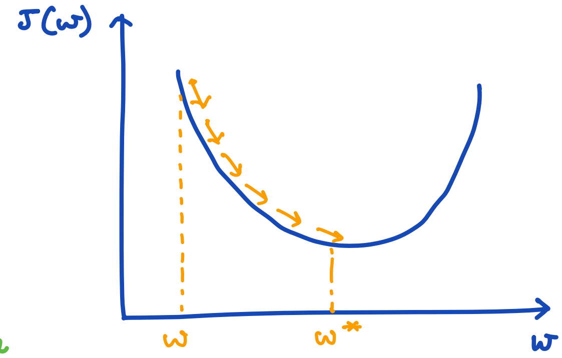
In the last lecture we talked about loss function to train a perceptron and how perceptron acts as a binary classifier. In this video our aim is to derive weight update rule that will help us to find optimal set of parameters w for the perceptron and we will use gradient descent update rule to update the parameters.

$$J(w) = \underbrace{-\sum_{i=1}^{m} y^{(i)} \log \hat{y}^{(i)} + (1-y^{(i)}) \log (1-\hat{y}^{(i)})}_{\text{Goal is to learn } w}$$

Loss function

J(w) is a convex fx^n

we start with any random w and we want to end up at some optimal $w^*$




This can be done by using Gradient update rule which repeatedly decreases our loss in the direction of reducing gradient.

$$w = [w_0 \quad w_1 \quad w_2 \quad \cdots\cdots \quad w_n]$$

n = no. of features

$$\frac{\partial J(w)}{\partial w} = \frac{\partial J}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial z}{\partial w}$$

(Chain Rule)

$\hat{y} = \sigma(z)$
$z = w^T \cdot x$

for 1 example later we will add $i$

$$= -\left(\frac{y}{\hat{y}} - \frac{1-y}{1-\hat{y}}\right)$$

$$\frac{\partial \hat{y}}{\partial z} \; ?$$

$$\hat{y} = \sigma(z) = \frac{1}{1+e^{-z}}$$

$$\frac{\partial \hat{y}}{\partial z} = \frac{e^{-z}}{(1+e^{-z})^2}$$

$$\frac{\partial \hat{y}}{\partial z} = \left(\frac{1}{1+e^{-z}}\right)\left(1 - \frac{1}{1+e^{-z}}\right)$$

$$\hat{y} = \sigma(z)$$

$$\hat{y}' = \sigma'(z)$$

$$\frac{\partial \hat{y}}{\partial z} = \sigma(z)\,(1-\sigma(z))$$

$$\hat{y}' = \sigma(z)\,(1-\sigma(z))$$

$$\hat{y}' = \hat{y}\,(1-\hat{y})$$

$$\frac{\partial z}{\partial w} \; ?$$

$$z = w^T x$$

$$z = w_0 + w_1 x_1 + \cdots\cdots w_i x_i + \cdots\cdots + w_n x_n$$

$$\frac{\partial z}{\partial w_i} = x_i$$

$$\frac{\partial J(w)}{\partial w_i} = \frac{\partial J}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z} \cdot \frac{\partial z}{\partial w_i}$$

$$= \left(-\frac{y}{\hat{y}} + \frac{1-y}{1-\hat{y}}\right)\left(\hat{y}\,(1-\hat{y})\right) x_i$$

$$= \frac{\left(-y + \cancel{y \cdot \hat{y}} + \hat{y} - \cancel{\hat{y} \cdot y}\right)}{\cancel{\hat{y}(1-\hat{y})}} \; \cancel{\left(\hat{y}(1-\hat{y})\right)} \; x_i$$

$$= (\hat{y} - y)\, x_i$$

$$\frac{\partial J(w)}{\partial w_i} = \left(\sigma(w^T \cdot x) - y\right) x_i$$

ith feature for given 1 example

$$w_j = w_j - \eta \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)}\right) x_j^{(i)}$$

update the jth gradient.