Stanford CS230 Notes

1 Introduction to Deep Learning

1.1 What is a Neural Network?

neuron, links.

1.2 Supervised Learning with Neural Networks

Supervised Learning

Examples： Standard NN, Convolutional NN, Recurrent NN

Structured Data： tabular data; Unstructured Data： audio/image/text

1.3 Why is Deep Learning taking off?

Amount of labeled data.

2 Basics of Neural Network Programming

2.1 Binary Classification

2.1.1 Binary Classification

image $\rightarrow$ 1 (cat) vs 0 (non cat)

2.1.2 Notation

m: number of examples

n_x: input size

n_y: output size

x: input, column vector

y: output, 0/1

X: input matrix, shape = (n_x, m)

Y: output matrix, shape = $(1, m)$

$x^{(i)}$: superscript (i) will denote the i^{th} example.

2.2 Logistic Regression

Given: $x \in \mathbb{R}^{n_x}$, $0 \leq \hat{y} \leq 1$

Parameters: $w \in \mathbb{R}^{n_x}$, $b \in \mathbb{R}$

Output:

$$z = w^T x + b \tag{1}$$

$$\hat{y} = \sigma(z) \tag{2}$$

$$\sigma(z) \approx \frac{1}{1 + e^{-z}} \tag{3}$$

$$z \approx \infty, \sigma(z) \approx \frac{1}{1 + 0} = 1 \tag{4}$$

$$z \approx -\infty, \sigma(z) \approx \frac{1}{1 + \infty} = 0 \tag{5}$$

Simplified Parameters: $x_0 = 1$, $x \in \mathbb{R}^{n_x+1}$

$$\theta_0 = b, \theta_1 ... \theta_{n_x} = w \tag{6}$$

$$\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ ... \\ \theta_{n_x} \end{bmatrix} \tag{7}$$

$$\hat{y} = \sigma(\theta^T x) \tag{8}$$

2.3 Logistic Regression cost function

Given $\{(x^{(1)}, y^{(1)}, ..., (x^{(m)}, y^{(m)}\}$, want $\hat{y}^{(i)} \approx y^{(i)}$.

Loss (error) function (mean square error/cross entropy):

$$\mathcal{L}(\hat{y}, y) = \frac{1}{2}(\hat{y} - y)^2 \tag{9}$$

$$\mathcal{L}(\hat{y}, y) = -(y \log \hat{y} + (1 - y) \log(1 - \hat{y})) \tag{10}$$

If y = 1: $\mathcal{L}(\hat{y}, y) = -\log \hat{y}$, want $\mathcal{L}$ small, want $\hat{y}$ large, want $\hat{y}$ equal to 1. If y = 0: $\mathcal{L}(\hat{y}, y) = -\log(1 - \hat{y})$, want $\mathcal{L}$ small, want $\hat{y}$ small, want $\hat{y}$ equal to 0.

Cost function:

$$J(w, b) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{L}(\hat{y}, y) \tag{11}$$

$$J(w, b) = -\frac{1}{m} \sum_{i=1}^{m} [y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log 1 - \hat{y}^{(i)}] \tag{12}$$

Given a random variable X with probability mass function $p_X(x)$, the self information of measuring X as outcome x is defined as:

$$I_X(x) = \log[p_X(x)] = \log(\frac{1}{p_X(x)}) \tag{13}$$

Shannon Entroy of X:

$$H(X) = \sum_x -p_X(x) \log p_X(x) \tag{14}$$
$$= \sum_x p_X(x) I_X(x) \tag{15}$$
$$= E[I_X(x)] \tag{16}$$

Cross Entropy of the the true distributions p and estimated distribution q:

$$H(p,q) = E_p[-\log q] = -\sum_{x \in \mathcal{X}} p(x) \log q(x) \tag{17}$$

2.4 Gradient Descent

Want to find w, b that minimize $J(w,b)$.

Repeat:

$$w := w - \alpha \frac{\mathrm{d}J(w,b)}{\mathrm{d}w} \tag{18}$$
$$b := b - \alpha \frac{\mathrm{d}J(w,b)}{\mathrm{d}b} \tag{19}$$

α: learning rate

2.5 Derivatives

$$f(a) = 3a, \frac{\mathrm{d}f(a)}{\mathrm{d}a} = 3 = \frac{\mathrm{d}}{\mathrm{d}a} f(a) \tag{20}$$
$$f(a) = a^2, \frac{\mathrm{d}}{\mathrm{d}a} f(a) = 2a \tag{21}$$
$$f(a) = a^3, \frac{\mathrm{d}}{\mathrm{d}a} f(a) = 3a^2 \tag{22}$$
$$f(a) = \log_e a = \ln a, \frac{\mathrm{d}}{\mathrm{d}a} f(a) = \frac{1}{a} \tag{23}$$
$$f(x) = \log_a x, \frac{\mathrm{d}}{\mathrm{d}x} f(x) = \frac{1}{x \ln a} \tag{24}$$
$$f(x) = a^x, \frac{\mathrm{d}}{\mathrm{d}x} f(x) = a^x \ln a \tag{25}$$
$$\log_a b = \frac{\log_c b}{\log_c a} = \frac{\ln a}{\ln b} \tag{26}$$