3.1

1.

$$Aw = b$$

$$w = A^{-1} b$$

$$A^{-1} = \frac{1}{|A|} \operatorname{adj} A$$

$$= \frac{1}{11}\begin{bmatrix} 3 & -1 \\ -1 & 4 \end{bmatrix}$$

$$w = \frac{1}{11}\begin{bmatrix} 3 & -1 \\ -1 & 4 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

$$w = \begin{bmatrix} 1/11 \\ 7/11 \end{bmatrix}$$

2.

$$w = \frac{1}{11}\begin{bmatrix} 3 & -1 \\ -1 & 4 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \end{bmatrix} \Rightarrow \begin{bmatrix} 3/11 & -1/11 \\ -1/11 & 4/11 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

3.

To verify: $Aw = b$

$$Aw = \begin{bmatrix} 4 & 1 \\ 1 & 3 \end{bmatrix}\begin{bmatrix} 1/11 \\ 7/11 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

which is b

Hence, verified.

3.2

1. $f(x) = x^2 + 3x + 1$

$$\frac{df}{dx} = 2x + 3$$

2. $g(x_1, x_2) = x_1^2 + 2x_1x_2 + 3x_2^2$

$$\frac{\partial g}{\partial x_1} = 2x_1 + 2x_2$$

$$\frac{\partial g}{\partial x_2} = 2x_1 + 6x_2$$

3.

$$\frac{dh}{dx} = \begin{bmatrix} 2x \\ 3 \end{bmatrix}$$

4.

$$\nabla g = \begin{bmatrix} 2x_1 + 2x_2 \\ 2x_1 + 6x_2 \end{bmatrix}$$

5.

$$g(x_1, x_2) = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} a & b \\ b & c \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= \begin{bmatrix} ax_1 + bx_2 & bx_1 + cx_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= \begin{bmatrix} ax_1^2 + bx_1x_2 + bx_1x_2 + cx_2^2 \end{bmatrix}$$

$$x_1^2 + 2x_1x_2 + 3x_2^2 = \begin{bmatrix} ax_1^2 + 2bx_1x_2 + cx_2^2 \end{bmatrix}$$

on comparing, $a = 1$, $b = 1$, $c = 3$.

$$\therefore A = \begin{bmatrix} 1 & 1 \\ 1 & 3 \end{bmatrix}$$

6.

$$\frac{\partial(g(x))}{\partial x} = \begin{bmatrix} 2x_1 + 2x_2 \\ 2x_1 + 6x_2 \end{bmatrix} \qquad g(x) = \begin{bmatrix} x_1^2 + 2x_1x_2 + 3x_2^2 \end{bmatrix}$$

7. (a) Let A be $\begin{bmatrix} a \\ b \end{bmatrix}$

$$A^T x = \begin{bmatrix} a & b \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} ax_1 + bx_2 \end{bmatrix}$$

$$\frac{\partial}{\partial x} \begin{bmatrix} ax_1 + bx_2 \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix} \Rightarrow A$$

$\therefore$ LHS = RHS

Hence, proved.

(b) $x^T x = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1^2 + x_2^2 \end{bmatrix}$.

$$\frac{\partial}{\partial x} [x_1^2 + x_2^2] = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix} = 2 \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= 2x$$

Hence, proved.

(c) $x^T A x = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$$= \begin{bmatrix} ax_1 + cx_2 & bx_1 + dx_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$= \begin{bmatrix} ax_1^2 + cx_1x_2 + bx_1x_2 + dx_2^2 \end{bmatrix}$$

$$\frac{\partial}{\partial x} [x^T A x] = \begin{bmatrix} 2ax_1 + (c+b)x_2 \\ 2dx_2 + (c+b)x_1 \end{bmatrix}$$

$$= \begin{bmatrix} 2a & b+c \\ b+c & 2d \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

$$A + A^T$$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} a & c \\ b & d \end{bmatrix}$$

$$= \underline{(A + A^T)\, x}$$

4.1

1.

$$P(\text{disease} \mid \text{positive}) = \frac{P(\text{positive} \mid \text{disease}) \times P(\text{disease})}{P(\text{positive} \mid \text{no disease}) \times P(\text{no disease}) + P(\text{positive} \mid \text{disease}) \times P(\text{disease})}$$

$$= \frac{0.01 \times 0.99}{0.01 \times 0.99 + 0.99 \times 0.05} = \frac{1}{6} = 0.167$$

4.2

1. Joint likelihood function

$$L(\mu, \sigma^2) = p(x_1, x_2, \dots, x_n \mid \mu, \sigma^2)$$

$$\cancel{L(\theta)} =$$

$$L(\mu, \sigma^2) = \prod_{i=1}^{N} P(x_i \mid \mu, \sigma^2)$$

and as $x_i \sim N(\mu, \sigma^2)$

$$p(x_i \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

$$L(\mu, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

2.

$$l(\mu, \sigma^2) = \log(L(\mu, \sigma^2)) =$$

$$\log\left(\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^N\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2$$

$$l(\mu, \sigma^2) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2$$

3.

$$\frac{\partial l}{\partial \mu} = \frac{+\cancel{2}1}{\cancel{2}\sigma^2}\sum_{i=1}^{N}(x_i-\mu) = 0 \Rightarrow \sum_{i=1}^{N}(x_i-\mu) = 0$$

Therefore, to maximize log-likehood $\underline{\mu = \frac{1}{N}\sum_{i=1}^{N} x_i}$

4.

$$\frac{\partial l}{\partial \sigma^2} = \frac{-N}{2} \times \frac{2\pi}{2\pi\sigma^2} + \frac{\cancel{1}}{\cancel{2}}\sum_{i=1}^{N}(x_i-\mu)^2\left(\frac{1}{(\sigma^2)^2}\right)$$

$$= \frac{-N}{2\sigma^2} + \frac{1}{2}\frac{1}{\sigma^4}\sum_{i=1}^{N}(x_i-\mu)^2$$

$$0 = \frac{1}{2\sigma^2}\left(-N + \frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i-\mu)^2\right)$$

$$\therefore \quad \sigma^2 = \underline{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}$$

5. $\hat{\mu}_{MLE} = \frac{1}{N} \sum_{i=1}^{N} x_i$

$\hat{\sigma}^2_{MLE} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$

5>

1. 10

2. Observing the first 2 pairs you find that they satisfy the relation $y = 2x$ and as the third pair satisfies this relation as well it feels like even if input is 5 the output would be 10.

3. The cubic fn. could be

$$f(x) = 2x + (x-2)(x-4)(x-9)$$

Therefore it does not necessarily satisfy $f(5) = 10$.

4. Assuming that it is a linear relation and hence it is satisfied by every x is what led us to arrive at the prediction of 10.

5. For a single input variable x, the hypothesis function is

$$\boxed{h(x) = \beta_0 + \beta_1 x}$$

where $h(x)$ is the predicted value of the dependent variable
β_0 is the intercept (value of y when $x = 0$)
β_1 is the slope

6. The multiple linear regression model takes the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_d X_d + \epsilon.$$

where X_j represents the jth predictor
β_j quantifies the association between the variable and response.

9. $\hat{y} = wx + b$

Let y_i be the actual observed value

$\hat{y}_i$ is the predicted value from the line for that x_i

Total squared error loss $\Rightarrow \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

$$RSS = \sum_{i=1}^{n} (y_i - wx_i - b)^2$$

10. $$\frac{\partial RSS}{\partial w} = -2 \sum_{i=1}^{n} (x_i (y_i - wx_i - b)) = 0.$$

$$\sum_{i=1}^{n} x_i (y_i - wx_i - b) = 0$$

$w\bar{x},$ $\quad = \frac{1}{\ } \sum y_i$

$$\sum_{i=1}^{n} \left(x_i y_i - w x_i^2 - b x_i \right) = 0 \Rightarrow (2)$$

$$\sum_{i=1}^{n} x_i y_i - w \sum_{i=1}^{n} x_i^2 - b \sum_{i=1}^{n} x_i = 0 . \Rightarrow (2) .$$

$$\frac{\partial (RSS)}{\partial b} = -2 \sum_{i=1}^{n} (y_i - w x_i - b) = 0 .$$

$$\begin{bmatrix} n\bar{x} = \sum x_i \\ n\bar{y} = \sum y_i \end{bmatrix}$$

$$\sum_{i=1}^{n} y_i - w \sum_{i=1}^{n} x_i - nb = 0$$

$$\boxed{\bar{y} - w\bar{x} = b} - (1)$$

$$\sum_{i=1}^{n} x_i y_i - w \sum_{i=1}^{n} x_i^2 - (\bar{y} - w\bar{x}) \sum_{i=1}^{n} x_i = 0 .$$

$$\sum_{i=1}^{n} x_i y_i - \bar{y} \sum_{i=1}^{n} x_i - w \sum_{i=1}^{n} x_i^2 + w \bar{x} \sum_{i=1}^{n} x_i$$

$$\left(n\bar{x} = \sum_{i=1}^{n} x_i \right.$$

$$\frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = w$$

$$\therefore w = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$b = \bar{y} - w\bar{x}$$

11. $$\underset{n\times 1}{\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}} = \underset{n\times d+1}{\begin{bmatrix} 1 & x_{11} & \cdots & x_{1d} \\ 1 & x_{21} & \cdots & x_{2d} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nd} \end{bmatrix}} \underset{d+1\times 1}{\begin{bmatrix} b \\ w_1 \\ \vdots \\ w_d \end{bmatrix}}$$

$$\boxed{\hat{y} = Xw}$$

12. Squared error loss $= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

$$= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\underline{L(w) = \| y - Xw \|^2} \qquad (\hat{y} = Xw)$$

13. $L(w) = \|y - Xw\|^2 \Rightarrow (y - Xw)^T (y - Xw)$

To minimise, take derivative / gradient & equate to 0.

$\Rightarrow$ ~~from derived property,~~

$$\Rightarrow y^T y - (Xw)^T y - y^T X w + (Xw)^T X w$$

$$\Rightarrow y^T y - \underbrace{y^T X w - w^T X^T y}_{\text{equal}} + w^T X^T X w$$

$$L(w) = y^T y - 2 y^T X w + w^T X^T X w$$

$$\cancel{\frac{\partial L(w)}{\partial w}} \Rightarrow \quad \cancel{0} \quad -2 (X^T y)^T w + w^T (X^T X) w$$

$y^T y$

$\left[\text{using } \frac{\partial (A^T w)}{\partial w} = A \text{ and } \frac{\partial (w^T A w)}{\partial w} = (A + A^T) w \right]$

$$\therefore \frac{\partial L(w)}{\partial w} = 0 - 2 X^T y + (X^T X + X^T X) w$$

$$\Rightarrow 2 X^T X w - 2 X^T y = 0$$

$$2 X^T (Xw - y) = 0.$$

$$X^T (Xw - y) = 0$$

$$\boxed{X^T X w = X^T y}$$

1 Assignment 2

1. (a)

$$P(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2)}}$$

$$Y = \begin{cases} 1 & \text{if } p > 0.5 \\ 0 & \text{if } p \leq 0.5 \end{cases}$$

(b) treating outcomes as bernoulli trials,

$$P(y^{(i)} | p^{(i)}) = (p^{(i)})^{y^{(i)}} (1-p^{(i)})^{1-y^{(i)}}$$

Likelihood fn:

$$L(\beta_0, \beta_1, \beta_2) = \prod_{i=1}^{n} (p^{(i)})^{y^{(i)}} (1-p^{(i)})^{1-y^{(i)}}$$

$$\Rightarrow \prod_{i=1}^{n} \left(\frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2)}}\right)^{y^{(i)}} \left(\frac{e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2)}}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2)}}\right)^{1-y^{(i)}}$$

Log-likelihood fn:

$$l(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^{n} \left[y^{(i)} \log(p^{(i)}) + (1-y^{(i)}) \log(1-p^{(i)}) \right]$$

2. (a) $P(Y) = \dfrac{1}{1+e^{-(-6+0.05(40)+3.5)}} = \underline{0.38}$

(b) $\cancel{0.5} = \dfrac{1}{\cancel{1+e^{-(}}}$ $\quad \log\left(\dfrac{p(x)}{1-p(x)}\right) = -6 + 0.05(x) + 3.5$

$-2.5 + 0.05(x_1) = 0$

$\underline{x_1 = 50}$ ∴ student should study for 50 hrs

3. Given $x = 4$,

using likelihood fns

(i) Dividend = Yes

$$p(x_i \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

$$p(4 \mid \mu_{(1)}, \sigma^2) = \frac{1}{\sqrt{2\pi(36)}} \exp\left(-\frac{(4-10)^2}{72}\right)$$

$$= 0.066 \times 0.606 \approx 0.04$$

(ii) Dividend = No

$$p(4 \mid \mu_{(0)}, \sigma^2) = \frac{1}{\sqrt{2\pi(36)}} \exp\left(-\frac{(4-0)^2}{72}\right)$$

$$= 0.666 \times 0.8 \approx 0.052.$$

Now to find $p(y=1 \mid x=4) = \dfrac{p(4 \mid y=1)\, P(y=1)}{P(4 \mid y=1) P(y=1) + p(4 \mid y=0)\, p(y=0)}$

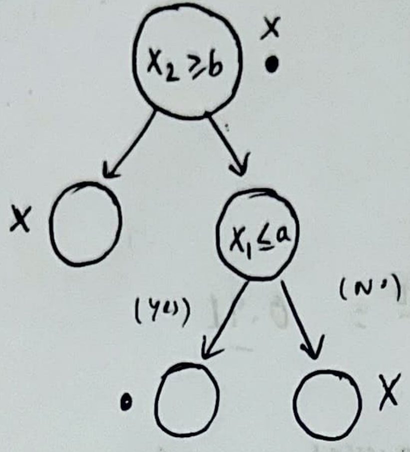
$$= \frac{0.8 \times 0.04}{0.8 \times 0.04 + 0.2 \times 0.052}$$

$$\approx 0.75.$$

$\therefore$ P(Company issues divided $\mid x = 4) \approx 0.75$.

2. Assignment 3

Q1.



Q2. Random forests uses multiple decision trees (hence called a forest) for regression problems and relies on bootstrapping & feature selection. In both these processes there is randomness involved (randomly sampling training data with replacement to create new datasets, random feature selection) and helps our model to be less sensitive to the original training data and reduces correlation between the trees.

Q3. Ensemble methods combine "weaker" models to form a stronger model so that if one base model is prone to error, it can be auto-corrected by others so that the final model is more robust and unlikely to be influenced by small changes in the training data.
Yes, combining the predictions of multiple decision trees can be called an ensemble method, as in the case of random forests, as it uses the predictions to provide a stronger output.

Q4.

1. True positives : 180
   False positives : 70
   True negatives : 730
   False Negative : 20

2. Accuracy = $\frac{180 + 730}{1000} = \underline{0.91}$

   Precision = $\frac{\text{Correct +ve guesses}}{\text{Total +ve guesses}} = \frac{180}{180 + 70} = \underline{0.72}$.

   Recall = $\frac{\text{Correct +ve guesses}}{\text{All positive labels}} = \frac{180}{180 + 20} = \underline{0.9}$

   Specificity = $\frac{\text{Correct -ve guesses}}{\text{All -ve labels}} = \frac{730}{730 + 70} = \underline{0.9}$

   F1 Score = $\frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision + recall}} = \frac{2 \times 0.72 \times 0.9}{1.62}$

   $= \underline{0.8} \Rightarrow 0.8$.

3. Specificity focuses on correct -ve guesses, so we would prioritise this metric to decrease false negatives.

4. Lower classification threshold would result in More positives so higher true +ves & false +ves.
   ∴ Recall & Precision are most likely to increase so F1 Score would also increase.

5. Yes, as accuracy only checks the sum of correctly evaluated guesses so they could have different confusion matrices with respect to the number of true +ves, ~~true~~ false -ves true -ves and so on individually.

We learnt about the math behind the linear regression model and derived the weight and bias, maximum likelihood estimation for a gaussian model, About logistic regression and why its preferred, classification methods, decision trees & random forests, confusion matrix and writing a simple code for executing the logistic regression model.