N. KAUSIK
CS21M037

Tutorial 4

1. A) True
As we are provided the target values.

2. B) Predicting dep vars using indep vars
As in regression we try predicting value of dependent variables as a function of independent variables.

3. B) Relation between x_1 and y is strong
As corr coeff $= -0.95$, absolute value $= 0.95$ which is high (strong relation)

4. D) $X^T X \theta = X^T Y$
As we need to minimise sum squared error,

Minimise $(Y - X\theta)^T (Y - X\theta)$

$= (Y^T - \theta^T X^T)(Y - X\theta)$

$= Y^T Y - Y^T X\theta - \theta^T X^T Y + \theta^T X^T X\theta$

$= Y^T Y - Y^T X\theta - (Y^T X\theta)^T + (X\theta)^T X\theta$

as $Y^T X\theta$ is a scalar, it is equal to $(Y^T X\theta)^T$

$= Y^T Y - 2Y^T X\theta + (X\theta)^T X\theta$

To find minima, diff wrt θ and equate to 0,

$$\Rightarrow \frac{\partial (Y^T Y)}{\partial \theta} - 2\frac{\partial}{\partial \theta}\left((X^T Y)^T \theta\right) + \frac{\partial}{\partial \theta}\left(\theta^T (X^T X)\theta\right) = 0$$

Using identities, $\frac{\partial (M)}{\partial \theta} = 0$,

$$\frac{\partial (M\theta)}{\partial \theta} = M^T,$$

$$\frac{\partial (\theta^T M\theta)}{\partial \theta} = M\theta + M^T\theta,$$

$\Rightarrow 0 - 2X^TY + X^TX\theta + (X^TX)^T\theta = 0$

$\Rightarrow -2X^TY + 2X^TX\theta = 0$

$\Rightarrow X^TX\theta = X^TY$,,

5. A) Vertical Offset
As horizontal axis is independent var and vertical axis is dependent var, we consider only vertical offset for fitting.

6. A) 1 and 2
As it is underfitting we can add more variables or increase complexity by introducing polynomial degree vars.
We remove variables only for Over-fitting.

7. D) None of these
We do regularisation only for overfitting. For underfitting we dont do any reg.

8. B) -0.21
Since, $\hat{y} = -2.29 + 1.7x$
Given point $(x, y) = (5, 6)$
For $x = 5$, $\hat{y} = -2.29 + 1.7 \times 5 = -2.29 + 8.5$
$\hat{y} = 6.21 \quad y = 6$
Residual $= y - \hat{y} = 6 - 6.21$
$= -0.21$,,

9. C) L2

10. B) Absolute value of magnitude
As in Lasso we use L1 → absolute value.

11. <u>L1 regularisation</u>
Variables are weighted by 1 or 0. It can remove variables which dont contribute to the target. It is like setting a Laplacian prior on the terms.

<u>L2 regularisation</u>
It will not completely remove variables as instead of 1 or 0 (binary), it tries to spread the error amongst all the variables. It is like setting a Gaussian prior on the terms.

12. Ridge regression is preferred over Lasso in cases where we want to get a solution without completely removing any feature that contributes to the prediction.
Lasso is preferred if we want to reduce dimensions of the data.
However Lasso solution is not invertible but ridge regression solution is invertible

13. $H = X(X^TX)^{-1}X^T$

To prove symmetric,

$H^T = \left(X(X^TX)^{-1}X^T\right)^T = (X^T)^T\left((X^TX)^{-1}\right)^T X^T$

$= X\left((X^TX)^T\right)^{-1}X^T = X(X^TX)^{-1}X^T = H$

as $H^T = H$, H is symmetric matrix.

To prove idempotent,

$HH = X(X^TX)^{-1}X^T \cdot X(X^TX)^{-1}X^T$

$= X\underbrace{\left[(X^TX)^{-1}(X^TX)\right]}_{A^{-1}A = I}(X^TX)^{-1}X^T$

$= X\, I\, (X^TX)^{-1}X^T$

$= X(X^TX)^{-1}X^T = H$

as $HH = H$, H is idempotent matrix.

14. By applying orthogonalisation on dimensions in linear regression, we can separately get the coeffs of regression by doing univariate regression on each orthogonal dimension.
As these dimensions are orthogonal, coeffs in one dimension dont depend on others.

15. If dimensions are nearly correlated, if we change one dimension, others also get affected. Hence their coeffs shoot up to large values and hence system becomes very unstable.

16. Forward Stepwise Selection,
Variables are added one after another to set of added variables to update the best fit. We consider performance measure as the Residual Error on a test data.

Forward Stagewise Selection,
Variable with maximum correlation with residual is selected at each stage. Regression is then done using the now selected variable and it is added to the predictor.