Machine Learning

CS F464

BITS Pilani
Dubai Campus

Dr. Pranav M. Pawar

Contents

- Regression
  - Introduction
  - Simple Linear Regression
  - Model Evaluation
  - Evaluation Metrics
  - Multiple Linear Regression
- Overfitting and Underfitting
- Linear Basis of Function Model
- Bias-variance trade-off
- Regularization

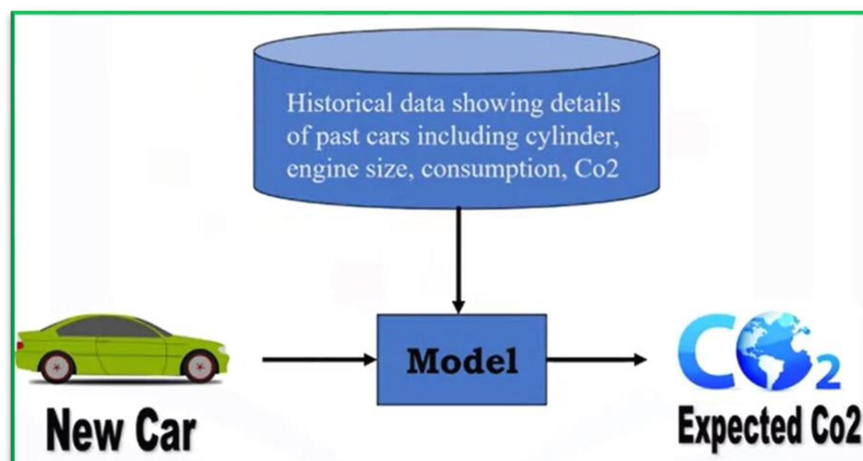
Regression

- Regression is the process of predicting a continuous values.
- Dependent variable can be seen as the state, target, or final goal we study and try to predict.
- Independent variables, also known as explanatory variables, can be seen as the causes of those states.
- A regression model relates Y or the dependent variable to a function of X
- Regression Model

X: Independent variable   Y: Dependent variable

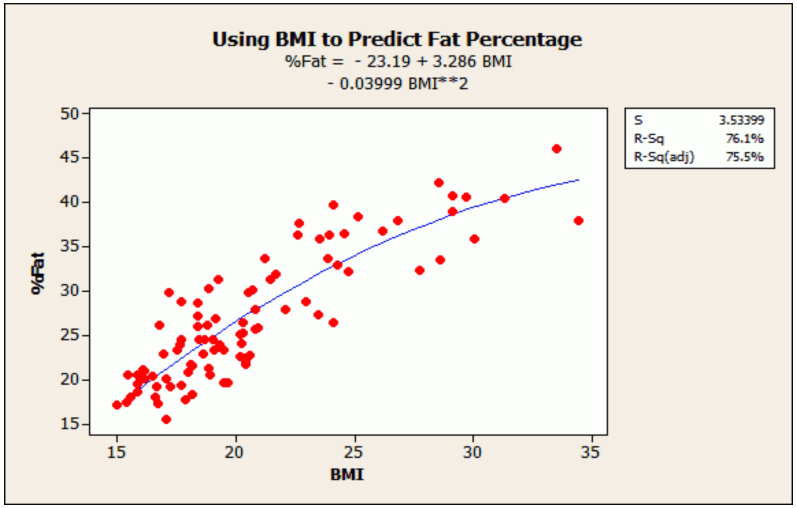
	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Continuous Values

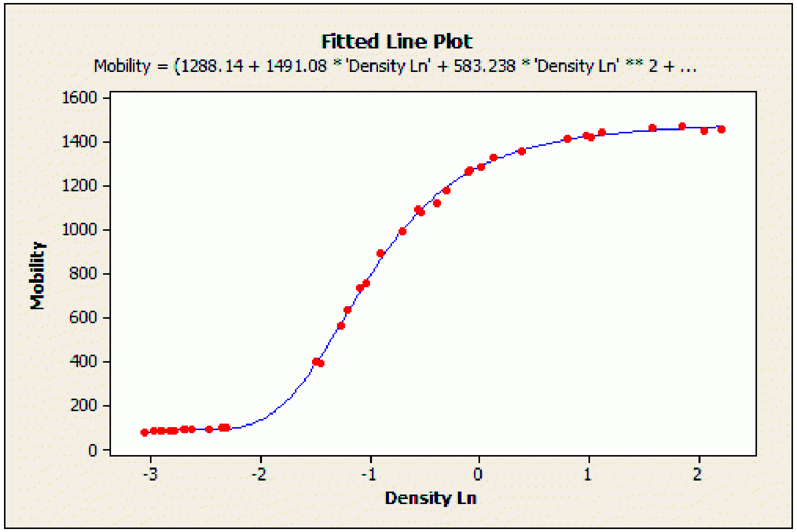


Regression Model

- Types of Regression Models
  - Linear Regression
    - Simple
    - Multiple



  - Non Linear Regression
    - Simple
    - Multiple



Linear Regression

- Linear regression is the approximation of a linear model used to describe the relationship between two or more variables.
- In simple linear regression, there are two variables, a dependent variable and an independent variable.

X: Independent variable

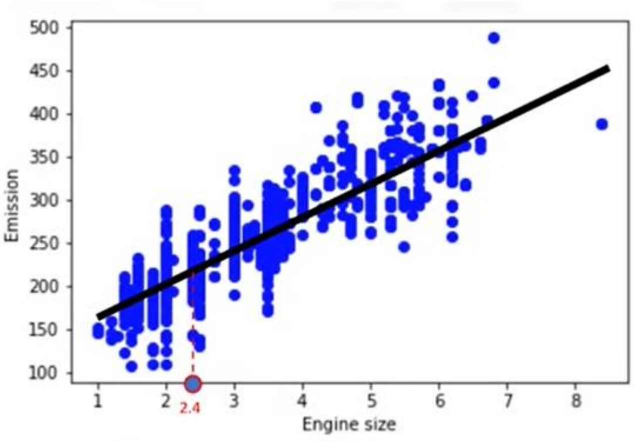
Y: Dependent variable

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

Continuous Values

Linear Regression

- Here, Scatter plot shows the relation between variables where changes in one variable cause changes in the other variable.
- With linear regression we can fit line through data.
- In case of simple linear regression equation of fit line,



$$\hat{y} = \underbrace{\theta_0 + \theta_1}_{\text{Parameters/Coefficients}} x_1$$

Response variable/Dependant variable

Parameters/Coefficients

Predictor variable/Independent variable

$y = mx + c$

$y = ax + b$

$y = w_1 x + w_0$

Linear Regression

- The aim of regression model is to find values coefficient which minimize errors.
  - $Error = y - \hat{y}$
  - $MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^{2}$
- How to estimate the parameters ?

Linear Regression

$$\hat{y} = \theta_0 + \theta_1 x_1$$

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267

X_1 (ENGINESIZE), y (CO2EMISSIONS)

$$\theta_1 = \frac{\sum_{i=1}^{s}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{s}(x_i - \bar{x})^2}$$

$$\bar{x} = (2.0 + 2.4 + 1.5 + \dots)/9 = 3.03$$

$$\bar{y} = (196 + 221 + 136 + \dots)/9 = 226.22$$

$$\theta_1 = \frac{(2.0 - 3.03)(196 - 226.22) + (2.4 - 3.03)(221 - 226.22) + \dots}{(2.0 - 3.03)^2 + (2.4 - 3.03)^2 + \dots}$$

$$\theta_1 = 39$$

$$\theta_0 = \bar{y} - \theta_1 \bar{x}$$

$$\theta_0 = 226.22 - 39 * 3.03$$

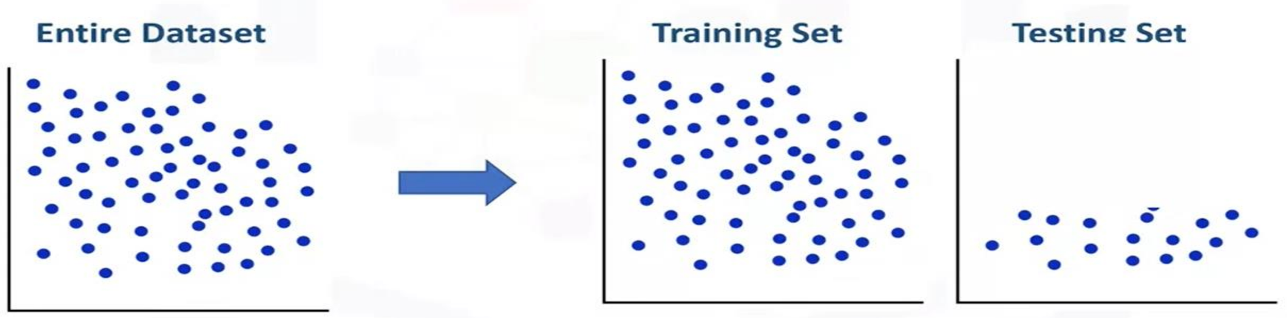
$$\theta_0 = 125.74$$

$$\boxed{\hat{y} = 125.74 + 39 x_1}$$

- Here, θ_0 is called as bias coefficient and θ_1 is the coefficient used for prediction.
- Now we can easily predict Y if X is given to us.

Model Evaluation (1)

- Once model is build we need to evaluate the model.
- Two methods for model evaluation,
  - Train and test on same dataset
  - Train/Test Split
- Train and Test on same dataset



  - High training accuracy and low out of sample accuracy
    - Training accuracy is the percentage of correct predictions that the model makes when using the test dataset.
      - Having a high training accuracy may result in an over-fit the data.
    - Out-of-sample accuracy is the percentage of correct predictions that the model makes on data that the model has not been trained on.

Model Evaluation (2)

- Train/Test Split

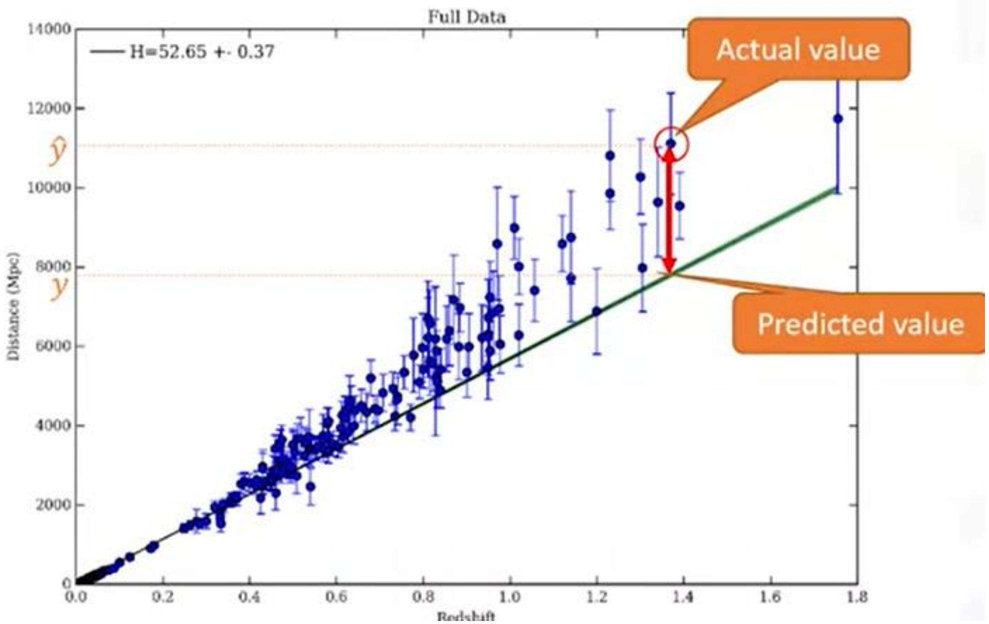
- Training and testing set are mutually exclusive.
- More out of sample accuracy.
- Truly out-of-sample testing
- Ensure that you train your model with the testing set afterwards.
- Highly dependent on dataset on which data is trained and tested.

Model Evaluation (3)

- Have more consistent out of sample accuracy.

Evaluation Metrics for Regression

- For measuring the accuracy of model.
- Accuracy is depends on amount of prediction errors.



$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j|$$

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

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

$$RAE = \frac{\sum_{j=1}^{n}|y_j - \hat{y}_j|}{\sum_{j=1}^{n}|y_j - \bar{y}|}$$

$$RSE = \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2}$$

$$\boxed{R^2 = 1 - RSE}$$

Higher the R^2, better the model fits the data.

Multiple Linear Regression

innovate achieve lead

- When should we use multiple linear regression?
  - Independent variables effectiveness on prediction
    - Does revision time, test anxiety, lecture attendance and gender have any effect on the exam performance of students?
  - Prediction impacts of changes
    - How much does blood pressure go up or down for every unit increase or decrease in the BMI of a patient?

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ ... + \theta_n x_n$$

$$\hat{y} = \theta^T X$$

$$\theta^T = [\theta_0, \theta_1, \theta_2, ...] \qquad X = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ ... \end{bmatrix}$$

X: Independent variable | Y: Dependent variable

	ENGINESIZE	CYLINDERS	FUELCONSUMPTION_COMB	CO2EMISSIONS
0	2.0	4	8.5	196
1	2.4	4	9.6	221
2	1.5	4	5.9	136
3	3.5	6	11.1	255
4	3.5	6	10.6	244
5	3.5	6	10.0	230
6	3.5	6	10.1	232
7	3.7	6	11.1	255
8	3.7	6	11.6	267
9	2.4	4	9.2	?

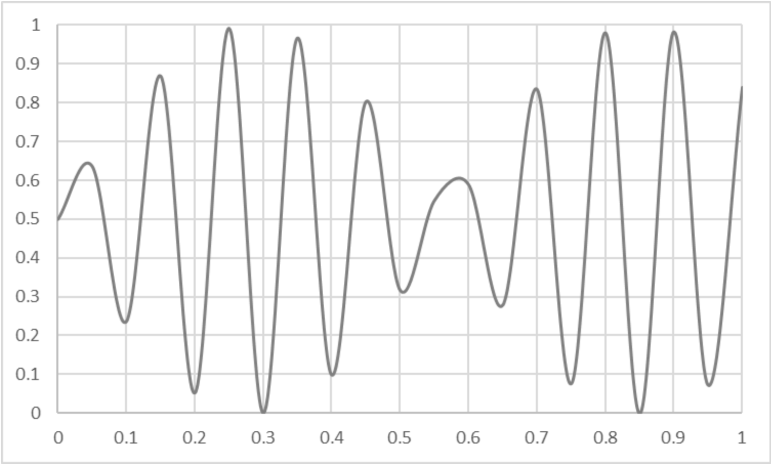
Overfitting and Underfitting (1)

- The central challenge in machine learning is to perform well on new, previously unseen inputs, and not just on the training inputs
- We want the generalization error (test error) to be low as well as the training error
- The model is developed from the training set the expected test error is usually greater than the expected training error.
- Factors determining model performance are
  - Make the training error small
  - Make the gap between train and test error small
- These two factors correspond to the two challenges in machine learning
  - Underfitting and overfitting

Overfitting and Underfitting (2)

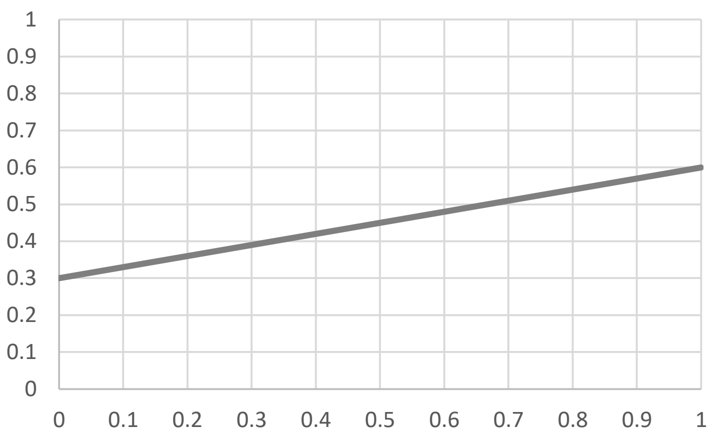
Overfitting

- Fitting the data too well
  - Features are noisy / uncorrelated to concept
  - Modeling process very sensitive (powerful)
  - Too much search



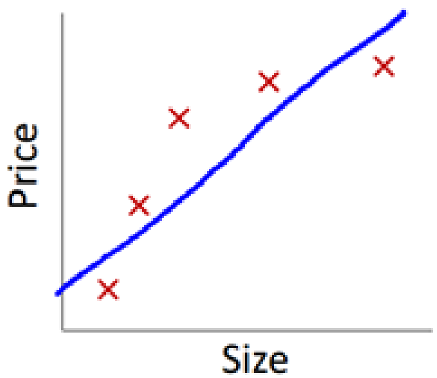
Underfitting

- Learning too little of the true concept
  - Features don't capture concept
  - Too much bias in model
  - Too little search to fit model

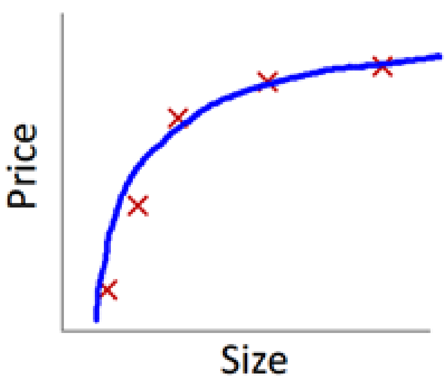


Overfitting and Underfitting in Linear Regression

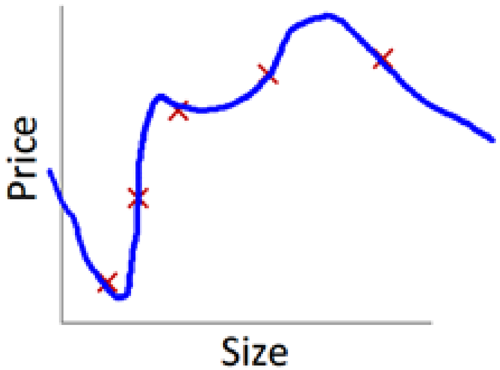
- In over fitting, if we have too many features, the learned hypothesis may fit the training set very well, but fail to generalize to new examples



$$\theta_0 + \theta_1 x$$



$$\theta_0 + \theta_1 x + \theta_2 x^2$$



$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

How to avoid overfitting or underfitting?

- Overfitting
  - Cross Validation
    - Use train-split method or k-fold cross validation.
  - Train with more data
    - Not always work.
  - Reduce features
  - Early stopping
  - Regularization
    - Is a technique used for tuning the function by adding an additional penalty term in the error function for reducing overfitting.
- Underfitting
  - Increase model complexity
  - Increase number of features, performing feature engineering
  - Remove noise from the data.
  - Increase the duration of training to get better results.

Linear Basis Function model (1)

- Simple linear model for regression for D input variables,

$$y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + \ldots + w_D x_D$$

  - Where $\mathbf{x} = (x_1, \ldots, x_D)^{\mathrm{T}}$. are the input variables.
  - It is linear in terms of both parameters (*w*) and input variables.
  - Limitation as it only consider linear input variables (straight line fit).

Linear Basis Function model (2)

- Linear combinations of fixed nonlinear functions of input variables,

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x})$$

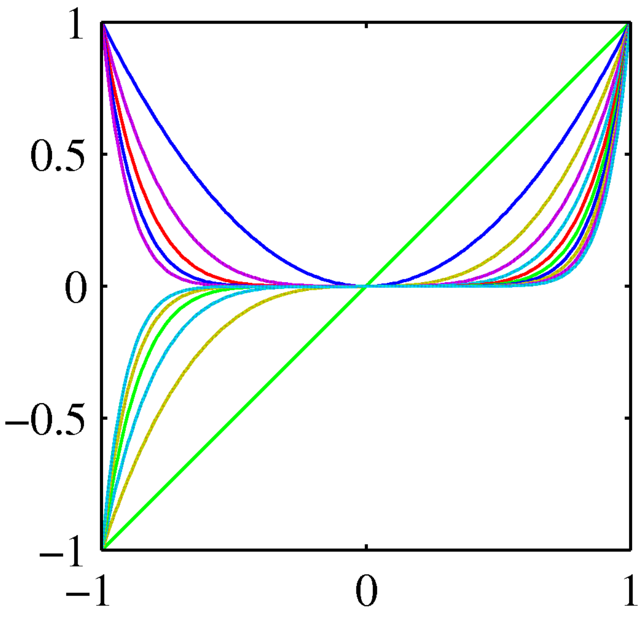
  - Where, $\emptyset_j(x)$ is know as bias function.
  - *M* is total number of parameters in model.
  - w_0 is know as bias parameter (fixed offset).
- The above can also be written as,

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

  - where, $\mathbf{w} = (w_0, \ldots, w_{M-1})^{\mathrm{T}}$ and $\boldsymbol{\phi} = (\phi_0, \ldots, \phi_{M-1})^{\mathrm{T}}$

Linear Basis Function model (3)

- Polynomial regression is a particular example of this model.
  - Example: for single input variable *x*, basis function is $\emptyset_j(x) = x^j$.
- Limitation of polynomial basis function
  - As they are global functions so changes in one region of input space affects others.
  - Can divide input space into regions
    - use different polynomials in each region
    - equivalent to spline functions (complex)
      - Splined functions are form by joining polynomial.

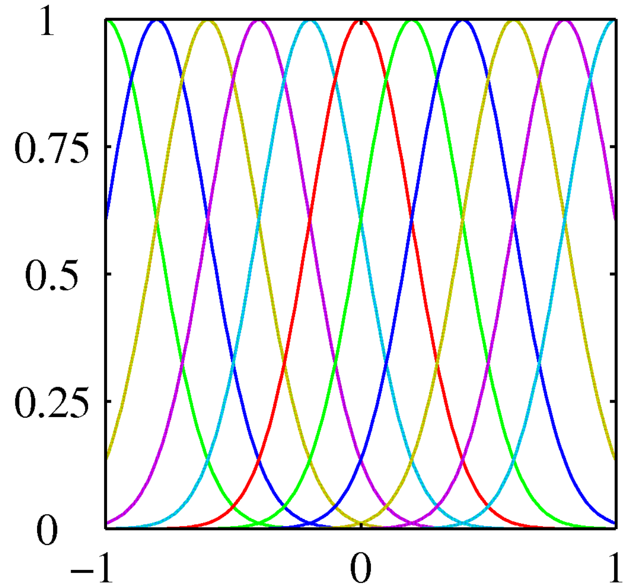


Linear Basis Function model (4)

- Gaussian basis functions:

$$\phi_j(x) = \exp\left\{-\frac{(x-\mu_j)^2}{2s^2}\right\}$$

- These are local; a small change in *x* only affect nearby basis functions. μ_j and *s* control location and scale (width).



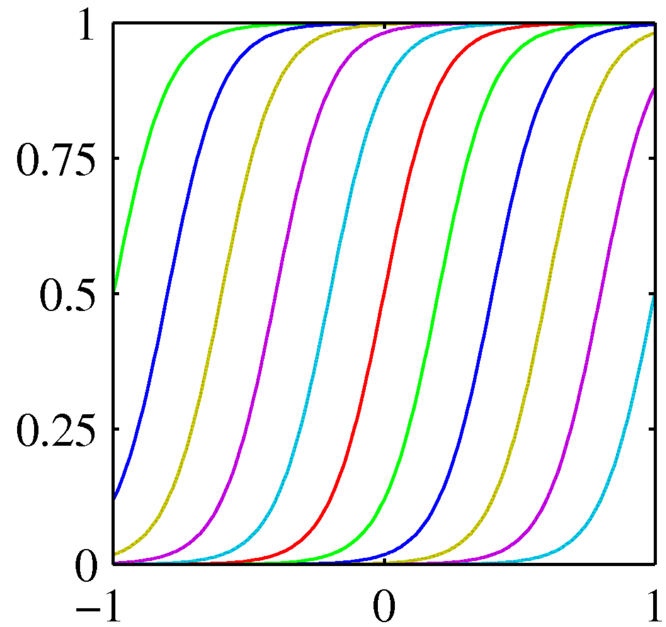
Linear Basis Function model (5)

- Sigmoidal basis functions:

$$\phi_j(x) = \sigma\left(\frac{x - \mu_j}{s}\right)$$

where

$$\sigma(a) = \frac{1}{1 + \exp(-a)}.$$



- Also these are local; a small change in x only affect nearby basis functions. μ_j and s control location and scale (slope).

Bias and Variance (1)

- Bias
  - The inability for machine learning method to capture the true relationship is called **Bias**.
  - Bias is the difference between the average prediction of our model and the correct value which we are trying to predict.
  - Let $f(x)$ be the true model and $\widehat{f(x)}$ be our estimate of the model.

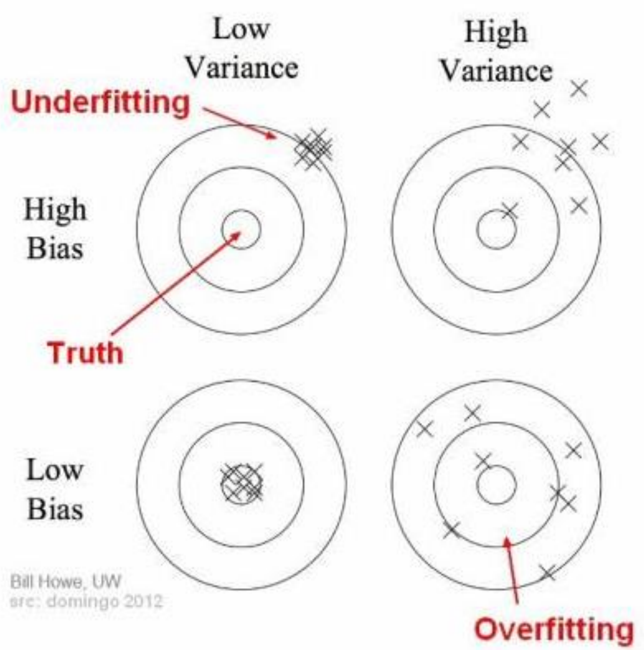
$$Bias\left(\widehat{f(x)}\right) = \mathrm{E}\left[\widehat{f(x)}\right] - f(x)$$

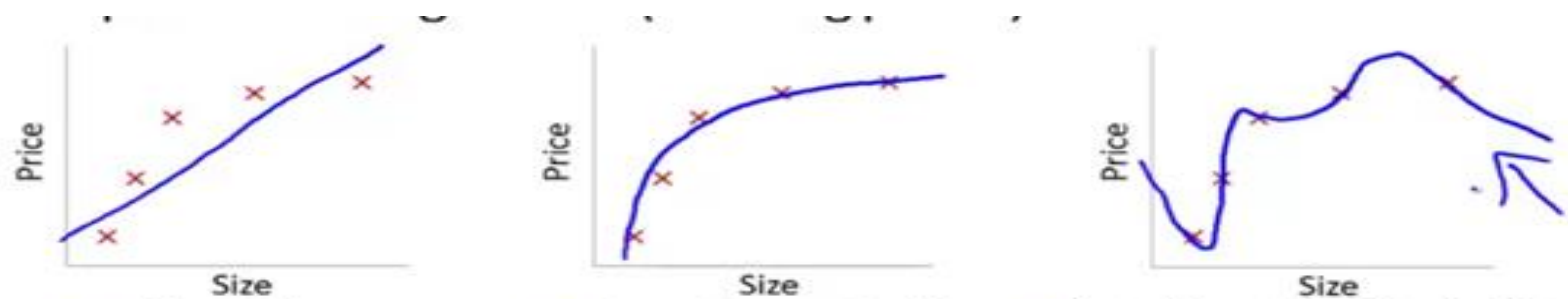
  - Model with high bias pays very little attention to the training data and oversimplifies the model causing leading to underfitting.
  - Simple model has high bias and complex model has low bias

Bias and Variance (2)

- Variance
  - Variance is the variability of model prediction for a given data point or a value which tells us spread of our data.
  - Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before. Such models perform very well on training data but has high error rates on test data.
  - $Variance\left(\widehat{f(x)}\right) = \mathrm{E}\left[\,(f(x) \;-\; \mathrm{E}\left[\widehat{f(x)}\right])^2\right]$
  - It tells how much the different $\left[\widehat{f(x)}\right]$ (trained on different samples of data) differ from each other.
  - Simple model has low variance and complex model has high variance.

Bias Variance Trade-off



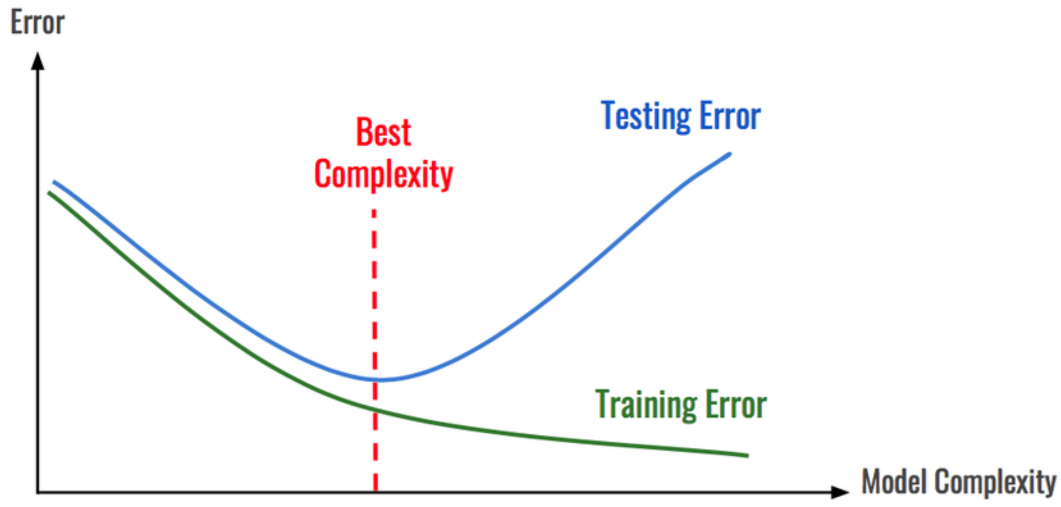


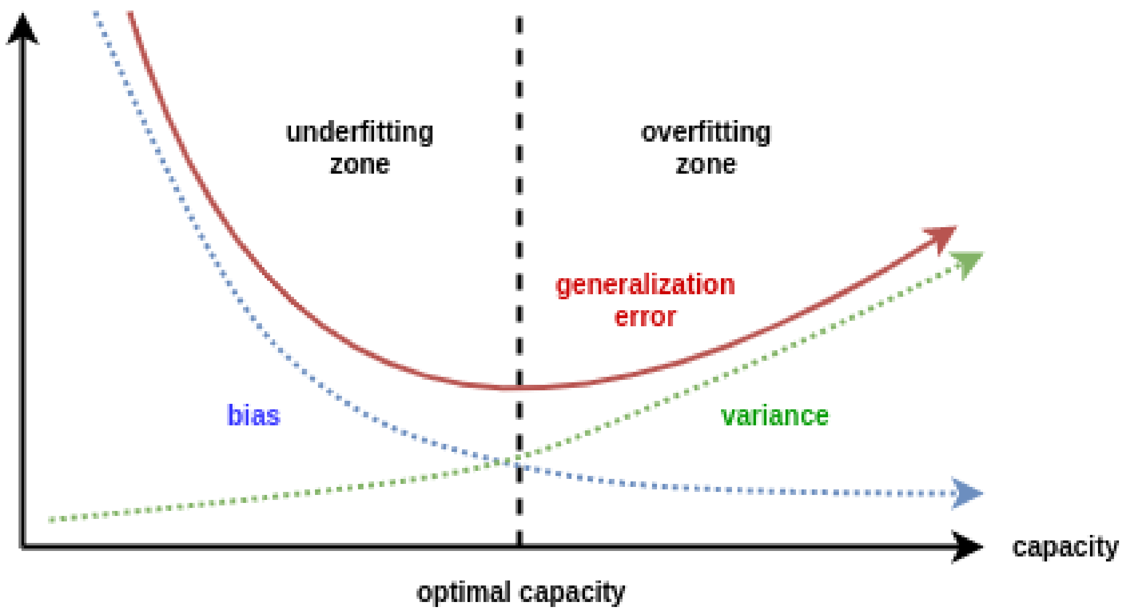
$\rightarrow \theta_0 + \theta_1 x$ "Underfit" "High bias"

$\rightarrow \theta_0 + \theta_1 x + \theta_2 x^2$ "Just right"

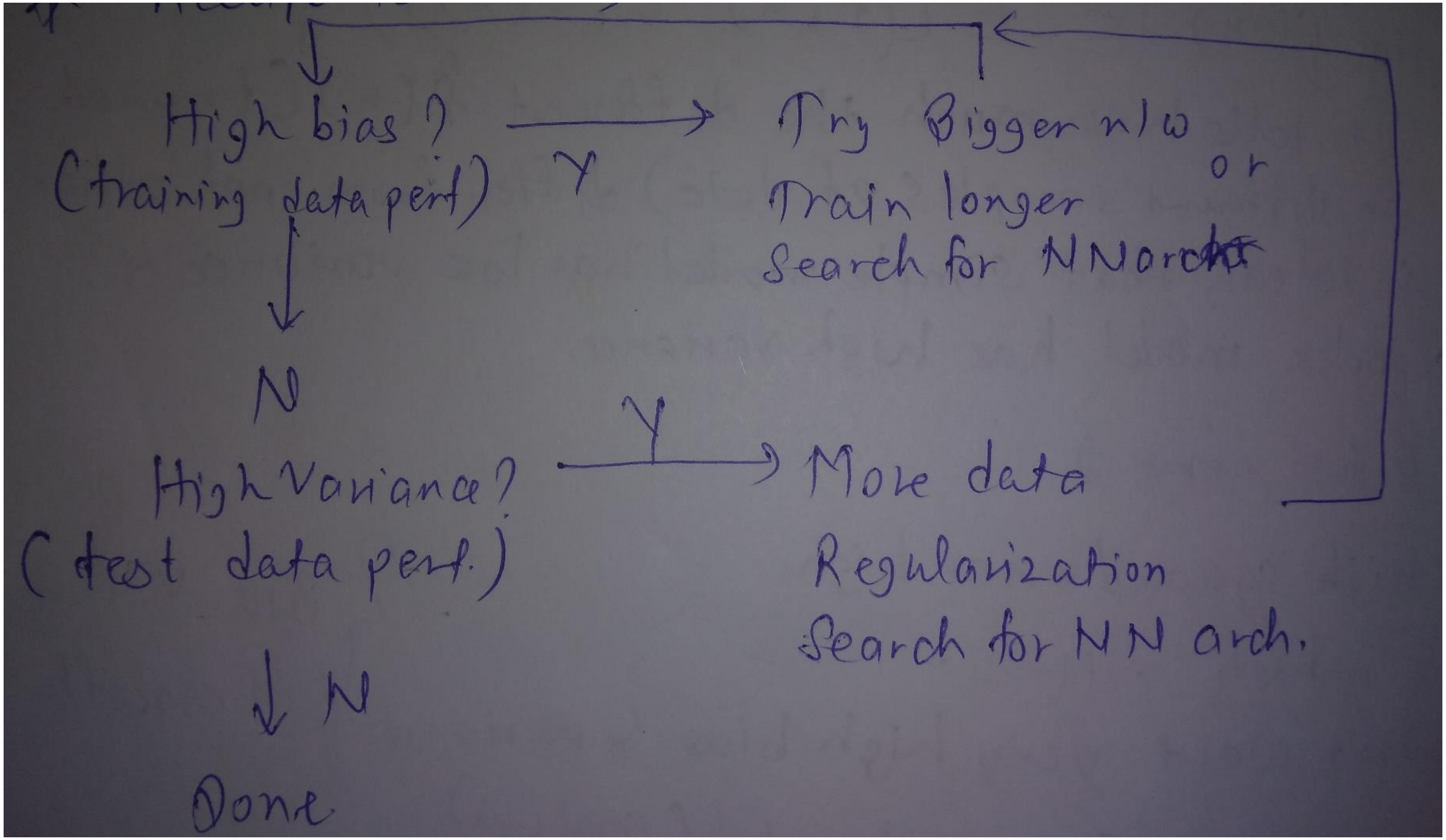
$\rightarrow \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$ "Overfit" "High variance"

Learning Curves (Train, Test Errors, Bias, Variance)





Addition in ML Recipe



Regularization (1)

- Regularisation is a technique used to reduce the errors by fitting the function appropriately on the given training set and avoid overfitting.
- It helps to reduce the variance.
- This technique discourages learning a more complex or flexible model, so as to avoid the risk of overfitting.
- Regularization is a technique used for tuning the function by adding an additional penalty term in the error function.

Regularization (2)

- Consider multiple linear regression model and loss function (RSS: Residual sum of squares),

$$Y \approx \beta 0 + \beta 1X1 + \beta 2X2 + \ldots + \beta pXp$$

$$\mathrm{RSS} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2.$$

- Ridge Regression (L2)

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \mathrm{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2$$

  - λ is the tuning parameter that decides how much we want to penalize the flexibility of our model.

Regularization (3)

- Lasso (L1)

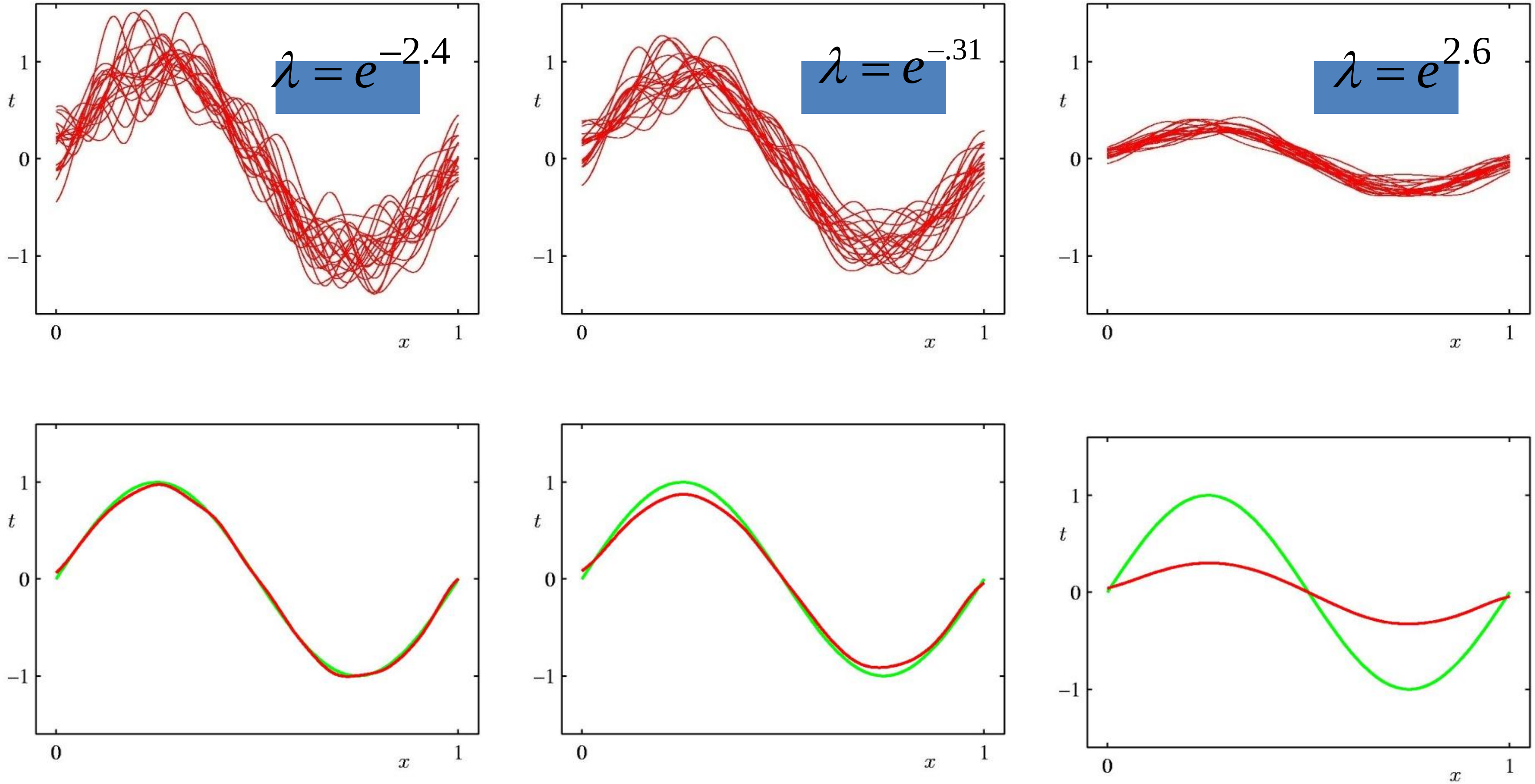
$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|.$$

- How regularization prevents overfitting?
  - The additional regularization factor make some weights approximately zero, which converts it from overfitting to underfitting and correct tuning of λ make it to right case.

Effect Regularization Parameters

high variance

low variance



low bias

high bias

Sources

- Chapter 1, Christopher M Bishop: Pattern Recognition & Machine Leaning, 2006 Springer.
- https://www.cse.unr.edu/~bebis/CS479/Lectures
- https://statisticsbyjim.com/regression/difference-between-linear-nonlinear-regression-models/
- “Machine learning ” course, Andrew Ng
- “Machine learning with Python course”, IBM
- Chapter 3, Christopher M Bishop: Pattern Recognition & Machine Leaning, 2006 Springer.

Thank You!