Can we do stochastic gradient descent in this case?

Questions

1. Can we do stochastic gradient descent in this case?
2. What to do during the test?

Questions

1. Can we do stochastic gradient descent in this case?
2. What to do during the test?
3. What about biases?

Outline

Gradient Descent with Momentum

Let $L(w)$ be a loss function that we want to minimize. The algorithm gradient descent with momentum is the following

$$v_0 = 0,$$

$$v_t = \beta v_{t-1} + (1-\beta)\nabla L(w_t),$$

$$w_{t+1} = w_t - \alpha v_t,$$

where α is the learning rate and $\beta \in [0,1)$ is the parameter of exponential moving average.

Gradient Descent with Momentum

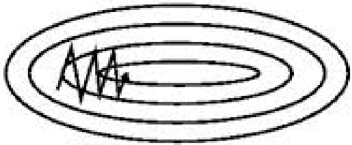

Image 2: SGD without momentum

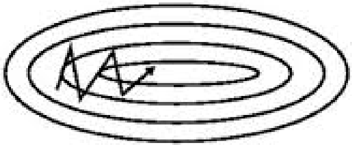

Image 3: SGD with momentum

RMSProp

Let $L(w)$ be a loss function that we want to minimize. The algorithm RMSprop is the following

$$v_0 = 0,$$
$$v_t = \beta v_{t-1} + (1-\beta)\left(\nabla L(w_t)\right)^2,$$
$$w_{t+1} = w_t - \alpha \frac{\nabla L(w_t)}{\sqrt{v_t} + \varepsilon},$$

where α is the learning rate and $\beta \in [0, 1)$ is the parameter of exponential moving average.

ADAM

Let $L(w)$ be a loss function that we want to minimize. The algorithm RMSprop is the following

$$m_0 = 0, v_0 = 0,$$

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) \nabla L(w_{t-1}),$$
$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t},$$

ADAM

Let $L(w)$ be a loss function that we want to minimize. The algorithm RMSprop is the following

$$m_0 = 0, v_0 = 0,$$

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) \nabla L(w_{t-1}),$$

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t},$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) (\nabla L(w_{t-1}))^2,$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t},$$

ADAM

Let $L(w)$ be a loss function that we want to minimize. The algorithm RMSprop is the following

$$m_0 = 0, v_0 = 0,$$

$$m_t = \beta_1 m_{t-1} + (1-\beta_1)\nabla L(w_{t-1}),$$
$$\hat{m}_t = \frac{m_t}{1-\beta_1^t},$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)\left(\nabla L(w_{t-1})\right)^2,$$
$$\hat{v}_t = \frac{v_t}{1-\beta_2^t},$$

$$w_t = w_{t-1} - \alpha\frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\varepsilon},$$

where α is the learning rate and $\beta_1, \beta_2 \in [0,1)$ are the parameters of exponential moving averages.

Outline

Data Augmentation

- Images:

Data Augmentation

- Images:
  - horizontal flips,

Data Augmentation

- Images:
  - horizontal flips,
  - crops,

Data Augmentation

- Images:
    - horizontal flips,
    - crops,
    - change contrast of colours,

Data Augmentation

- Images:
  - horizontal flips,
  - crops,
  - change contrast of colours,
  - add noise.
- Audios:

Data Augmentation

- Images:
  - horizontal flips,
  - crops,
  - change contrast of colours,
  - add noise.
- Audios:
  - change speed,

Data Augmentation

- Images:
  - horizontal flips,
  - crops,
  - change contrast of colours,
  - add noise.
- Audios:
  - change speed,
  - change pitch,

Data Augmentation

- Images:
  - horizontal flips,
  - crops,
  - change contrast of colours,
  - add noise.
- Audios:
  - change speed,
  - change pitch,
  - add noise.

Outline

What is convolution?

Definition 3

Convolution of the functions $f, g : \mathbb{R} \to \mathbb{R}$ is defined as the integral of the product of the two functions after one is reversed and shifted:

$$(f * g)(t) =: \int\limits_{-\infty}^{+\infty} f(x)\, g(t - x)\, dx.$$

What is convolution?

Definition 3

Convolution of the functions $f, g : \mathbb{R} \to \mathbb{R}$ is defined as the integral of the product of the two functions after one is reversed and shifted:

$$(f * g)(t) =: \int_{-\infty}^{+\infty} f(x)\, g(t - x)\, dx.$$

It easy to see that $f * g = g * f$.

What is convolution?

Definition 3

Convolution of the functions $f, g : \mathbb{R} \to \mathbb{R}$ is defined as the integral of the product of the two functions after one is reversed and shifted:

$$(f * g)(t) =: \int_{-\infty}^{+\infty} f(x)\, g(t-x)\, dx.$$

It easy to see that $f * g = g * f$.

Definition 4

Convolution of the sequences of real numbers $\{f_n\}_{n=-\infty}^{+\infty}, \{g_n\}_{n=-\infty}^{+\infty}$ is the following sequence:

$$z_n =: \sum_{k=-\infty}^{+\infty} f_k g_{n-k}.$$

2D Convolution

Definition 5

Convolution of the functions $f, g : \mathbb{R}^2 \to \mathbb{R}^2$ *is the following function:*

$$(f * g)(t, \tau) =: \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y)\, g(t - x, \tau - y)\, dx dy.$$

2D Convolution

Definition 5

Convolution of the functions $f, g : \mathbb{R}^2 \to \mathbb{R}^2$ *is the following function:*

$$(f * g)(t, \tau) =: \int\limits_{-\infty}^{+\infty} \int\limits_{-\infty}^{+\infty} f(x, y)\, g(t - x, \tau - y)\, dxdy.$$

It easy to see that $f * g = g * f$.

2D Convolution

Definition 5

Convolution of the functions $f, g : \mathbb{R}^2 \to \mathbb{R}^2$ *is the following function:*

$$(f * g)(t, \tau) =: \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y)\, g(t - x, \tau - y)\, dxdy.$$

It easy to see that $f * g = g * f$.

Definition 6

Let $f(x, y)$ *is an image and* $w(s, t)$ *is a kernel where* $s \in [a, b], t \in [c, d]$, $x, y, s, t, a, b, c, d \in \mathbb{Z}$. *The convolution between kernel* w *and image* f *is the following function*

$$(w * f)(x, y) = \sum_{s=a}^{b} \sum_{t=c}^{d} w(s, t)\, f(x - s, y - t)$$

2D Convolution