1. LINEAR REGRESSION

$$MSE = E[(f(x) - y)^2]$$

let $f(x) = \theta$

$MSE = E[(\theta - y)^2]$

Using Law of iterative Expectations.

$E[(\theta - y)^2] = E[E[(\theta - y)^2 | x]]$

Find θ^* that minimizes conditional MSE for all x ($E[(\theta - y)^2 | x]$) $\Rightarrow$ It will also minimize MSE

- Take derivative with respect to θ and set it $= 0$.

$$\Rightarrow \frac{\partial E[(\theta - y)^2 | x]}{\partial \theta} = E\left[\frac{\partial (\theta - y)^2}{\partial \theta} \middle| x\right]$$

$$= E[2(\theta - y)(1) | x]$$

$$= 2E[(\theta - y)|x] = 0$$

$$\Rightarrow E[(\theta - y)|x] = 0$$

$\boxed{f(x) = \theta}$ substituting back.

$$\Rightarrow E[f(x)|x] - E[y|x] = 0$$

$$\Rightarrow E[f(x)|x] = E[y|x]$$

$f(x)$ is a function of x and expectation is for fixed values of x therefore $E[f(x)|x] = f^*(x)$

$$\boxed{f^*(x) = E(y|x)}$$

Gaussian Noise Regression Model.

a) $$p(t|X,w,\beta) = \prod_{n=1}^{N} N\left(t_n \,|\, w^T\phi(x_n), B_n^{-1}\right)$$

$$= \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi B_n^{-1}}} \exp\left\{ \frac{-\left(t_n - w^T\phi(x_n)\right)^2}{2B_n^{-1}} \right\}$$

The log-likelihood.

$$\log(p(t|X,w,\beta)) = \log\left[\prod_{n=1}^{N} \frac{\sqrt{\beta_n}}{\sqrt{2\pi}} \exp\left\{\frac{-\beta_n\left(t_n - w^T\phi(x_n)\right)^2}{2}\right\}\right]$$

$$= \sum_{n=1}^{N}\left[\frac{1}{2}\log\beta_n - \frac{1}{2}\log(2\pi) - \frac{\beta_n}{2}\left(t_n - w^T\phi(x_n)\right)^2\right]$$

$$= \frac{N\log\beta_n}{2} - \frac{N}{2}\log(2\pi) - \frac{\beta_n}{2}\sum_{n=1}^{N}\left(t_n - w^T\phi(x_n)\right)^2$$

$$\log(p(t|X,w,\beta)) = \frac{N}{2}\log\beta_n - \frac{N}{2}\log(2\pi) - \frac{\beta_n}{2}\sum_{n=1}^{N}\left(t_n - w^T\phi(x_n)\right)^2$$

b). Sum of squared error is maximum likelihood under a Gaussian Noise Model. Here it is B_n times Sum of Squared Error.

$\times\!\longleftarrow\!\times$

Weighted Linear Regression

a) $p(y|X, \beta, \sigma^2) = \prod_{i=1}^{N} \mathcal{N}(y_i \mid x_i^T\beta, \sigma_i^2)$

$$\log(p(y|X, \beta, \sigma^2)) = \log\left[\prod_{i=1}^{N}\left[\frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[\frac{(-y_i - x_i^T\beta)^2}{2\sigma_i^2}\right]\right]\right]$$

$$= \sum_{i=1}^{N}\left[-\log\sqrt{2\pi\sigma_i^2} - \frac{1}{2\sigma_i^2}(y_i - x_i^T\beta)^2\right]$$

$$= \sum_{i=1}^{N}\left[\frac{-1}{2}\log(2\pi) + \frac{1}{2}\log(\sigma_i^2) - \frac{1}{2\sigma_i^2}(y_i - x_i^T\beta)\right]$$

$$= \frac{-N}{2}\log(2\pi) + \frac{N}{2}\log\sigma_i^2 - \sum_{i=1}^{N}\frac{1}{2\sigma_i^2}(y_i - x_i^T\beta)^2$$

$$\boxed{\beta^* = \underset{\beta}{\operatorname{argmax}}\left[\sum_{i=1}^{N}\frac{1}{2\sigma_i^2}(y_i - x_i^T\beta)\right]}$$

b) Let $\omega_i = \frac{1}{\sigma_i^2}$ as they are inversely proportional.

Small error $\rightarrow$ large weight
large error $\rightarrow$ small weight.

$$\sum_{i=1}^{N} \frac{\omega_i}{2} \left(y_i - x_i^T \beta \right)^2$$

$$\nabla \log p(y|X, \beta, \sigma_i^2) = \sum_{i=1}^{N} \frac{\cancel{2}\omega_i}{\cancel{2}} \left(y_i - x_i^T \beta \right)(x_i^T)$$

$$0^T = \sum_{i=1}^{N} \omega_i x_i^T y_i - \beta^T \sum_{i=1}^{N} \omega_i x_i x_i^T$$. Taking derivative and setting it $= 0$.

$$(0^T)^T = (Y^T W^T X - \beta^T W^T X^T X)^T$$. Writing in vector form and taking Transpose on both sides.

$$0 = YWX^T - \beta WXX^T$$

$$\beta WXX^T = YWX^T$$

$$\boxed{\beta = (YWX^T)(WXX^T)^{-1}}$$

$X \longleftarrow X$

2. Regularization.

a) We cannot compute according to (3) because $m \gg N$ which makes $(X^T X)^{-1}$ <u>not</u> <u>invertible</u>.

b) $$J_R(\beta) = (X\beta - y)^T (X\beta - y) + \lambda \|\beta\|^2$$

$$= (X\beta - y)^T (X\beta - y) + \lambda I \beta^T \beta$$

Take partial derivative with respect to β and set it $= 0$

$$\nabla J_R(\beta) = 0$$

$$0 = \cancel{2} X^T (X\beta - y) + \cancel{2} \lambda \beta$$

$$0 = X^T X \beta - X^T y + \lambda I \beta$$

$$X^T y = X^T X \beta + \lambda I \beta$$

$$X^T y = \beta (X^T X + \lambda I)$$

$$\boxed{\beta^* = (X^T X + \lambda I)^{-1} X^T y}$$

3. Classification

Logistic regression

$$\log(\text{odds}) = -10 + 2 * x_1$$

a)

$$\log\left(\frac{P}{1-P}\right) = -10 + 2 * x_1$$

$$e^{\log(P/1-P)} = e^{-10+2x_1}$$

$$\frac{P}{1-P} = e^{-10+2(6)}$$

$$P = e^2(1-P)$$

$$P = e^2 - e^2 P$$

$$P + e^2 P = e^2$$

$$P(1+e^2) = e^2$$

$$P = \frac{e^2}{1+e^2}$$

$\boxed{P = 0.88}$ is the probability of a cell being unaffected with diameter = 6.

b)

$$\log\left(\frac{0.9}{1-0.9}\right) = -10 + 2x_1$$

$$\frac{\log\left[\frac{0.9}{1-0.9}\right) + 10}{2} = x_1$$

$$\boxed{x_1 = 6.098}$$

Softmax for Multi-Class Classification

a). Activation Functions.

$$\text{Type A} = 2x_1 + 5x_2 + 5$$

$$\text{Type B} = 5x_1 + 10x_2 + 1.5$$

$$\text{Type C} = 5x_1 + 2x_2 + 1$$

Class probabilities.

$$p(C_A|x) = \frac{\exp(2x_1 + 5x_2 + 5)}{\exp[(2x_1+5x_2+5) + (5x_1 + 10x_2 + 1.5) + (5x_1 + 2x_2 + 1)]}$$

$$p(C_B|x) = \frac{\exp(5x_1 + 10x_2 + 1.5)}{\exp[(2x_1+5x_2+5) + (5x_1 + 10x_2 + 1.5) + (5x_1 + 2x_2 + 1)]}$$

$$p(C_c|x) = \frac{\exp(5x_1 + 2x_2 + 1)}{\exp[(2x_1+5x_2+5) + (5x_1 + 10x_2 + 1.5) + (5x_1 + 2x_2 + 1)]}$$

b) $x_1 = 10,\ x_2 = 2$ => plugging values in above equations:

$$p(C_A|x_1, x_2) = 1.4 \times 10^{-16}$$

$$p(C_B|x_1, x_2) = 0.999$$

$$p(C_C|x_1, x_2) = 6.825 \times 10^{-8}$$

predicted type is Type B as the class probability of Type B is highest that is $\boxed{0.999}$.