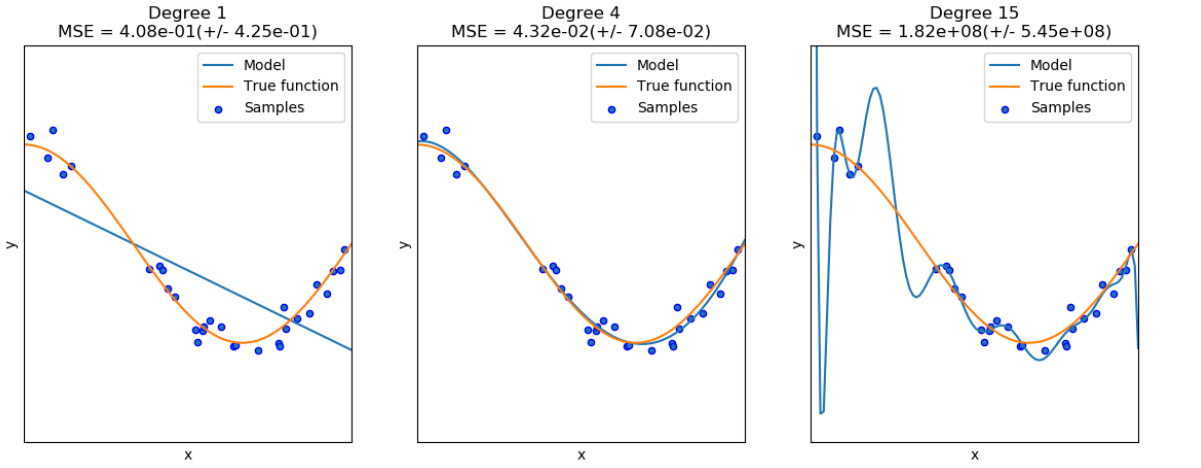
Q1. Explain Underfitting and Overfitting in Brief.

Underfitting:

(a) A statistical model or a machine learning algorithm is said to have underfitting when it cannot capture the underlying trend of the data. It has poor performance on the training da ta and ***poor generalization*** to other data.
(b) Underfitting destroys the accuracy of our machine learning model.
(c) Its occurrence simply means that our model or the algorithm does not fit the data well eno ugh.
(d) It usually happens when we have less data to build an accurate model and also when we try to build a linear model with a non-linear data.
(e) In such cases the rules of the machine learning model are too easy and flexible to be appli ed on such a minimal data and therefore the model will probably make a lot of wrong pre dictions.
(f) Underfitting can be avoided by using more data and also reducing the features by feature selection.

For example, when a given model yields a large training MSE and a large test MSE, we are said to be underfitting the data



Underfitting (High Bias) **Optimal** **Overfitting (High Variance)**

Overfitting:

(a) A statistical model is said to be overfitted, when we train it with a lot of data. It has good performance on the training data, ***poor generliazation*** to other data
(b) When a model gets trained with so much of data, it starts learning from the noise and ina ccurate data entries in our data set.
(c) Then the model does not categorize the data correctly, because of too much of details and noise. Thus model has ***high Variance***.
(d) The causes of overfitting are the non-parametric and non-linear methods because these ty pes of machine learning algorithms have more freedom in building the model based on th e dataset and therefore they can really build unrealistic models.
(e) Overfitting can be avoided by using methodologies like cross-validation, early stopping, pruning, regularization.

For example, when a given model yields a small training MSE but a large test MSE, we are said to be overfitting the data

Q3.What is Regularization? When you will use it? Effect of choosing different values of Lambda during Regression?

a) Overfitting or high variance is caused by hypothesis function that fits the trained data but does not generalize well to predict new data. It is usually caused by a complicated function that creates a lot of unnecessary curves and angles unrelated to the data.
b) Regularization takes care of the overfitting by keeping all the features but reducing the magnitude of parameters θ_j.
c) Regularization tries to push the coefficients for many variables to zero and hence reduce cost term.
d) The genral formula for regularization is as follows:

$$J(\theta) = \frac{1}{2}\left[\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2\right] + \lambda\sum_{j=1}^{n}\left(\theta_j^2\right)$$

Where λ is the regularization parameter. It determines how much the costs of our theta parameters are inflated

The effect of different values of lambda on regression

Linear regression with regularization

Model: $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \frac{\lambda}{2m}\sum_{j=1}^{m}\theta_j^2$$

The equation above describes fitting a high order polynomial with regularization (used to keep parameter values small)
Consider three cases

λ = large

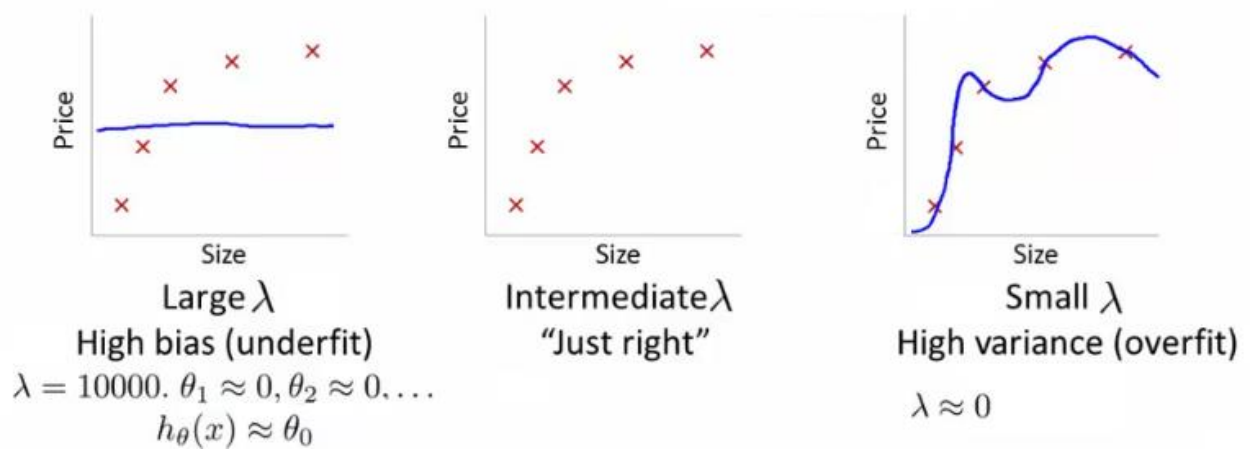
All θ values are heavily penalized
So most parameters end up being close to zero
So hypothesis ends up being close to 0
So high bias --- underfitting data

λ = intermediate

Only this values gives the fitting which is reasonable

λ = small

Lambda = 0
So we make the regularization term 0
So high variance ---- overfitting data



The effect of lambda on cross validation error and training error can be shown by the following graph