Q1) If the w vector is orthogonal to the subspace spanned by the datapoint,

$$w^T x_i = 0 \quad \forall i$$

Therefore, SE for a prediction will be,

$$SE = \|w^T x_i - y_i\|_2^2 = \|0 - y_i\|_2^2 = \|y_i\|^2$$

Q2) The argument for why an option is right or wrong is as follows:

a) $\underline{h(x_i) = \bar{y} \ \forall i}$; Here, if the predictions are always $\bar{y}$ SSE won't be zero as for an individual point the $SE \geq 0$. Therefore, $SSE \geq 0$.

b) $\underline{h(x_i) = w^T x_i \ \forall i}$; This is the standard regression form which may or may not always give a perfect linear mapping.

c) $\underline{h(x_i) = c}$; As with option (a), in this case too for each datapoint, $SE \geq 0$. Therefore, $SSE \geq 0$.

d) $\underline{h(x_i) = y_i}$; In this option, as the prediction is always equal to the label, $SE = 0$. $\therefore SSE = 0$.

Q3) $w = \phi(X)\begin{bmatrix}1.3 & 0.6 & -0.2 & -0.7\end{bmatrix}^T$

$k\left(x, \begin{bmatrix}0\\0\\0\\0\end{bmatrix}\right) = \left(x^T\begin{bmatrix}0\\0\\0\\0\end{bmatrix}+1\right)^3 = (0+1)^3 = 1$

$\therefore K = \begin{bmatrix}1\\1\\1\\1\end{bmatrix}$ $\quad\therefore y_{pred} = K^T\alpha = \begin{bmatrix}1 & 1 & 1 & 1\end{bmatrix}\begin{bmatrix}1.3\\0.6\\-0.2\\-0.7\end{bmatrix}$

$$\boxed{\therefore y_{pred} = 1}$$

Q4) Given, $||w^g - w^*|| < ||w^{sg} - w^*||$

Therefore, w^g is closer to w^* than w^{sg}.

Hence Option (a) ie Gradient Descent gives lesser training error the SGD.

Q5 and 6) $X = \begin{bmatrix}-1 & 0 & 2\end{bmatrix}$ $y = \begin{bmatrix}1\\-1\\1\end{bmatrix}$

Adding the bias feature to the dataset, we get

Q5) $y_i = \beta_0$, $X = \begin{bmatrix}1 & 1 & 1\end{bmatrix}$

$\therefore y_{pred} = (XX^T)^{-}Xy = \left(\begin{bmatrix}1 & 1 & 1\end{bmatrix}\begin{bmatrix}1\\1\\1\end{bmatrix}\right)^{-}\begin{bmatrix}1 & 1 & 1\end{bmatrix}\begin{bmatrix}1\\-1\\1\end{bmatrix}$

$= 3^{-} \times 1 = \underline{\underline{0.33}}$

Q6) $y_i = \beta_1 x_i$, $X = \begin{bmatrix}-1 & 0 & 2\end{bmatrix}$

$\therefore y_{pred} = (XX^T)^{-}X_y = \left(\begin{bmatrix}-1 & 0 & 2\end{bmatrix}\begin{bmatrix}-1\\0\\2\end{bmatrix}\right)^{-}\begin{bmatrix}-1 & 0 & 2\end{bmatrix}\begin{bmatrix}1\\-1\\1\end{bmatrix}$

$= \frac{1}{5} \times 1 = \underline{\underline{0.2}}$

Q7)

$$\hat{w}_{ridge} = (XX^T + \lambda I)^{-1} Xy = \left([-3\ 5\ 4]\begin{bmatrix}-3\\5\\4\end{bmatrix} + 50 \right)^{-1} [-3\ 5\ 4]\begin{bmatrix}-10\\20\\20\end{bmatrix}$$

$$= \frac{1}{100} \times 330 = 3.3$$

$$\hat{w}_{MLE} = (XX^T)^{-1} Xy = \left([-3\ 5\ 4]\begin{bmatrix}-3\\5\\4\end{bmatrix} \right)^{-1} [-3\ 5\ 4]\begin{bmatrix}-10\\20\\20\end{bmatrix}$$

$$= \frac{1}{50} \times 330 = 6.6$$

$$\therefore \frac{\hat{w}_{ridge}}{\hat{w}_{MLE}} = \frac{3.3}{6.6} = \underline{\underline{0.5}}$$

Q8) When (x_2, y_2) is in validation set,

$$\tilde{X} = \begin{bmatrix}1 & 1\\0 & 3\end{bmatrix} \quad \tilde{y} = \begin{bmatrix}2\\1\end{bmatrix}$$

Solving the above using simultaneous equations,

$$w_0 = 2 \qquad \therefore w_0 = 2, \ w_1 = -1/3$$

$$w_0 + 3w_1 = 1$$

Q 9-13)	1	0
Root :	300	200
Left C:	50	150
Right C:	250	50

Q9) From the table, we can say that the left child is labelled 0.

Q10) $p = \frac{300}{500} = 0.6$

$$\text{Entropy} = -(p \log_2 p + (1-p)\log_2(1-p))$$

$$= -(0.6 \log_2 0.6 + 0.4 \log_2 0.4)$$

$$= \underline{\underline{0.97}}$$

Q 11) $P = \frac{50}{200} = 0.25$ $\text{Entropy} = -(0.25 \log_2 0.25 + 0.75 \log_2 0.75)$

$= \underline{\underline{0.811}}$

Q 12) $P = \frac{250}{300} = 0.866$ $\text{Entropy} = -(0.86 \log_2 0.866 + 0.14 \log_2 0.14)$

$= \underline{\underline{0.65}}$

Q 13) $IG = \text{Entropy (Parent)} - [r_{LC} \text{ Entropy}(LC) + r_{RC} \text{ Entropy}(RC)]$

$r_{LC} = \frac{200}{500} = 0.4$ $\quad r_{RC} = \frac{300}{500} = 0.6$

$IG = 0.97 - (0.4 \times 0.811 + 0.6 \times 0.65)$

$\therefore IG = \underline{\underline{0.256}}$

Q 14) $x_1 \in (0, 4)$ $\quad \therefore$ Volume of $S = 4 \times 3 \times 2 = 24$

$x_2 \in (0, 3)$

$x_3 \in (0, 2)$

Q 15) S1 is true because $k=1$ and a point is its own neighbor. Therefore, no point in the training set is misclassified.

S2 is false because the model is overfit.

Q 16) For a linear classifier like Perceptron, Label $= 1$ if $w^T x_i \geq 0$ else 0.

$\therefore$ Using the above equation, we can verify that options $\underline{\underline{\text{(a) and (b)}}}$ are correct.

Q17) For Naive Bayes, the number of parameters to be estimated are, Outputs $\times$ d + (outputs - 1)

Here, outputs = 3 d = 3

$\therefore$ Parameters = $3 \times 3 + (3-1) = 9 + 2 = \underline{\underline{11}}$

Q 18 - 19)

Q 18) $\hat{p}^0_3 = P(f_3 = 1 \mid y = 0) = \dfrac{P(f_3 = 1, y = 0)}{P(y=0)}$

$= \dfrac{1/4}{1/2} = \dfrac{1}{2} = \underline{\underline{0.5}}$

$\therefore \hat{p}^0_1 = 0 \quad \hat{p}^0_2 = 0.5$

$\hat{p}^1_1 = 0.5 \quad \hat{p}^1_2 = 0 \quad \hat{p}^1_3 = 0.5$

Q19) $P\left(y=0 \,\middle|\, x = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}\right) = \hat{p}^0_1 \times (1 - \hat{p}^0_2) \times (1 - \hat{p}^0_3) \times (1 - \hat{p})$

$= 0$

$P\left(y=1 \,\middle|\, x = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}\right) = \hat{p}^1_1 \times (1 - \hat{p}^1_2) \times (1 - \hat{p}^1_3) \times (1 - \hat{p})$

$= 0.5 \times 1 \times 0.5 \times 0.5$

$= \underline{\underline{0.125}}$

$\therefore P\left(y=1 \,\middle|\, x = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}\right) > P\left(y=0 \,\middle|\, x = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}\right)$

$\therefore$ Label = 1

Q20) $P(y=1 \mid x) = \dfrac{P(x \mid y=1) \times p(y=1)}{p(x)}$

As we don't know the value for $p(y=1)$ and $p(y=0)$, we can't predict the label for x

$\therefore$ Option (c) is correct

Q21) $P(y \mid x) = \dfrac{P(x, y)}{P(x)}$

$\therefore$ Option a is correct.
