30-Aug-2020: Welcome Video, and What is Machine Learning

- Machine learning: algorithms; supervised, unsupervised, reinforcement, recommender. In this course, also will learn best practices.

31-Aug-2020: Supervised Learning, and Unsupervised Learning

- Supervised learning: right answers are given
- Regression: predicts continuous variable output; Classification: predicts discrete values
- Classification can have $1, \dots, N, \dots, \infty$ attributes. E.g. benignness/malignancy based on age, or age and tumor size, etc.
- Unsupervised learning a.k.a. clustering: Right answers aren't given. For example, news that links to different sources for the same topic.
- Cocktail party algorithm: separates two voices in a conversation, with two microphone recordings. Singular value decomposition is key to this algorithm.
- When learning machine learning, use Octave

1-Sep-2020: Model Representation, and Cost Function

- Training set notation: m is number of training examples, x are input examples, and y are the output variables. Together, (x, y) form a training example. Also denoted $(x^{(i)}, y^{(i)})$.
- In a linear regression, $h_\theta(x) = \theta_0 + \theta_1 x \equiv h(x)$.
- Cost function is

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

- Want to minimize J w.r.t. θ_0 and θ_1.

4-Sep-2020: Cost Function, Intuition I&II; Gradient Descent

- Intuition I; Let $\theta_0 = 0$, then $\min_{\theta_1} J(\theta_1)$ is what we want
- Ex: $h_\theta(x) = \theta_1 x$ and let $(x, y) = \{(1,1), (2,2), (3,3)\}$.

$$\begin{aligned} J(\theta_1) &= \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 \\ &\rightarrow \text{ If } \theta_1 = 0,\ h_\theta(x) \equiv 0 \\ J(0) &= \frac{1}{2 \times 3} (1 + 4 + 9) \\ &= \frac{14}{6} \end{aligned}$$

- $J(\theta_1)$ is parabolic
- We want $\min_\theta J(\theta)$; here, $\theta_1 = 1$ satisfies this criterion
- Intuition II; Let θ_0, θ_1 be free in $J(\theta_0, \theta_1)$ and $h_\theta(x)$.
- $J(\theta_0, \theta_1)$ is a parabloid
- Gradient Descent; Use gradient descent to find (θ_0, θ_1) that minimizes $J(\theta_0, \theta_1)$.
- Differing starting guesses can give different local minima.
- Gradient descent algorithm:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1) \quad \text{for } j = 1, 2$$

- Simultaneously update θ_0, θ_1, α is called the learning rate.
- Ex: $\theta_0 = 1, \theta_2 = 2$ and $\theta_j := \theta_j + \sqrt{\theta_0 \theta_1}$.

$$\begin{aligned} \theta_0 &:= \theta_0 + \sqrt{\theta_0 \theta_1} \\ &= 1 + \sqrt{1 \times 2} \\ &= 1 + \sqrt{2} \\ \theta_1 &= \theta_2 + \sqrt{\theta_0 \theta_1} \\ &= 2 + \sqrt{1 \times 2} \quad \text{note here that we used the old value of } \theta_0 \\ &= 2 + \sqrt{2} \end{aligned}$$

5-Sep-2020: Gradient Descent Intuition, Gradient Descent for Linear Regression

- Gradient Descent Intuition: For simplicity, assume $\theta_0 = 0$
- One variable: $\theta_1 := \theta_1 - \alpha \frac{d}{d\theta_1} J(\theta_1)$; Newton-Raphson
- If α is too small, convergence may be very slow. If too large, it may miss the minimum.
- If θ_1 is already at a local minimum, g.d. leaves θ_1 unchanged since the derivative is zero.
- Gradient Descent for Linear Regression: We need derivatives

$$\frac{\partial}{\partial\theta_0} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} \left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right)$$
$$\frac{\partial}{\partial\theta_1} J(\theta_0, \theta_1) = \frac{1}{m} \sum_{i=1}^{m} \left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right) \times x^{(i)}$$

- So, gradient descent finds the new θ variables as

$$\theta_0 := \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right)$$
$$\theta_1 := \theta_1 - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(\theta_0 + \theta_1 x^{(i)} - y^{(i)}\right) \times x^{(i)}$$

- This is called "batch gradient descent"; batch implies looking at all the training examples. This is represented by the $\sum_{i=1}^{m}$.
- Quiz Linear Regression with One Variable: 2) $m = \Delta y / \Delta x = (1 - 0.5)/(2 - 1) = 0.5 \implies y = 0.5x + b$; y-intercept is clearly zero since (0,) is a data point.
- 3) $h_\theta(x)$; $\theta_0 = -1$, $\theta_1 = 2$; $h_\theta(6) = -1 + 2 \times 6 = 11$