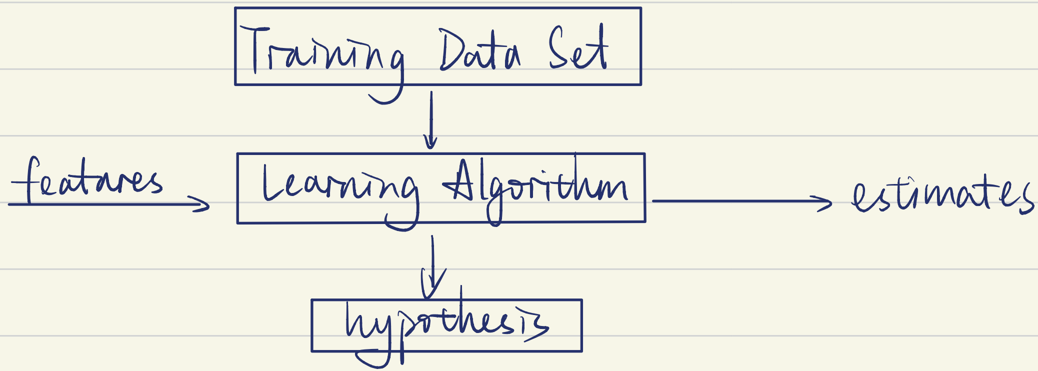
Lecture 2 Linear Regression and Gradient Descent

1.



hypothesis in linear regression represents as: $h(x) = \theta_0 + \theta_1 x_1 + \cdots$

equals: $h(x) = \sum_{j=0}^{n} \theta_j x_j$, where $x_0 = 0$

Notations:

m: # training example $\qquad$ n: # features

$x^{(i)}$: the i^{th} training example

2. Choose θ st. $h(x) \approx y$ for training examples

equals: $\underset{\theta}{\text{minimize}} \frac{1}{2} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 = J(\theta) \leftarrow$ cost function

use gradient descent:

① start with some θ (say $\theta = \vec{0}$)

② keep changing θ to reduce $J(\theta)$

$$\theta_j := \theta_j - \boxed{\alpha} \cdot \frac{\partial}{\partial \theta_j} J(\theta), \quad j = 0, 1, 2 \cdots$$

↳ learning rate

Batch Gradient Descent (use the whole data set)

Disadvantage: Computational Expensive

<proof>: when there's only one example:

$$\frac{\partial}{\partial \theta_j} \frac{1}{2} (h_\theta(x) - y)^2 = \frac{1}{2} \cdot 2 (h_\theta(x) - y) \cdot \frac{\partial}{\partial \theta_j} (h_\theta(x) - y)$$

$$= (h_\theta(x) - y) \cdot \frac{\partial}{\partial \theta_j} (h_\theta(x) - y) = (h_\theta(x) - y) \cdot x_j$$

⇒ Repeat until convergence:

$$\theta := \theta + \alpha \sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)})) x^{(i)}$$

3. Stochastic Gradient Descent (more computational efficient)

Repeat {
  For $j = 1$ to m {
    $\theta := \theta - \alpha(h_\theta(x^{(i)}) - y^{(i)})x^{(i)}$
  }
}

a slightly noisier and more random path
will never quite converge, oscillating around the minimum

4. Normal Equation

Notation: $\nabla_A f(A) = \begin{bmatrix} \frac{\partial f}{\partial A_{11}} & \cdots & \frac{\partial f}{\partial A_{1d}} \\ \vdots & \ddots & \\ \frac{\partial f}{\partial A_{n1}} & \cdots & \frac{\partial f}{\partial A_{nd}} \end{bmatrix}$ for a $f: \mathbb{R}^{n \times d} \to \mathbb{R}$

Derivations: If A is a square matrix ($A \in \mathbb{R}^{n\times n}$)

$\operatorname{tr} A$ = sum of diagnosed entries $= \sum_i^n A_{ii}$

$\operatorname{tr} A = \operatorname{tr} A^T$ | $\operatorname{tr} AB = \operatorname{tr} BA$

If $f(A) = \operatorname{tr} AB$ | $\operatorname{tr} ABC = \operatorname{tr} CAB$

$\Rightarrow \nabla_A f(A) = B^T$ | $\nabla_A \operatorname{tr} AA^TC = CA + C^TA$ $\quad (\frac{\partial}{\partial a} a^2 c = 2ac)$

For Least squares:

$$X_{m\times d} = \begin{bmatrix} -(x^{(1)})^T- \\ \vdots \\ -(x^{(m)})^T- \end{bmatrix} \qquad \vec{y} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

$$\therefore h_\theta(X) - \vec{y} = X\theta - \vec{y} = \begin{bmatrix} (x^{(1)})^T\theta \\ \vdots \\ (x^{(m)})^T\theta \end{bmatrix} - \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(m)} \end{bmatrix}$$

$$\therefore J(\theta) = \frac{1}{2}(h_\theta(X) - \vec{y})^2 = \frac{1}{2}(X\theta - \vec{y})^T(X\theta - \vec{y}) \qquad (z^Tz = \sum_i z^2)$$

<proof> Normal Equation:

$$\nabla_\theta J(\theta) = \nabla_\theta \frac{1}{2}(X\theta - \vec{y})^T (X\theta - \vec{y})$$

$$= \frac{1}{2}\nabla_\theta \left((X\theta)^T X\theta - (X\theta)^T \vec{y} - \vec{y}^T (X\theta) + \vec{y}^T \vec{y}\right)$$

$$= \frac{1}{2}\nabla_\theta \left(\theta^T (X^T X)\theta - \vec{y}^T (X\theta) - \vec{y}(X\theta)\right)$$

$$= \frac{1}{2}\nabla_\theta \left(\theta^T (X^T X)\theta - 2(X^T \vec{y})^T \theta\right)$$

$$= \frac{1}{2}\left(2X^T X\theta - 2X^T \vec{y}\right)$$

$$= X^T X\theta - X^T \vec{y}$$