Video 15

Wednesday, October 21, 2020 7:23 PM

Gradient descent for linear regression

$y = \vec{w}^T \vec{x} + b$

scalar, weights, input, bias

output target is a scalar

$\left(\vec{x}^{(i)}, t^{(i)}\right)$

input (vector), target scalar

Loss function

(quantifies the error between the model's prediction & the target)

For one training example:

$\mathcal{L}(y, t) = \frac{1}{2}(y - t)^2$

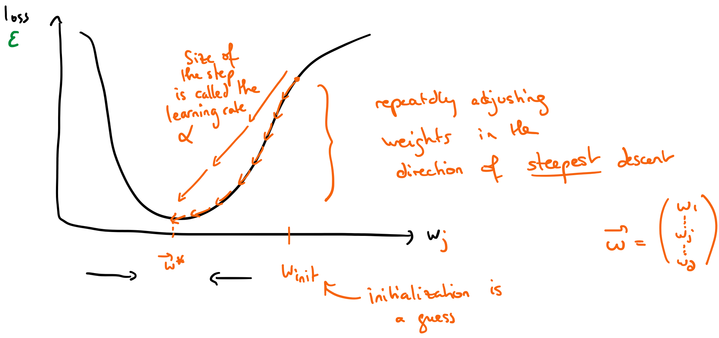
model prediction, target, for notation convenience

For entire dataset: N training examples

$$\mathcal{E}(\vec{w}, b) = \frac{1}{2N} \sum_{i=1}^{N} \left(y^{(i)} - t^{(i)}\right)^2$$

Gradient descent

- will allow us to minimize the loss function
- more broadly applicable (e.g., can be used w/ non-linear models)
- Grad descent is an iterative algorithm
  - ⇒ we are going to apply an update to the weights repeatedly
  - ⇒ until some criterion is met



$\vec{w} = \begin{pmatrix} w_1 \\ \vdots \\ w_j \\ \vdots \\ w_D \end{pmatrix}$

$\frac{\partial \mathcal{E}}{\partial w} < 0$ — If we increase w, we will decrease loss $\mathcal{E}$

$\frac{\partial \mathcal{E}}{\partial w} > 0$ — If we increase w, we will increase the loss $\mathcal{E}$

Update rule (direction of steepest descent):

$$w_j \leftarrow w_j - \alpha \frac{\partial \mathcal{E}}{\partial w_j}$$

If $\frac{\partial \mathcal{E}}{\partial w} > 0$, $-\frac{\partial \mathcal{E}}{\partial w} < 0$, we decrease w, we decrease the loss $\mathcal{E}$

If $\frac{\partial \mathcal{E}}{\partial w} < 0$, $-\frac{\partial \mathcal{E}}{\partial w} > 0$, we increase w, we decrease the loss $\mathcal{E}$

Learning rate

α: the larger it is, the faster $\vec{w}$ changes.

α is a hyperparameter which we need to tune using our training / validation splits

Typically we use small values of $\alpha \approx 0.01$ or 0.0001

Update rule in vectorized form:

(direction of fastest decrease of $\mathcal{E}$)

$$\vec{w} \leftarrow \vec{w} - \alpha \frac{\partial \mathcal{E}}{\partial \vec{w}}$$

$$\frac{\partial \mathcal{E}}{\partial \vec{w}} = \begin{pmatrix} \partial \mathcal{E}/\partial w_1 \\ \partial \mathcal{E}/\partial w_2 \\ \vdots \\ \partial \mathcal{E}/\partial w_D \end{pmatrix}$$

here model has D weights.

direction of fastest increase of $\mathcal{E}$