CSE 2027-Fundamental of Data Analysis

Module 5 – Prediction

- Introduction: Overview,
- Classification,
- Regression,
- Building a Prediction Model,
- Applying a Prediction Model,
- Simple Linear Regression,
- Simple Non Linear Regression.

PRESIDENCY UNIVERSITY
GAIN MORE KNOWLEDGE REACH GREATER HEIGHTS
Private University Estd. in Karnataka State by Act No. 41 of 2013
PRESIDENCY GROUP OVER 40 YEARS OF ACADEMIC WISDOM

Machine learning is divided into two main categories

Machine learning

Supervised learning

Algorithms trained with **labelled data**

Unsupervised learning

Algorithms trained with **unlabelled data**

How does supervised learning work?

Input (x) ❯ **Algorithm** that learns the mapping function from the input to the output = **Output (y)**

A supervised learning technique: Regression

How does regression work?

- Regression models use an algorithm to understand the relationship between **a dependent variable** (input) and **an independent variable** (output).

- They are helpful for **predicting numerical values** based on different features' values. E.g., temperature forecast based on wind, humidity and pressure.

Regression aims

to build a relationship between each feature and the output for predictions

Linear relationships **Linear regression**

Linear regression uses a best fitting straight line – **"regression line"**

$$y = wX + b$$

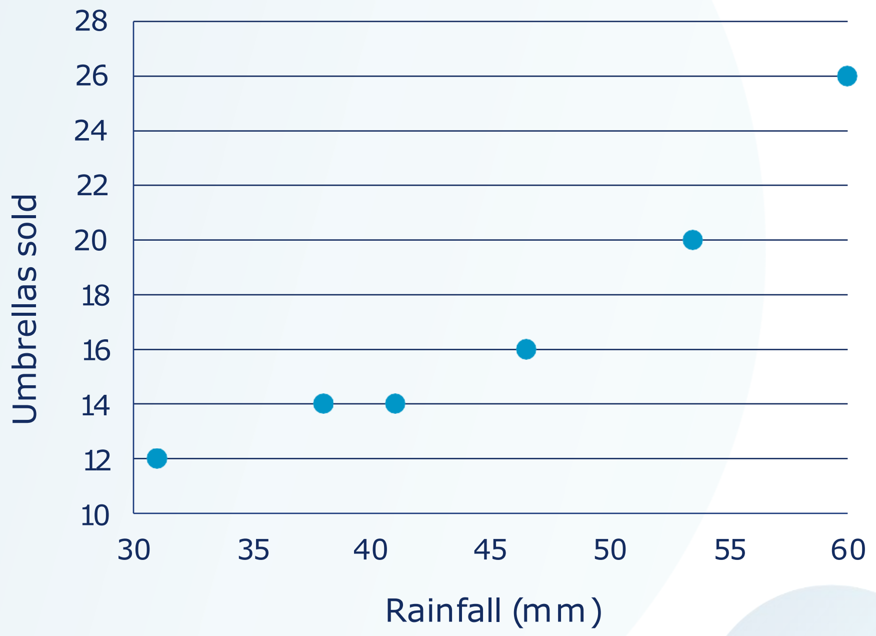
- y: dependent variable
- w: weight; the slope of the gradient of the line - indicates the impact X has on Y
- X: independent variable; used to predict the value of Y
- b: Bias; is the value of Y when there is no X or X is zero

A simple linear regression model

Simple linear regression only has one Y variable and one X variable:

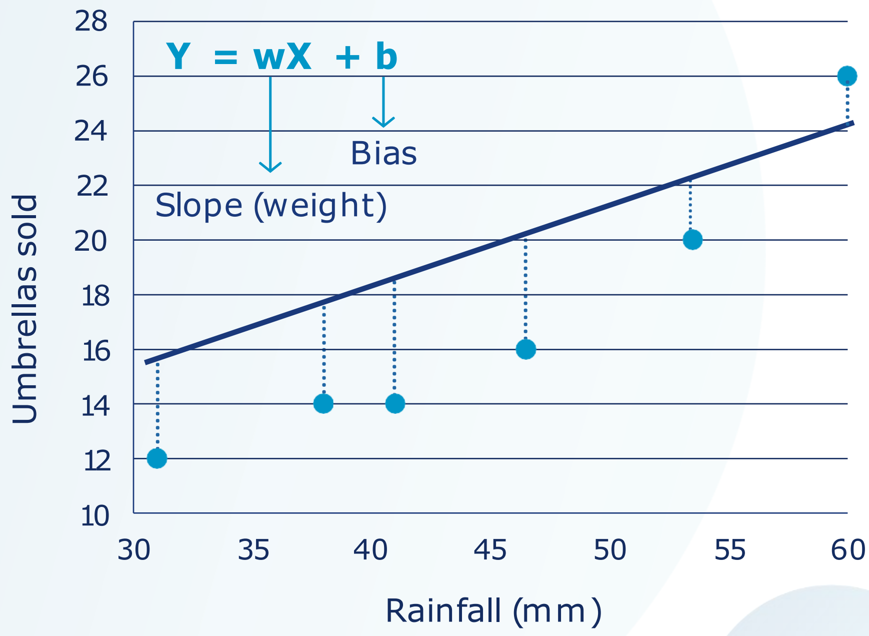
- **The independent variable x:** rainfall measured in millimeters
- **The dependent variable y:** the number of umbrellas sold

▶ We can predict the number of umbrellas, or Y, for any quantity of rain.



How can we calculate the regression line?

1. We draw a line to represent the relationship
2. We measure the distances between the line and each datapoint (the residuals)
3. We sum up the residuals
4. We adjust the weight & the bias to minimize this sum



Multiple features call for multiple linear regression

Multiple features ▶ **Multiple linear regression**

The aim is to **predict output variable** using multiple features

$$y = w_1x_1 + w_2x_2 + \dots + b$$

- Multiple linear regression can have many independent variables to one dependent variable
- Datasets with multiple features like the number of bedrooms, age of the building, covered area, etc.

How can we evaluate the performance of a regression model?

- We use performance evaluation metrics
- The most commonly used evaluation metrics is taking the difference between predicted and actual value of some test points:
- The mean of the squared difference is taken – Mean Squared Error (MSE)
- The size of the error is measured by taking the square root of MSE - Root Mean Squared Error (RMSE)

Evaluating the performance of a regression model using MSE & RMSE

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

RMSE=(MSE)$^{1/2}$

MSE = Mean squared error
n = Number of data points
Y_i = Observed values
$\hat{y}_i$ = Predicted values

Simple Non-Linear Regression

- In situations where the relationship between two variables is nonlinear, a simple way of generating a regression equation is to transform the nonlinear relationship to a linear relationship using a mathematical transformation.
- A linear model can then be generated.
- Once a prediction has been made, the predicted value is transformed back to the original scale.
- For example, in Table 7.10 two columns show a nonlinear relationship.

Simple Non-Linear Regression

- Plotting these values results in the scatterplot in Figure 7.13.
- There is no linear relationship between these two variables and hence we cannot calculate a linear model directly from the two variables.
- To generate a model, we transform x or y or both to create a linear relationship.
- In this example, we transform the v variable using the following formula:

$$y' = \frac{-1}{y}$$

PRESIDENCY UNIVERSITY

GAIN MORE KNOWLEDGE REACH GREATER HEIGHTS

Private University Estd. in Karnataka State by Act No. 41 of 2013

PRESIDENCY GROUP OVER 40 YEARS OF ACADEMIC WISDOM

Simple Non-Linear Regression

Table 7.10. Table of observations for variables x and y

x	y
3	4
6	5
9	7
8	6
10	8
11	10
12	12
13	14
13.5	16
14	18
14.5	22
15	28
15.2	35
15.3	42

Simple Non-Linear Regression

- We now generate a new column, y’ (Table 7.12). If we now plot x against y’, we can see that we now have an approximate linear relationship (see Figure 7.14).

Table 7.12. Prediction of y using a nonlinear model

x	y	$y' = -1/y$	Predicted y'	Predicted y
3	4	−0.25	−0.252	3.96
6	5	−0.2	−0.198	5.06
9	7	−0.143	−0.143	6.99
8	6	−0.167	−0.161	6.20
10	8	−0.125	−0.125	8.02
11	10	−0.1	−0.107	9.39
12	12	−0.083	−0.088	11.33
13	14	−0.071	−0.070	14.28
13.5	16	−0.062	−0.061	16.42
14	18	−0.056	−0.052	19.31
14.5	22	−0.045	−0.043	23.44
15	28	−0.036	−0.033	29.81
15.2	35	−0.029	−0.023	33.45
15.3	42	−0.024	−0.028	35.63

Simple Non-Linear Regression

- Using x we can now calculate a predicted value for the transformed value of y (y').
- To map this new prediction of y' we must now perform inverse transformation, that is, -1/y'.
- In Table 7.12, we have calculated the predicted value for y' and transformed the number to Predicted y.
- The Predicted y values are close to the actual y values.

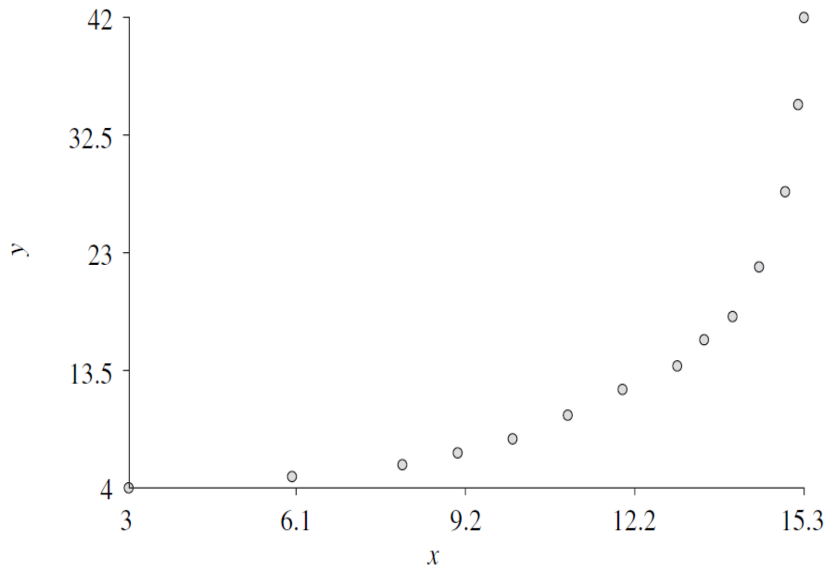
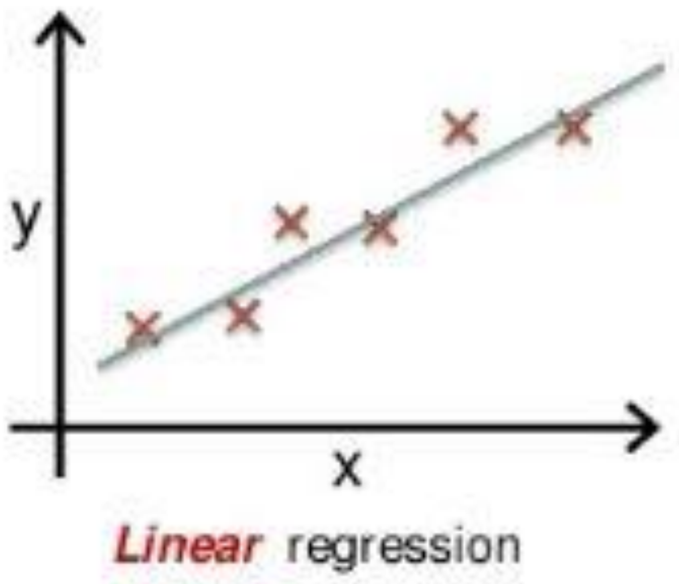


Figure 7.13. Scatterplot showing the nonlinear relationship between *x* and *y*

PRESIDENCY UNIVERSITY
GAIN MORE KNOWLEDGE REACH GREATER HEIGHTS
Private University Estd. in Karnataka State by Act No. 41 of 2013
PRESIDENCY GROUP OVER 40 YEARS OF ACADEMIC WISDOM

Simple Non-Linear Regression

- Some common nonlinear relationships are shown in Figure 7.15.
- The following transformation may create a linear relationship for the charts shown:
  - **Situation a:** Transformations on the x, y or both x and y variables such as **log or square root**
  - **Situation b:** Transformation on the x variable such as square **root, log or -1/x.**
  - **Situation c:** Transformation on the y variable such as square **root, log or -1/y.**
- **This approach of creating simple nonlinear models can only be used when there is a clear transformation of the data to a linear relationship.**



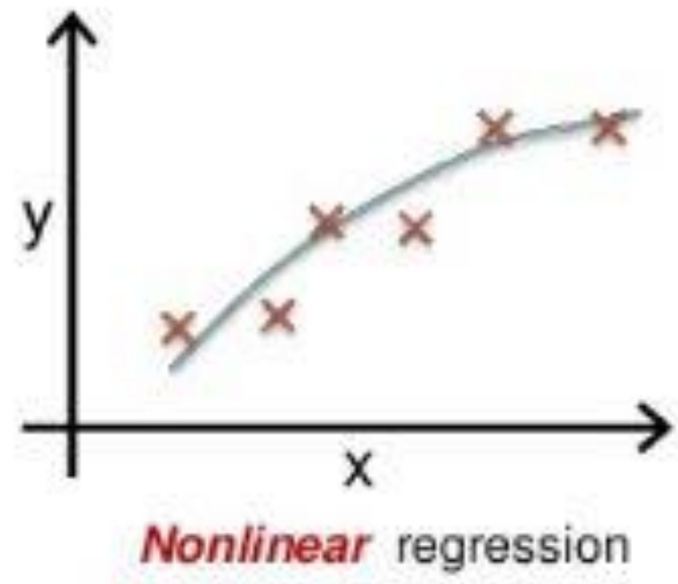


Fig 7.14 and 7.15

A supervised learning technique: classification

What is classification?

- Classification is the process of categorizing a given set of data into classes. The pre-defined classes act as our labels, or ground truth.
- The model uses the features of an object to predict its labels. E.g., filtering spam from non-spam emails or classifying types of fruits based on their color, weight and size.

What types of problems does classification solve?

There are two types of classification problems

Binary

The output is restricted to two classes

Multi-class

The output has more than two classes

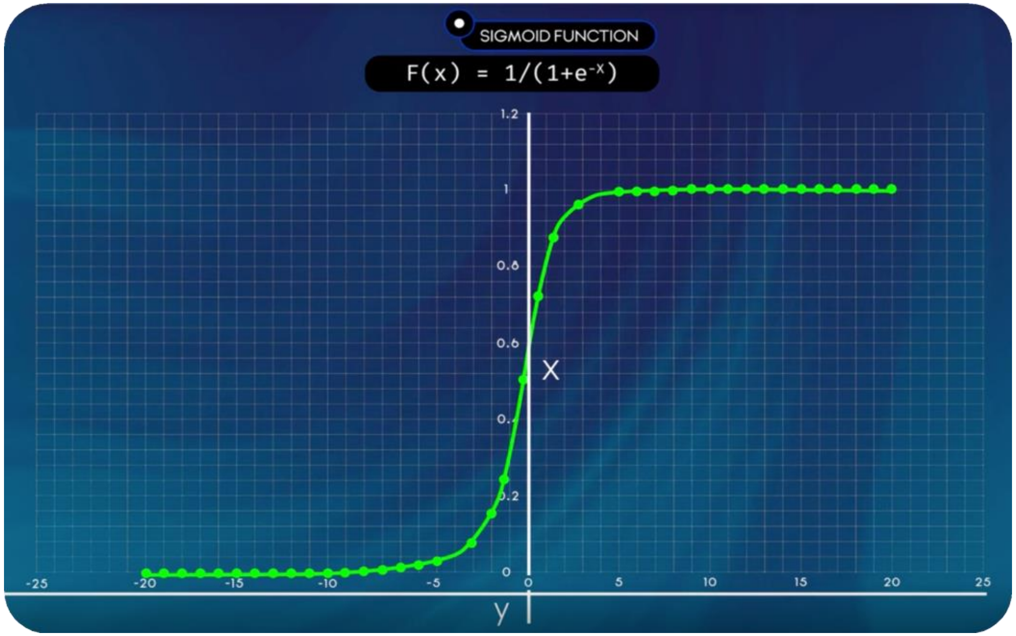
To solve classification problems: logistic regression

What is logistic regression?

Logistic regression is a linear regression but for classification problems. Unlike linear regression, logistic regression **doesn't need a linear relationship** between input and output variables.

Logistic regression uses a logistic function: sigmoid function

The sigmoid function takes any real input, and outputs a value between zero and one.



How can we measure the performance of a logistic regression classifier?

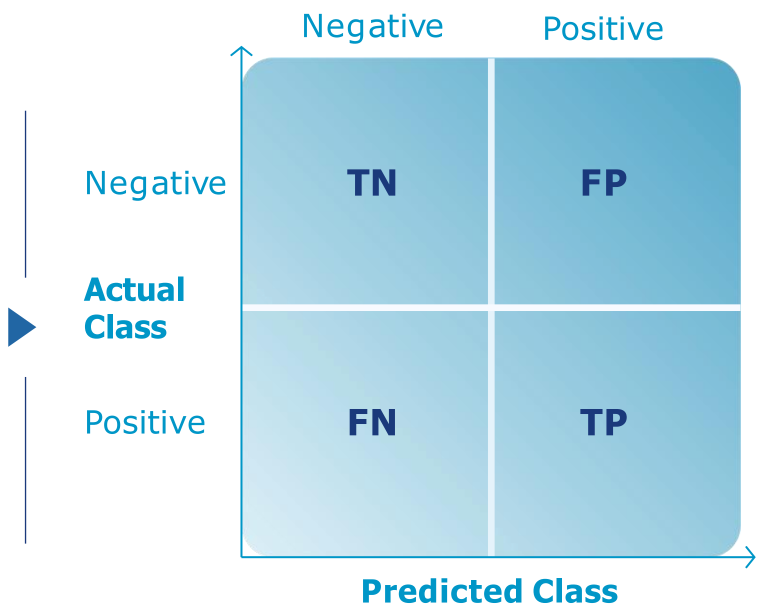
- Once we have the predicted results from our classification model (classifier), the results are compared with the actual label (ground truth)
- Then the performance of the model is being evaluated using the **confusion matrix**

CONFUSION MATRIX

		PREDICTED CLASS	
		NEGATIVE	POSITIVE
ACTUAL CLASS	NEGATIVE	TN	FP
	POSITIVE	FN	TP

Applying the confusion matrix to measure the model performance

- **True positives (TP)** - results which were predicted as positive & ground truth were also positive.
- **False positives (FP)** - instances predicted as positives but actually were negative.
- **True negatives (TN)** - instances predicted as negatives & their ground truth was also negative.
- **False negatives (FN)** - instances predicted as negative but their ground truth was positive.



The evaluation metrics

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)}$$

$$Precision = \frac{(TP)}{(TP+FP)}$$

$$Recall = \frac{(TP)}{(TP+FN)}$$

$$F1\ Score = \frac{2 * Precision * Recall}{Precision + Recall}$$

Support vector machine (SVM)

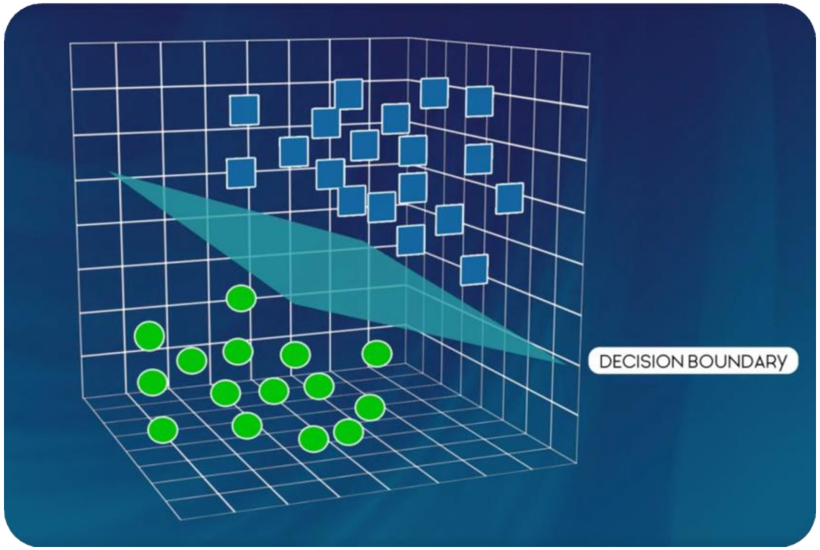
- What is support vector machine (SVM)?

- Support vector machine (SVM), is a supervised ML technique that can be used to solve classification and regression problems. It is, however, mostly used for classification.
- In this algorithm, each feature & data points are plotted in the space. Then, the SVM model finds boundaries to separates different data samples into specific classes.

A practical example: finding a 2D plane that differentiates two classes

Let's say we have a dataset of different animals of two classes: birds & fish

- **There are only three features:** body weight, body length, and daily food consumption
- We draw a **3D grid** and plot all these points
- A SVM model will try to find a 2D plane that differentiates the 2 classes



PRESIDENCY UNIVERSITY
GAIN MORE KNOWLEDGE REACH GREATER HEIGHTS
Private University Estd. in Karnataka State by Act No. 41 of 2013
PRESIDENCY GROUP OVER 40 YEARS OF ACADEMIC WISDOM

If there are more than three features, we would have a hyper-space

A **hyper-space** is a space with higher than 3 dimensions like 4D, 5D etc., and a separating line in a dimension higher than 3, is called a **hyper-plane**.

- If the hyper-planes are linear, the SVM is called **Linear Kernel SVM**
- For nonlinear hyper-planes, a **Polynomial Kernel** or other advanced SVMs are used

What is a Prediction Model

- **Predictive models** are used in many situations where **an estimate** for forecast is required.
- Ex: To project sales or forecast the weather.

- **A Predictive model** will calculate an estimate for one or more variables (responses), based on other variables (descriptors).
- Ex: A dataset of cars is used to build a predictive model to estimate car fuel efficiency (MPG).

PRESIDENCY UNIVERSITY
GAIN MORE KNOWLEDGE REACH GREATER HEIGHTS
Private University Estd. in Karnataka State by Act No. 41 of 2013
PRESIDENCY GROUP OVER 40 YEARS OF ACADEMIC WISDOM

A portion of the observations are shown in the below table.

Table of cars with known MPG values

Names	Cylinders	Displacement	Horsepower	Weight	Acceleration	MPG
Chevrolet Chevelle Malibu	8	307	130	3,504	12	18
Buick Skylark 320	8	350	165	3,693	11.5	15
Plymouth Satellite	8	318	150	3,436	11	18
AMC Rebel SST	8	304	150	3,433	12	16
Ford Torino	8	302	140	3,449	10.5	17
Ford Galaxie 500	8	429	198	4,341	10	15
Chevrolet Impala	8	454	220	4,354	9	14
Plymouth Fury III	8	440	215	4,312	8.5	14
Pontiac Catalina	8	455	225	4,425	10	14
AMC Ambassador DPL	8	390	190	3,850	8.5	15

- A model to predict the **car fuel efficiency** was built using:
  - The **MPG** variable as the **response** and,
  - The **Cylinders, Displacement, Horsepower, Weight** and **Acceleration** variables as **descriptors.**

- Once the model has been built, it can be used to make predictions for **car fuel efficiency.**

Ex: The observations in the below table could be presented to the model & the model would predict the MPG column.

Table 7.2. Table of cars where MPG is to be predicted

Names	Cylinders	Displacement	Horsepower	Weight	Acceleration	MPG
Dodge Challenger SE	8	383	170	3,563	10	
Plymouth Cuda 340	8	340	160	3,609	8	
Chevrolet Monte Carlo	8	400	150	3,761	9.5	
Buick Estate Wagon (SW)	8	455	225	3,086	10	
Toyota Corona Mark II	4	113	95	2,372	15	
Plymouth Duster	6	198	95	2,833	15.5	
AMC Hornet	6	199	97	2,774	15.5	
Ford Maverick	6	200	85	2,587	16	
Datsun Pl510	4	97	88	2,130	14.5	
Volkswagen 1131 Deluxe Sedan	4	97	46	1,835	20.5	

- There are many methods for building prediction models and they are often characterized based on the response variable.
- When the response is a categorical variable, the model is called a **classification model.**
- When the response is a continuous variable, then the model is called a **regression model.**

- Below table summarizes some of the methods available:

Table 7.3. Different classification and regression methods

Classification	Regression
Classification trees	Regression trees
k-Nearest Neighbors	k-Nearest Neighbors
Logistic regression	Linear regressions
Naïve Bayes classifiers	Neural networks
Neural networks	Nonlinear regression
Rule-based classifiers	Partial least squares
Support vector machines	Support vector machines

- There are two distinct phases, each with a unique set of processes and issues to consider:
  - **Building**
  - **Applying**

Building a Prediction Model

- **Building:**
  - The prediction model is built using existing data called training set.
  - This training set contains examples with values for the descriptor and response variables.
  - The training set is used to determine and qualify the relationships between the input descriptors and the output response variables.
  - This set will be divided into observations used to build the model and assess the quality of any model built.

Building a Prediction Model

A. Preparing the Data set

- It is important to prepare a data set prior to modeling.
- Preparation should include the operations outlined such as characterizing, cleaning, and transforming the data.
- Particular care should be taken to determine whether subsetting the data is needed to simplify the resulting models

Building a Prediction Model

B. Designing a Modelling Experiment:

- Building a prediction model is an experiment.
- It will be necessary to build many models for which you do not necessarily know which model will be the 'best'.
- This experiment should be appropriately designed to ensure an **optimal result.**
- There are three major dimensions that should be explored:

Building a Prediction Model

1. **Different models:**
   - There are many different approaches to building prediction models.
   - A series of alternative models should be explored since all models work well in different situations.
   - The initial list of modeling techniques to be explored can be based on the criteria previously defined as important to the project.

Building a Prediction Model

2. Different descriptor combinations:

- Models that are based on a single descriptor are called simple models, whereas those built using a number of descriptors are called multiple (or multivariate) models.
- Correlation analysis as well as other statistical approaches can be used to identify which descriptor variables appear to be influential.
- A subject matter expert or business analyst may also provide insight into which descriptors would work best within a model.

Building a Prediction Model

3. **Model parameters:**
   - Most predictive models can be optimized by fine tuning different model parameters.
   - Building a series of models with different parameter settings and comparing the quality of each model will allow you to **optimize the model**.
   - For example, when building a neural network model there are a number of settings, which will influence the quality of the models built such as the **number of cycles or the number of hidden layers.**

Building a Prediction Model

- Evaluating the 'best' model depends on the **objective of the modeling process defined** at the start of the project.
- Other issues, for example, the **ability to explain how a prediction was made**, may also be important and should be taken into account when assessing the models generated.
- Wherever possible, when two or more models give comparable results, the simpler model should be selected.

Building a Prediction Model

C. Separating Test and Training Sets:

- The goal of building a predictive model is to generalize the relationship between the input descriptors and the output responses.
- The quality of the model depends on how well the model is able to predict correctly for a given set of input descriptors.
- If the model generalizes the input/output relationships too much, the accuracy of the model will be low. - **Overfitting**

Building a Prediction Model

C. Separating Test and Training Sets:

- If the model does not generalize the relationships enough, then the model will have difficulties making predictions for observations not included in the data set used to build the model.
- Hence, when assessing the quality of the model, it is important to use a data set to build the model, which is different from the data set used to test the accuracy of the model.
- There are a number of ways for achieving this separation of test and training set.

Applying a Prediction Model

- **Applying:**
  - Once a model has been built, a data set with no output response variables can be fed into this model and the model will produce an estimate for this response.
  - A measure that reflects the confidence in this prediction is often calculated along with an explanation of how the value was generated.

Applying a Prediction Model

- Once a model has been built and verified, it can be used to make predictions.
- Along with the presentation of the prediction, there should be some indications of the confidence in this value.

- During the data preparation step of the process, the descriptors and/or the response variables may have been translated to facilitate analysis.
- Once a prediction has been made, the variables should be translated back into their original format prior to presenting the information to the end user.
- For example, the log of the variable Weight was taken in order to create a new variable log(Weight) since the original variable was not normally distributed.
- This variable was used as a response variable in a model.
- Before any results are presented to the end user, the log(Weight) response should be translated back to Weight by taking the inverse of the log and presenting the value using the original weight scale.

Applying a Prediction Model

- When applying these models to new data, some criteria will need to be established as to which model **the observation will be presented to.**
- For example, a series of models predicting house prices in different locations such as **coastal, downtown, and suburbs were built.**
- When applying these models to a new data set, the observations should be applied only to the appropriate model.

Thank you!