1

i). The conditional probability of Y_1 and Y_2 is

$$f_{Y_1|Y_2}(y_1|y_2) = \frac{1}{\sqrt{2\pi(1-\rho^2)}\,\sigma_1} e^{\left(\frac{-(y_1-a)^2}{2\sigma^2(1-\rho^2)}\right)}$$

where $a = \mu_1 + \rho \frac{\sigma_1}{\sigma_2}(y_2 - \mu_2)$.

The expected value of $Y_1 | Y_2$ is

$$E[Y_1|Y_2] = \mu_1 + \frac{\rho\sigma_1}{\sigma_2}(y_2 - \mu_2)$$

$$E[Y_1|Y_2] = \left[\mu_1 - \mu_2 \rho \frac{\sigma_1}{\sigma_2}\right] + \rho \frac{\sigma_1}{\sigma_2} y_2$$

This is a linear combination of Y_2, which mimics linear regression with one predictor. The variance is defined by $\sigma[Y_1|Y_2] = \sigma^2(1-\rho^2)$

1

a) Joint probability distribution of Y_1 and Y_2 is

$$f_{Y_1,Y_2}(y_1, y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot e^{\frac{-1}{2(1-\rho^2)}\left[\left(\frac{y_1-\mu_1}{\sigma_1}\right)^2 + \left(\frac{y_2-\mu_2}{\sigma_2}\right)^2 - 2\rho\left(\frac{y_1-\mu_1}{\sigma_1}\right)\left(\frac{y_2-\mu_2}{\sigma_2}\right)\right]}$$

b) ~~[illegible]~~

$$f_{Y_1|Y_2}(y_1, y_2) = \frac{f_{Y_1,Y_2}(y_1, y_2)}{f_{Y_2}(y_2)}$$

$f_{Y_2}(y_2)$ (Marginal distribution) is.

$$\int_{-\infty}^{\infty} f_{Y_1,Y_2}(y_1, y_2)\, dy_1 = \frac{1}{\sqrt{2\pi}\sigma_2} e^{-\frac{(y_2-\mu_2)^2}{2\sigma_2^2}}$$

Applying Bayes theorum we get.

$$f_{Y_1,Y_2}(y_1,y_2) = \frac{\dfrac{1}{\sqrt{2}\,\pi\,\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot e^{\frac{-1}{2(1-\rho^2)}\left[\left(\frac{y_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{y_1-\mu_1}{\sigma_1}\right)\left(\frac{\cancel{y_2}-\mu_2}{\sigma_2}\right) + \left(\frac{y_2-\mu_2}{\sigma_2}\right)^2\right]}}{\dfrac{1}{\sqrt{2\pi}\,\sigma_2}\, e^{-\frac{(y_2-\mu_2)^2}{2\sigma_2^2}}}$$

$$= \frac{1}{\sqrt{2\pi(1-\rho^2)}\,\sigma_1}\, e^{-\frac{\left(y_1-\left(\mu_1+\frac{\sigma_1}{\sigma_2}(y_2-\mu_2)\right)\right)^2}{2\sigma_1^2(1-\rho^2)}}$$

The conditional probability ~~is~~ of $Y_1 \mid Y_2$ is normally distributed having mean $\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y_2-\mu_2)$ and $\sigma^2 = (1-\rho^2)\sigma_1^2$

Q.
2
a) Quadratic loss is defined by.

$$L = \sum_{i=1}^{N} e_i^2.$$

Taking second derivative.

$$\frac{\partial L}{\partial e} = 2e \; ; \quad \frac{\partial^2 L}{\partial e^2} = 2$$

Since the $\frac{\partial^2 L}{\partial e^2} \geq 0$ this loss is convex.

b) L_1 norm $L = \sum_{i=1}^{N} |e_i|$

$$\frac{\partial L}{\partial e} = \text{sgn}(e) \; ; \quad \frac{\partial^2 L}{\partial e^2} = 0$$

Since $\frac{\partial^2 L}{\partial e^2} = 0$ this loss function is also convex.

c) $L = \sum_{i=1}^{N} \ell(e_i)$ where $\ell(e) = \begin{cases} \frac{1}{2}e^2 & \text{if } |e| \leq \delta \\ \delta|e| - \frac{1}{2}\delta^2 & \text{if } |e| > \delta \end{cases}$

$$\frac{\partial L}{\partial e} = \begin{cases} e & , \text{if } |e| \leq \delta \\ \delta \operatorname{sgn}(e) & , \text{if } |e| > \delta \end{cases}$$

$$\frac{\partial^2 L}{\partial e^2} \begin{cases} 1 & ; \text{if } |e| \leq \delta \\ 0 & \text{if } |e| > \delta \end{cases}$$

, both of which are greater than or equal to 0.

Therefore this function is convex as well.

3

a). $y_i = \theta_0 + \theta_1 x_i + e_i \quad , i = 1 \dots N.$

Our goal is to minimise the cost function

$$J(\theta_0 \dots n) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2.$$

where m is the number of datapoints

Our hypothesis $h(x) = \theta_0 x_0 + \theta_1 x_1 \dots \theta_n x_n$

$$h(x) = \theta^T X.$$

where $\theta = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}$ and $X = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$

The cost function converted into matrix notation, we get.

$$J(\theta) = 1/2m \left(X\theta - y \right)^T \left(X\theta - y \right)$$

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

$$= \frac{(X\theta)^T X\theta - (X\theta)^T y - y^T (X\theta) + y^T y}{2m}$$

$$= \frac{\theta^T X^T X\theta - 2(X\theta)^T y + y^T y}{2m}$$

Differentiating and setting it to 0

$$= \frac{\left(2X^T X\theta - 2X^T y \right)}{2m} = 0$$

$$2X^T X\theta - 2X^T y = 0$$

$$\theta = (X^T X)^{-1} X^T y$$

b) Gradient descent is used to find optimal θ's in a given linear equation of form

$$h(x) = \theta_n x_n + \dots + \theta_2 x_2 + \theta_1 x_1 + \theta_0 x_0$$

such that the cost function $J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} h_\theta(x^i - y^i)^2$ is minimised

In order to minimise the above cost function, we need to calculate the slope of the function and adjust θ ~~m~~ ~~area~~ such that we 'move down' the cost slope.

This is done by differentiating the cost function, which gives us

$$\frac{\partial J(\theta_0, \theta_1)}{\partial \theta_1} = \frac{1}{m} \sum_{i=0}^{m} (h_\theta(x^i) - y^i) \cdot x^i$$

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

The above slopes are multipled by a set learning rate α and subtracted by θ_0 and θ_1

This gives us

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

$$\theta_1 = \theta_1 - \frac{\alpha}{m} \sum_{i=1}^{m} \left(h_\theta(x^i) - y^i \right) . x^i$$

Both θ_0 and θ_1 are updated simultaneously until convergence.

c) Stochastic gradient descent is, for the most part, the same as batch gradient descent.

The major difference is in the way stochastic processes its iterations. Whereas batch gradient descent processes all datapoints at once, for each iteration, stochastic processes one row per iteration.

Formally its specified as follows.

while $|\text{new } \theta - \text{old } \theta| > \text{convergence}$:

{

for $i=0 : i<n ; i{+}{+}$.

{ $\theta_{J=0 \text{ to } m} = \theta_J - \alpha \left(h_\theta(x^i) - (y^i) \right) x_J^i$

}

}