

Types of Learning

- Supervised
 - *Structured/labeled data*
 - Ex: picture \rightarrow label
 - Ex: picture \rightarrow digit number
- Unsupervised
 - *Unstructured data*
 - Dataset of pictures(not labeled)
- m denotes # of training examples

Binary Classification

- Classifies whether the input (is/has) or (isn't/doesn't have) a particular thing.
- Ex: Given an image, tell whether it has a cat in it or not 1(cat) vs 0(non-cat).
- In general, computers store images in 3 matrices(red, green, and blue) each of size width x height
- Images are stretched into a single column vector, usually starting with all the red pixels then all the green ones, then all the blue ones.
- n_x = the dimension of the input features
 - Ex: $64 \times 64 \times 3 = n_x = 12288$
- A single training example is denoted as:
 - (x, y) where $x \in \mathbb{R}^{n_x}$ and $y \in \{0, 1\}$
- m training examples: $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$
 - Other common notation:
 - * $m = m_{train}$
 - * $m_{test} = \#$ test examples
- X is used to denote a matrix of all the inputs
 - $X \in \mathbb{R}^{n_x \times m}$
 - $X.shape = (n_x, m)$
- Y is used to denote a row vector of all the labels
 - $Y = [y^{(1)}, y^{(2)}, \dots, y^{(m)}]$
 - $Y \in \mathbb{R}^{1 \times m}$
 - $Y.shape = (1, m)$

Logistic Regression

- Given x , we want $\hat{y} = P(y = 1 | x)$
- Parameters:
 - $x \in \mathbb{R}^{n_x}$
 - $b \in \mathbb{R}$
- Output:
 - $\sigma(w^T x + b)$

- We use the *sigmoid* function to make sure that the output is between 0 and 1, and centered on 0.5

Logistic Regression Cost Function

- Given $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(m)}, y^{(m)})\}$, we want $\hat{y}^{(i)} \approx y^{(i)}$
- Loss(error) function:
 - Computes the error for a single training example
 - $L(\hat{y}, y) = -(y \log \hat{y} + (1 - y) \log (1 - \hat{y}))$
 - If $y = 1$: $\mathcal{L}(\hat{y}, y) = -\log \hat{y} \leftarrow$ wants \hat{y} to be large
 - If $y = 0$: $\mathcal{L}(\hat{y}, y) = -\log (1 - \hat{y}) \leftarrow$ wants \hat{y} to be small
- Cost function:
 - Averages the loss of the entire training set
 - $J(w, b) = \frac{1}{m} \sum_{i=1}^m L(\hat{y}^{(i)}, y^{(i)})$

Gradient Descent

- $w := w - \alpha \frac{dJ(w, b)}{dw}$
- $b := b - \alpha \frac{dJ(w, b)}{db}$
- α is called the **Learning Rate**

Neural Network Representation

- Each layer is denoted by $a^{[n]}$
 - a is a column vector representing activations of neurons
 - The superscript n represents the layer that we are looking at
- The input layer is denoted $a^{[0]}$

Activation Functions

- Activation functions are denoted $g^{[n]}$
 - g represents the function
 - n represents the layer
- Previously, we used Sigmoid(σ)
 - Output is always between 0 and 1
 - Centered on 0.5
- Tanh is another alternative
 - Output is always between -1 and 1
 - Centered on 0
- Tanh is usually much better except for the output layer for binary activation
- ReLU = $\max(0, z)$
 - Rectified Linear Unit
- Leaky ReLU is another option
- No activation function is called a **linear Activation** function.
 - Its very uncommon to use a linear activation function.

- Sometimes used for the output layer when $y \in \mathbb{R}$

L-layer Deep Neural Network

- Forward Pass
 - $Z^{[\ell]} = W^{[\ell]} \cdot A^{[\ell-1]} + b^{[\ell]}$
 - $A^{[\ell]} = g^{[\ell]}(Z^{[\ell]})$
- Backward Pass
 - $dZ^{[\ell]} = dA^{[\ell]} * g^{[\ell]'}(Z^{[\ell]})$
 - $dW^{[\ell]} = \frac{1}{m} dZ^{[\ell]} \cdot A^{[\ell-1]T}$
 - $db^{[\ell]} = \frac{1}{m} \text{np.sum}(dZ^{[\ell]}, \text{axis}=1)$
 - $dA^{[\ell]} = W^{[\ell]T} \cdot dZ^{[\ell]}$

Train/Dev/Test Sets

- For small datasets, you have to allocate a larger % of examples for testing and validation.
 - Ex: 70/30 train/test split
- For larger datasets, a much larger % of examples are used for testing.
 - Ex: 99/1/1 train/dev/test split
- Always make sure the dev and test sets come from the same distribution.

Bias/Variance

- High **Bias** means *underfitting*.
 - Ex: 15% training error and 16% dev error
- High **Variance** means *overfitting*.
 - Ex: 1% training error and 11% dev error.
- High bias and high variance can both be present.
 - Ex: 15% training error and 30 % dev error.
- In between high bias and high variance is **just right**
 - Ex: 0.5% training error and 1% dev error.

Basic Recipe for Machine Learning

- Issues with high bias?
 - Increase the size of your network
 - Train longer
- High variance?
 - Try to get more data
 - Regularization
 - More appropriate architecture

Regularization

- **L2 Regularization**
 - *Sum of all the weights squared*

$$\|w\|_2^2 = \sum_{j=1}^{N_x} w_j^2$$