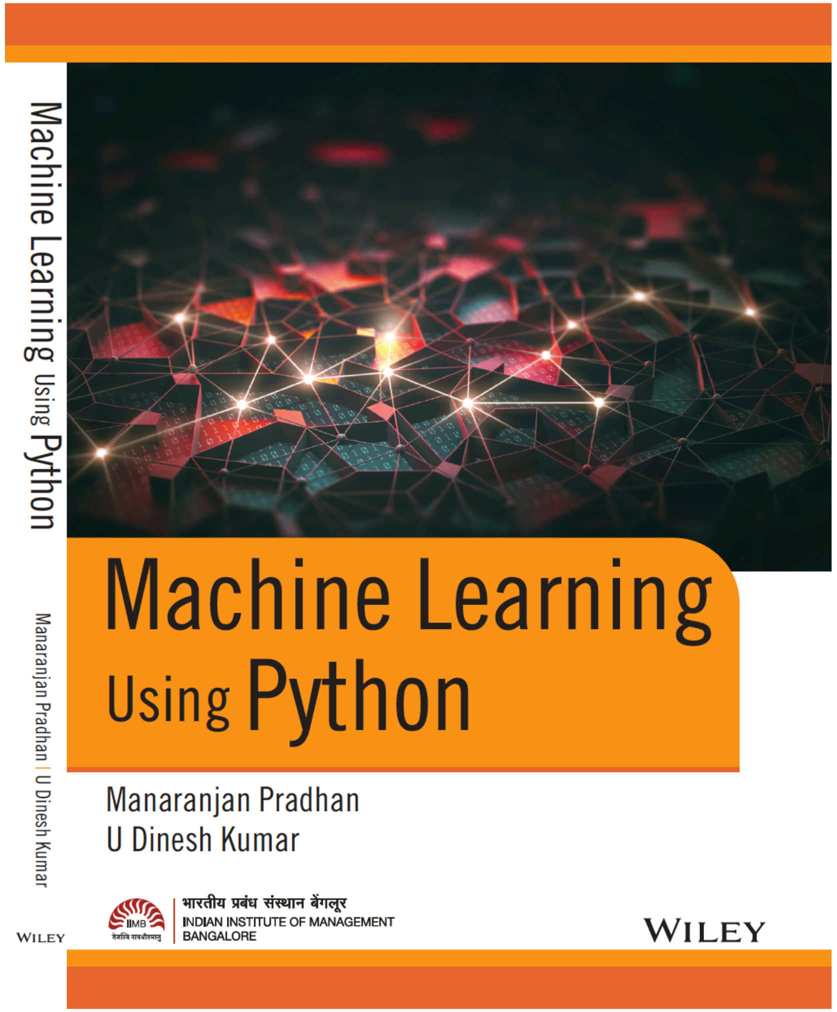


Chapter 06: Advanced Machine Learning

Prepared By: Purvi Tiwari

Learning Objectives

- Understanding the foundations of machine learning algorithms
- Learning the difference between supervised and unsupervised learning algorithms.
- Understanding and developing the gradient descent algorithm.
- Applying machine learning algorithms available in *scikit-learn* to regression and classification problems.
- Understanding the concepts of underfitting – overfitting and use of regularization.
- Understanding ensemble techniques such as Random Forest, Bagging and Boosting.
- Learning feature selection using machine learning models.

Overview

- Machine learning (ML) algorithms are a subset of artificial intelligence (AI) that imitates human learning process.
- ML algorithms develop multiple models and each model is analogues to an experience.
- In ML algorithms, several models are developed which can run into several hundred and each data and model is treated as learning opportunity.
- According to Mitchell (2006)

  Machine learns with respect to a particular task T, performance metric P follows experience E, if the system reliably improves its performance P at task T following experience E.

Overview (Cntd.)

- Slearning depends heavily on validation of model assumption and hypothesis testing, whereas the objective of machine learning is to improve prediction accuracy.
- Two types of ML algorithms
    1. Supervised Learning – the datasets have the values of features and the corresponding outcome variable. Example – Linear regression and logistic regression.
    2. Unsupervised learning – the datasets will have only feature values, but not the outcome variables. The algorithm learns the structure in the features. Example – Clustering and factor analysis

How Machines Learn

- In supervised learning, the algorithm learns using a function called loss function, cost function or error function.
- It is a function of predicted output and the desired output.

The image part with relationship ID rId2 was not found in the file.

- h(X) is the predicted output and y is the desired output, and n is the total number of recorded for which the predictions are made.
- The objective is to learn the values of parameters that minimizes the cost function.

Gradient Descent Algorithm

- Most widely used optimization technique in ML.
- The functional form of a simple linear regression model

The image part with relationship ID rId3 was not

- Where β_0 is called bias, β_1 is the feature weight, ε_i is the error in prediction.
- The predicted value of Y_i is written as $\hat{Y}_i$ and is given by

The image part with relationship ID rId3 was not found in the file.

- Where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimated values of β_0 and β_1

Gradient Descent Algorithm (Cntd.)

- The error is given by

The image part with relationship ID rId3 was not found in the file.

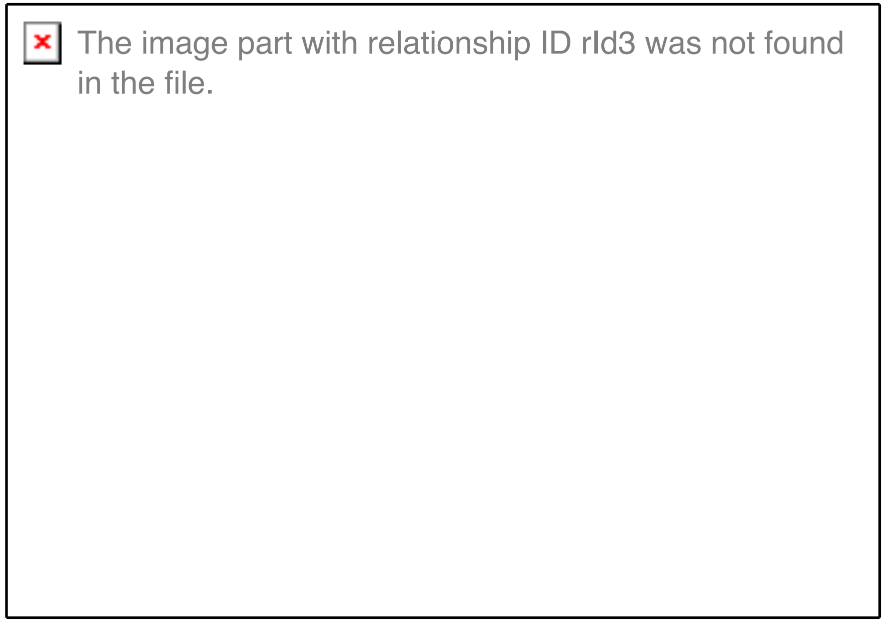
- The cost function for the linear regression model is the total error across all N records and given by

The image part with relationship ID rId3 was not found in the file.

- Error is a function of β_0 and β_1

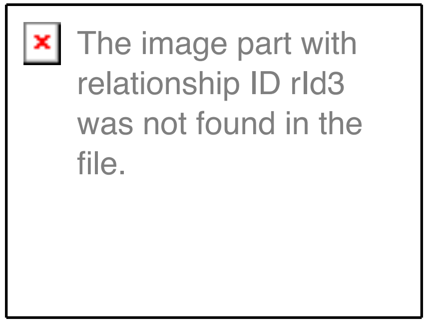
Gradient Descent Algorithm (Cntd.)

- Error is a pure convex function and has a global minimum.
- The gradient descent algorithm starts at a random point and moves toward the optimal solution.



Gradient Descent Algorithm (Cntd.)

- Steps for finding the optimal values of β_0 and β_1
  1. Randomly guess the initial value of β_0 and β_1.
  2. Calculate the estimated value of the outcome variable $\hat{Y}_i$ for initialized values of bias and weights.
  3. Calculate the mean squared error function (MSE).
  4. Adjust the β_0 and β_1 values by calculating the gradients of the error function



Where $\propto$ is the learning rate and the magnitude of the update is applied to bias and weights at each iteration.

Gradient Descent Algorithm (Cntd.)

- The partial derivatives of MSE with respect to β_0 and β_1

The image part with relationship ID rId3 was not found in the file.

5. Repeat steps 1 to 4 for several iterations until the error stops reducing further or the change in cost is infinitesimally small.

- The values of β_0 and β_1 at the minimal cost points are best estimates of the model parameters.

Developing Gradient Descent Algorithm

- For Linear Regression Model
- Dataset – *Advertising.csv*
- The dataset has the following elements:
  1. TV – Spend on TV advertisements
  2. Radio – Spend on radio advertisements
  3. Newspaper – Spend on newspaper advertisements
  4. Sales – Sales revenue generated

Developing Gradient Descent Algorithm (Cntd.)

- Loading the dataset

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Developing Gradient Descent Algorithm (Cntd.)

- Setting X (Features) and Y (Outcome) variables

The image part with relationship ID rId2 was not found in the file.

- Standardize X an Y

The image part with relationship ID rId2 was not found in the file.

Developing Gradient Descent Algorithm (Cntd.)

- Implementing the Gradient Descent Algorithm
  1. **Method 1:** to randomly initialize the bias and weights.
  2. **Method 2:** to calculate the predicted value of Y, that is, Y given the bias and weights.
  3. **Method 3:** to calculate the cost function from predicted and actual values of Y.
  4. **Method 4:** to calculate the gradients and adjust the bias and weights.

Developing Gradient Descent Algorithm (Cntd.)

Method 1: Random Initialization of the Bias and Weights

- The method randomly initialize the bias and weights.

The image part with relationship ID rId2 was not found in the file.

Developing Gradient Descent Algorithm (Cntd.)

- Initializing the parameters

The image part with relationship ID rId2 was not found in the file.

Developing Gradient Descent Algorithm (Cntd.)

Method 2: Predict Y values from the Bias and Weights

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Developing Gradient Descent Algorithm (Cntd.)

Method 3: Calculate the Cost Function - MSE

- Computing mean squared error (MSE) by

1. Calculating differences between the estimated and actual Y.
2. Calculating the square of the above residuals, and sum over all records.
3. Dividing it with number of observations.

Developing Gradient Descent Algorithm (Cntd.)

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Developing Gradient Descent Algorithm (Cntd.)

Method 4: Update the Bias and Weights

- Most important method, where the bias and weights are adjusted based on the gradient of cost function.
- The bias and weights are updates using the following gradients

The image part with relationship ID rId3 was not found in the file.

- Where ∝ is the learning parameter that decides the magnitude of the update to be done to the bias and weights.

Developing Gradient Descent Algorithm (Cntd.)

- The parameters passed to the function are:
  1. x, y: the input and output variables
  2. y_hat: predicted value with current bias and weights
  3. b_0, w_0: current bias and weights
  4. learning rate: learning rate to adjust the update step

The image part with relationship ID rId2 was not found in the file.

Developing Gradient Descent Algorithm (Cntd.)

- Updating bias and weights once after initializing.

The image part with relationship ID rId2 was not found in the file.

Developing Gradient Descent Algorithm (Cntd.)

- Finding the Optimal Bias and Weights
  - The updates to the bias and weights need to be done iteratively, until the cost is minimum.
  - There are two approaches to stop the iterations:

1. Run a fixed number of iterations and use the bias and wrights as optimal values at the end these iterations.
2. Run iterations until the change in cost is small, that is, less than a predefined value.

Developing Gradient Descent Algorithm (Cntd.)

The image part with relationship ID rId2 was not found in the file.

Developing Gradient Descent Algorithm (Cntd.)

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Developing Gradient Descent Algorithm (Cntd.)

- Plotting the cost Function against the Iterations

The image part with relationship ID rId2 was not found in the file.

Developing Gradient Descent Algorithm (Cntd.)

- The cost is still reducing and has not reached the minimum point. More iterations can be run to verify if the cost is reaching a minimum point or not.

The image part with relationship ID rId2 was not found in the file.

Developing Gradient Descent Algorithm (Cntd.)

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Developing Gradient Descent Algorithm (Cntd.)

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Scikit-learn Library for Machine Learning

- Open-source Pyhton library for building machine learning models.
- scikit-learn provides a comprehensive set of algorithms for the following kind of problems:
    1. Regression
    2. Classification
    3. Clustering
- It provides an extensive set of methods for data pre-processing and feature selection.

Steps for Building Machine Learning Models

- Steps to be followed for building, validating a machine learning model and measuring its accuracy are as follows:
  1. Identify the features and outcome variable in the dataset.
  2. Split the dataset into training and test sets.
  3. Build the model using training set.
  4. Predict outcome variable using a test set.
  5. Compare the predicted and actual values of the outcome variable in the test set and measure accuracy using measures such as mean absolute percentage error (MAPE) or root mean square error (RMSE).

Steps for Building Machine Learning Models

- Splitting Dataset into Train and Test Datasets

The image part with relationship ID rId2 was not found in the file.

Steps for Building Machine Learning Models

- Building Linear Regression Model with Train Dataset
- Steps for building a model in *sklearn* are
  1. Initialize the model.
  2. Invoke fit() method on the model and pass the input (X) and output (Y) values.
  3. Fit() will run the algorithm and return the final estimated model parameters.

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Steps for Building Machine Learning Models

- After the model is built, the model parameters such as intercept and coefficients can be obtained as follows

The image part with relationship ID rId2 was not found in the file.

Steps for Building Machine Learning Models

- Associating the coefficient values with the variable names

The image part with relationship ID rId2 was not found in the file.

  - The model indicates that for every unit change in TV spending, there is an increase of 0.44 units in sales revenue.
  - The weights are different than what we estimated earlier as we have not used standardized values in this model.

Steps for Building Machine Learning Models

- Making Prediction on Test Set

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Steps for Building Machine Learning Models

- Making Prediction on Test Set

The image part with relationship ID rId2 was not found in the file.

Steps for Building Machine Learning Models

- Measuring Accuracy

The image part with relationship ID rId2 was not found in the file.

- R-Squared Value

The image part with relationship ID rId2 was not found in the file.

Steps for Building Machine Learning Models

- RMSE Calculation

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Steps for Building Machine Learning Models

- Bias-Variance Trade-Off
  - Models errors can be decomposed into two components: bias and variance.
  - Avoid model overfitting or underfitting.
  - High bias can lead to building underfitting model, whereas high variance can lead to overfitting models.
- Understanding with example - dataset *curve.csv*

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Steps for Building Machine Learning Models

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Steps for Building Machine Learning Models

- It can be observed that the relation between y and x is not linear.
- Need to try various polynomial forms of x and verify the model.

The image part with relationship ID rId2 was not found in the file.

Steps for Building Machine Learning Models

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Steps for Building Machine Learning Models

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Steps for Building Machine Learning Models

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Steps for Building Machine Learning Models

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Steps for Building Machine Learning Models

- Following code is used to build models with degrees ranging from 1 to 15
- Storing the degree and error details in different columns of DataFrame names *rmse_df.*
  1. *degree:* Degree of the model.
  2. *rmse_train*: RMSE error on train set.
  3. *rmse_test:* RMSE error on test set.

Steps for Building Machine Learning Models

The image part with relationship ID rId2 was not found in the file.

Steps for Building Machine Learning Models

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Steps for Building Machine Learning Models

The image part with relationship ID rId2 was not found in the file.

Steps for Building Machine Learning Models

The image part with relationship ID rId2 was not found in the file.

Steps for Building Machine Learning Models

The image part with relationship ID rId2 was not found in the file.

- Key Observations

1. Error on the test set are high for the model with complexity of degree 1 and degree 15.
2. Error on the test set reduces initially, however increases after a specific level of complexity.
3. Error on the training set decreases continuously.

Steps for Building Machine Learning Models

- K-Fold Cross-Validation

1. A robust validation approach that can be adopted to verify if the model is overfitting.
2. The model, which generalizes well and does not overfit, should not be very sensitive to any change in underlying training samples.
3. It builds and validate multiple models by resampling multiple training and validation sets from the original dataset.

Steps for Building Machine Learning Models

- The following steps are used in K-fold cross-validation:

1. Split the training dataset into k subsets of equal size. Each subset will be called a fold. Let the folds be labelled as f, f, …, f. Generally, the value of k is taken to be 5 or 10.
2. For i=1 to k
   a) Fold f is used as validation set and all the remaining k-1 folds as training set.
   b) Train the model using the training set and calculate the accuracy of the model in fold f.
3. Calculate the final accuracy by averaging the accuracies in the test data across all k models.

Steps for Building Machine Learning Models

- The average accuracy value shows how the model will behave in the real world.
- The variance of these accuracies is an indication of the robustness of the model.

The image part with relationship ID rId2 was not found in the file.

Advanced Regression Models

- Dataset – IPL dataset
- Linear Regression model – the model will predict SOLD PRICE of a player based on past performance measures of the players.

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Advanced Regression Models (Cntd.)

- Getting the features

The image part with relationship ID rId2 was not found in the file.

- Encoding the categorical variables

The image part with relationship ID rId2 was not found in the file.

Advanced Regression Models (Cntd.)

- Displaying all the feature names along with the new dummy features.

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Advanced Regression Models (Cntd.)

- Creating variables X and Y

The image part with relationship ID rId2 was not found in the file.

- Standardization of X and Y

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Advanced Regression Models (Cntd.)

- Splitting the dataset into Train and Test

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Advanced Regression Models (Cntd.)

- Build the model

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Advanced Regression Models (Cntd.)

- Storing the beta coefficients and respective columns in a DataFrame

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Advanced Regression Models (Cntd.)

- Plotting the Coefficient Values

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

1. AVE, ODI-RUNS-S, SIXERS are top three highly influential features which determine the player’s SOLD PRICE.
2. Higher ECON, SR-B and AGE have negative effect.
3. Interestingly, higher test runs (T-runs) and highest score (HS) have negative effect on the SOLD PRICE.

Note that few of these counter-intuitive sign for coefficients could be due to multi-collinearity. For example – SR-B is expected to have a positive effect on the SOLD PRICE.

Advanced Regression Models (Cntd.)

- Calculate RMSE

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Advanced Regression Models (Cntd.)

- Applying Regularization
  - Regularization deals with overfitting
  - Overfitting is typically caused by inflation of the coefficients.
  - Regularization applies penalties on parameters if they inflate to large values and keeps them from being weighted too heavily.
  - The coefficients are penalized by adding the coefficient terms to the cost function.
  - Optimizer controls the coefficient values to minimize the cost function.
  - Following are the twoo approaches that can be used for adding a penalty to the cost function:

1. L1 Norm
2. L2 Norm

Advanced Regression Models (Cntd.)

1. L1 Norm – Summation of the absolute value of the coefficients, called Least Absolute Shrinkage and Selection Operator (LASSO Term).

The image part with relationship ID rId3 was not found in the file.

Where ∝ is the multiplier term

2. L2 Norm – Summation of the squared value of the coefficients, called Ridge Term.

The image part with relationship ID rId3 was not found in the file.

Advanced Regression Models (Cntd.)

- Effect of LASSO and Ridge constraint applied to the cost function

The image part with relationship ID rId2 was not found in the file.

Advanced Regression Models (Cntd.)

- Ridge Regression
  1. *sklearn.linear_model* provides Ridge regression for building linear models by applying L2 penalty.
  2. Ridge regression takes the following parameters:

  a) *alpha* – float – is the regularization strength and must be positive.
  Larger values of alpha imply stronger regularization

  b) *max_iter* – int – is the maximum umber of iterations for the gradient solver.

The image part with relationship ID rId2 was not found in the file.

Advanced Regression Models (Cntd.)

- Ridge Regression

The image part with relationship ID rId2 was not found in the file.

- The difference in RMSE on train and test has reduced because of the penalty effect. The difference can be further reduced by applying stronger penalty (for example – apply large alpha value as 2.0)

The image part with relationship ID rId2 was not found in the file.

Advanced Regression Models (Cntd.)

- LASSO Regression
  1. *Sklearn.linear_model* provides LASSO regression for building linear models by applying L1 penalty.

  a) *alpha* – float – multiplies the L1 term. Default value is set to 1.0

  b) *max_iter* – int – Maximum number of iterations for gradient solver.

The image part with relationship ID rId2 was not found in the file.

Advanced Regression Models (Cntd.)

- It can be noticed that the model in not overfitting and the difference between train and test RMSE is very small.
- LASSO reduces some of the coefficient values to 0, which indicates that these features are not necessary for explaining the variance in the outcome variable.

The image part with relationship ID rId2 was not found in the file.

Advanced Regression Models (Cntd.)

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

- The LASSO regression indicates that the features listed under “columns” are not influencing factors for predicting the SOLD PRICE as the respective coefficients are 0.0

Advanced Regression Models (Cntd.)

- Elastic Net Regression – it combines both L1 and L2 regularizations to build a regression model.
- The corresponding function is given by

The image part with relationship ID rId3 was not found in the file.

- *ElasticNet* takes following two parameters:
    1. *Alpha:* constant that multiplies the penalty terms. Default is set to 1.0. (alpha = $\sigma + \gamma$), where σ (L2) and γ(L1) are two hyperparameters.

Advanced Regression Models (Cntd.)

2. *l1_ratio:* The ElasticNet mixing parameter, with 0 <= l1_ratio <= 1

Where

l1_ratio = 0 implies that the penalty is an L2 penalty.

l1_ratio = 1 implies that the penalty is an L1 penalty.

l1_ratio < 0 implies that the penalty is a combination of L1 and L2.

The image part with relationship ID rId2 was not found in the file.

Advanced Machine Learning Algorithms

- We will take a binary classification problem and demonstrate it through ML algorithms such as
  1. K-Nearest Neighbors (KNN),
  2. Random Forest, and
  3. Boosting
- Dataset : bank marketing dataset, available at the University of California, Irvine machine learning repository *(*http://archive.ics.uci.edu/ml/datasets/Bank+Marketing*)*

Advanced Machine Learning Algorithms (Cntd.)

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Advanced Machine Learning Algorithms (Cntd.)

The image part with relationship ID rId2 was not found in the file.

Advanced Machine Learning Algorithms (Cntd.)

- Dealing with Imbalanced Datasets
  - A dataset is imbalanced when there is no equal representation of all classes in data.
  - In our dataset the proportion of customers who responded to the telemarketing is approximately 11.5% and the remaining 88.5% did not respond.
- Number of records of customers who has or has not opened the account

The image part with relationship ID rId2 was not found in the file.

Advanced Machine Learning Algorithms (Cntd.)

- Dealing with Imbalanced Datasets
  - A dataset is imbalanced when there is no equal representation of all classes in data.
  - In our dataset the proportion of customers who responded to the telemarketing is approximately 11.5% and the remaining 88.5% did not respond.
- Number of records of customers who has or has not opened the account

The image part with relationship ID rId2 was not found in the file.

Advanced Machine Learning Algorithms (Cntd.)

- Resampling techniques to to deal with imbalanced datasets
  1. Upsampling – Increase the instances of under-represented minority class by replicating the existing observations in the dataset. It is also called oversampling.
  2. Downsampling – Reduce the instances of over-represented majority class by removing the existing observations from the dataset and is also called undersampling.
- *Sklearn.utils* has resample method to help with upsampling. It takes three parameters:
  1. The original sample set
  2. *replace*: implements resampling with replacements. If false, all resampled examples will be unique.
  3. *n_samples:* number of samples to generate.

Advanced Machine Learning Algorithms (Cntd.)

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Advanced Machine Learning Algorithms (Cntd.)

- After upsampling, the case of subscribed and unsubscribes customers is 67:33
- Before using the dataset, the examples can be shuffles to make sure they are not in particular order.

The image part with relationship ID rId2 was not found in the file.

Advanced Machine Learning Algorithms (Cntd.)

The image part with relationship ID rId2 was not found in the file.

Advanced Machine Learning Algorithms (Cntd.)

- encoding all the categorical features into dummy features and assign to X.

The image part with relationship ID rId2 was not found in the file.

Advanced Machine Learning Algorithms (Cntd.)

- Splitting into train and test data

The image part with relationship ID rId2 was not found in the file.

Advanced Machine Learning Algorithms (Cntd.)

- Building Logistic Regression Model

1. Logistics regression is a classification model.
2. The cost function is called log loss (log likelihood) or binary cross-entropy function and is given by

The image part with relationship ID rId2 was not found in the file.

Advanced Machine Learning Algorithms (Cntd.)

- Building Logistic Regression Model

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Advanced Machine Learning Algorithms (Cntd.)

- Confusion matrix

The image part with relationship ID rId2 was not found in the file.

Advanced Machine Learning Algorithms (Cntd.)

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Advanced Machine Learning Algorithms (Cntd.)

- Classification Report – the *classification_report* function in *sklearn.metrics* gives a detailed report of precision, recall and F1-score for each class.

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Advanced Machine Learning Algorithms (Cntd.)

- ROC and AUC Score

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Advanced Machine Learning Algorithms (Cntd.)

- Creating DataFrame *test_results_df* to store the actual labels and predicted probabilities for class label 1.

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Advanced Machine Learning Algorithms (Cntd.)

- ROC AUC score can be obtained using *metrice.roc_auc_score().*

The image part with relationship ID rId2 was not found in the file.

Advanced Machine Learning Algorithms (Cntd.)

- Plotting ROC Curve

The image part with relationship ID rId2 was not found in the file.

Advanced Machine Learning Algorithms (Cntd.)

- Plotting ROC Curve (Cntd.)

The image part with relationship ID rId2 was not found in the file.

Advanced Machine Learning Algorithms (Cntd.)

- Plotting ROC Curve (Cntd.)

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Advanced Machine Learning Algorithms (Cntd.)

- Plotting ROC Curve (Cntd.)

The image part with relationship ID rId2 was not found in the file.

Advanced Machine Learning Algorithms (Cntd.)

- K-nearest Neighbors (KNN) Algorithm

1. A non-parametric, lazy learning algorithm used for regression and classification problems.
2. ML algorithms are of two types: parametric and non-parametric
   - Parametric models estimate a fixed number of parameters from the data and strong assumptions of the data. The data is assumed to be following a specific probability distribution. Logistic regression is an example of a parametric model.
   - Non-parametric models do not make any assumptions on the underlying data distribution (such as normal distribution). KNN memorizes the data classifies new observations by comparing the training data.

Advanced Machine Learning Algorithms (Cntd.)

- KNN algorithm finds observations in the training set, which are similar to the new observation. These observations are called neighbors.
- Fore better accuracy, a set of neighbors (K) can be considered for classifying a new observation.
- The class for the new observation can be predicted to be same class that majority of the neighbors belong to.

Advanced Machine Learning Algorithms (Cntd.)

The image part with relationship ID rId2 was not found in the file.

Advanced Machine Learning Algorithms (Cntd.)

- The neighbors are founded by computing distance between observations. Euclidean distance is one of the widely used metrics.

The image part with relationship ID rId2 was not found in the file.

- Where O and O are two observations in the data. X, X are the values of feature for records 1 and 2, respectively, X and X are the values of feature X for records 1 and 2, respectively.
- Other distance measures are Minkowski distance, Jaccard Coefficient and Gower's distance.

Advanced Machine Learning Algorithms (Cntd.)

- *sklearn.neighbors* provides KNeighborsClassifier algorithm for classification problems. It takes the following parameters:
  1. N_neighbors : int – Number of neighbors to use by default. Default is 5.
  2. Metric: string – the distance metrics. Default ‘Minkowski’.
  3. Weights: str – Default is uniform where all points in each neighborhood are weighted equally. Else the distance which weighs points by the inverse of their distance.

The image part with relationship ID rId2 was not found in the file.

Advanced Machine Learning Algorithms (Cntd.)

- KNN Accuracy

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Advanced Machine Learning Algorithms (Cntd.)

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Advanced Machine Learning Algorithms (Cntd.)

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

- The recall of positive cases has improved from 0.25 to 0.75 in the KNN model.
- K in KNN Is called hyperparameters and the process of finding optimal value for a hyperparameter is called hyperparameter tuning.

Advanced Machine Learning Algorithms (Cntd.)

- GridSearch for Most Optimal Parameters
  - *sklearn.model_selection* provides a feature called GridSearch_CV, which searches through a set of given hyperparameter values and reports the most optimal one.
  - It does k-fold cross-validation for each value of hyperparameter to measure accuracy and avoid overfitting
  - Can be used for any machine learning algorithm to search for. Optimal values for its hyperparameters.
- GridSearchCV takes the following parameters:
  - Estimator – scikit-learn model, which implements estimator interface.
  - Param_grid – a distionary with parameter names as keys and lists of parameter values to search for
  - Scoring – the accuracy measure.
  - Cv – integer – the number of folds in K-fold.

Advanced Machine Learning Algorithms (Cntd.)

The image part with relationship ID rId2 was not found in the file.

Advanced Machine Learning Algorithms (Cntd.)

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Advanced Machine Learning Algorithms (Cntd.)

The image part with relationship ID rId2 was not found in the file.

The image part with relationship ID rId2 was not found in the file.

Ensemble Methods

- Learning algorithms that take a set of estimators or classifiers and classify new data points using strategy such as majority vote.
- Also used for regression problems, where the prediction of new data is simple average or weighted average of all the predictions from the set of regression models.
- Multiple datasets are needed for building multiple classifiers.
- In practice, strategy such as bootstrapped samples are drawn from the initial training set and given to each classifier.

Ensemble Methods

- Sometimes bootstrapping involves sampling features along with sampling observations.
- Each resampled set contains a subset of features available in the original set.
- Sampling features help to find important features.
- Records that are not part of specific sample are used for testing the model accuracy. Such records are called Out-of-Bag records
- Process of bootstrapping samples from original set to build multiple models and aggregating the results for final prediction is called Bagging.
- The most widely used bagging technique is Random Forest.