MATH5110

APPLIED LINEAR ALGEBRA AND MATRIX ANALYSIS

PRUDENTIAL LIFE INSURANCE

Abhilasha Jain
Ailin Dolson-Fazio

INDEX

INDEX

DATA DES

DATA DESCRIPTION

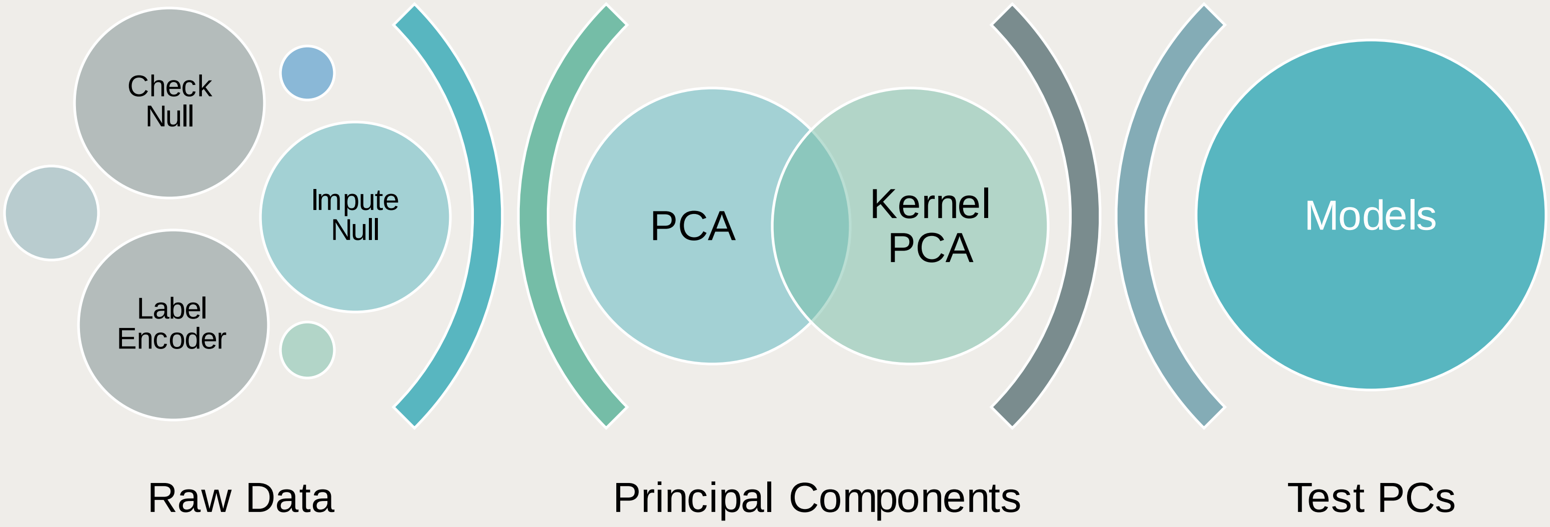
Variable	Description
Id	A unique identifier associated with an application.
Product_Info_1-7	A set of normalized variables relating to the product applied for
Ins_Age	Normalized age of applicant
Ht	Normalized height of applicant
Wt	Normalized weight of applicant
BMI	Normalized BMI of applicant
Employment_Info_1-6	A set of normalized variables relating to the employment history of the applicant.
InsuredInfo_1-6	A set of normalized variables providing information about the applicant.
Insurance_History_1-9	A set of normalized variables relating to the insurance history of the applicant.
Family_Hist_1-5	A set of normalized variables relating to the family history of the applicant.
Medical_History_1-41	A set of normalized variables relating to the medical history of the applicant.
Medical_Keyword_1-48	A set of dummy variables relating to the presence of/absence of a medical keyword being associated with the application.
Response	This is the target variable, an ordinal variable relating to the final decision associated with an application

PROBLEM

PROBLEM

- Predict customer risk based on the input parameters.
- Risk can be in range 1-8.
- 1 is lowest and 8 is the highest risk.

PREPROCESSING DATA



ML MODELS

MODELS USED

CLASSIFICATION	REGRESSION
LOGISTIC REGRESSION	SGD REGRESSOR
SUPPORT VECTOR MACHINES	LINEAR REGRESSION
RANDOM FOREST	XGB CLASSIFIER

LOGISTIC REG

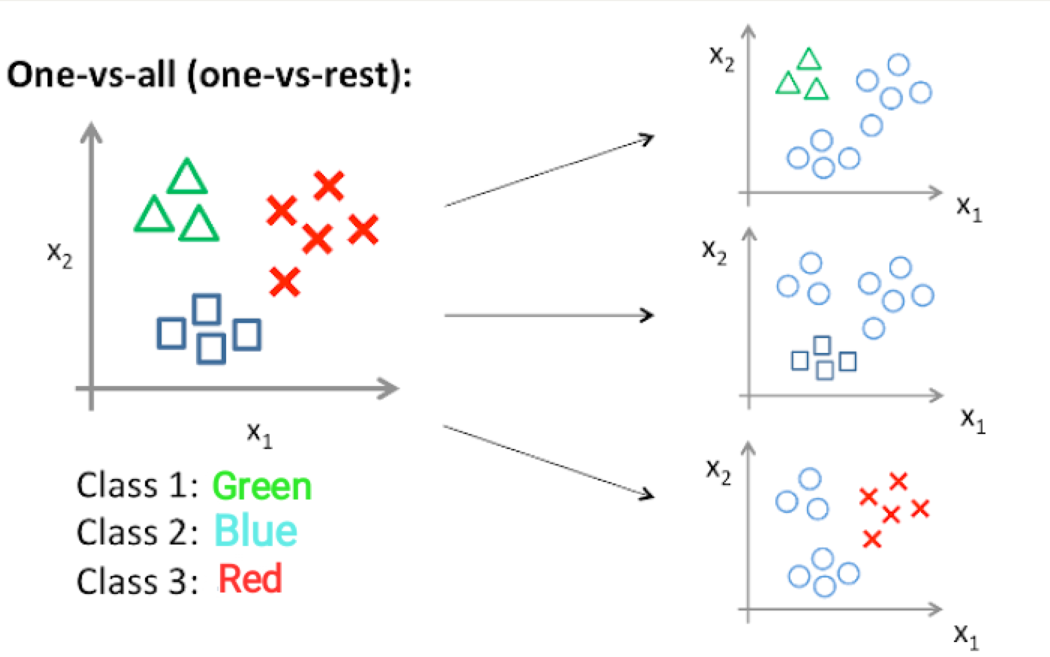
LOGISTIC REGRESSION

Method for using binary classification algorithms for multi-class classification

Divided the problem into smaller binary classifications.

One vs. All:- N-class instances then N binary classifier models

One vs. One:- N-class instances then N* (N-1)/2 binary classifier models



LINEAR REG

LINEAR REGRESSION

The method aims at finding the best fit line for predicting dependent variable (y) based on the independent variables (x).

Hypothesis function for linear Regression:

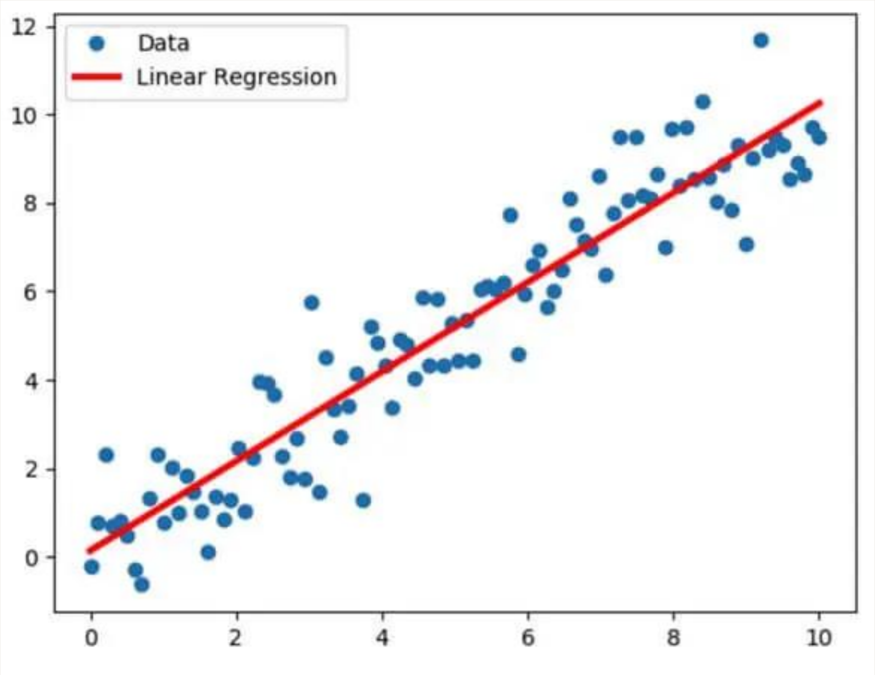
$$y = \theta_1 + \theta_2 * x$$

y = label to predict (dependent variable)

x = input training data (independent variable)

θ_1= intercept of line

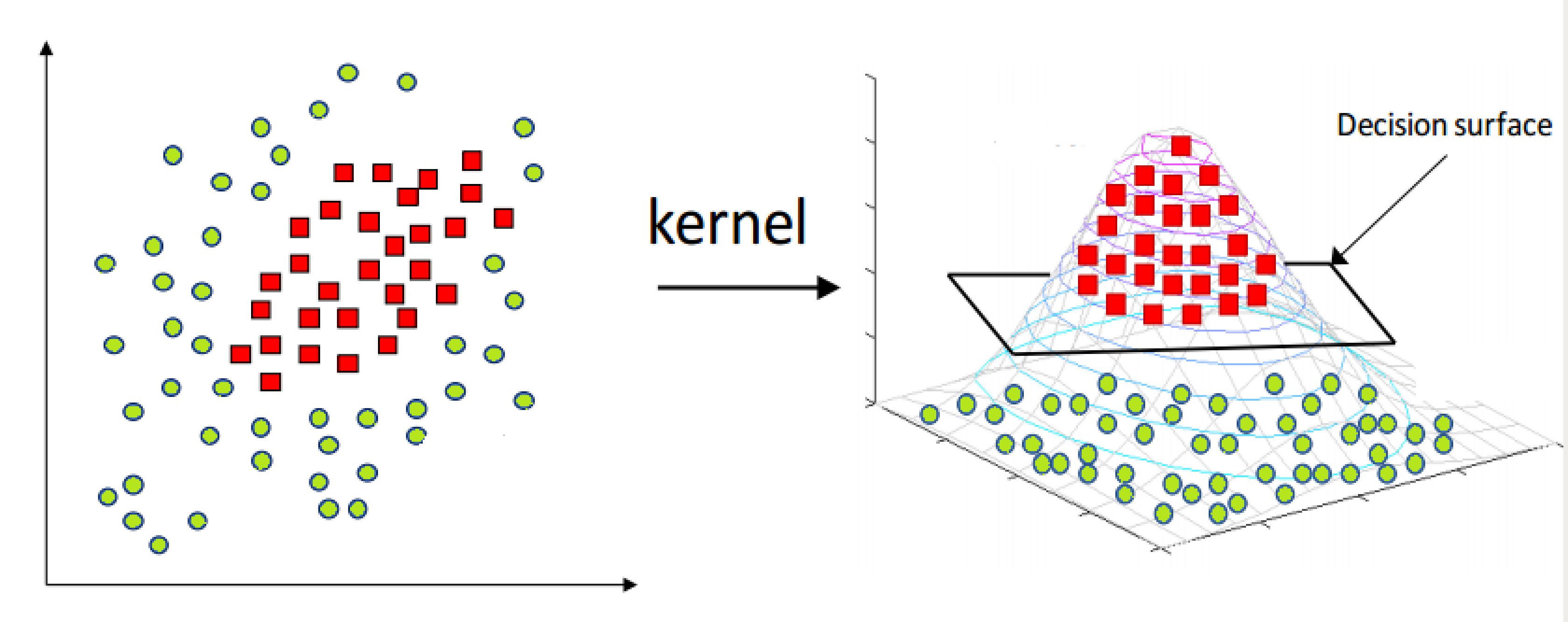
θ_2= co-efficient of x.



SVMs

SUPPORT VECTOR MACHINES

- SVC is capable of performing multiclass classification.
- For Multiclass classification SVC implements “One-versus-one” approach.
- This implies, number of classifiers constructed are: (n_classes - 1) / 2
- SVM utilizes kernels to raise data to higher dimensions for classification.



SGD REGRESS

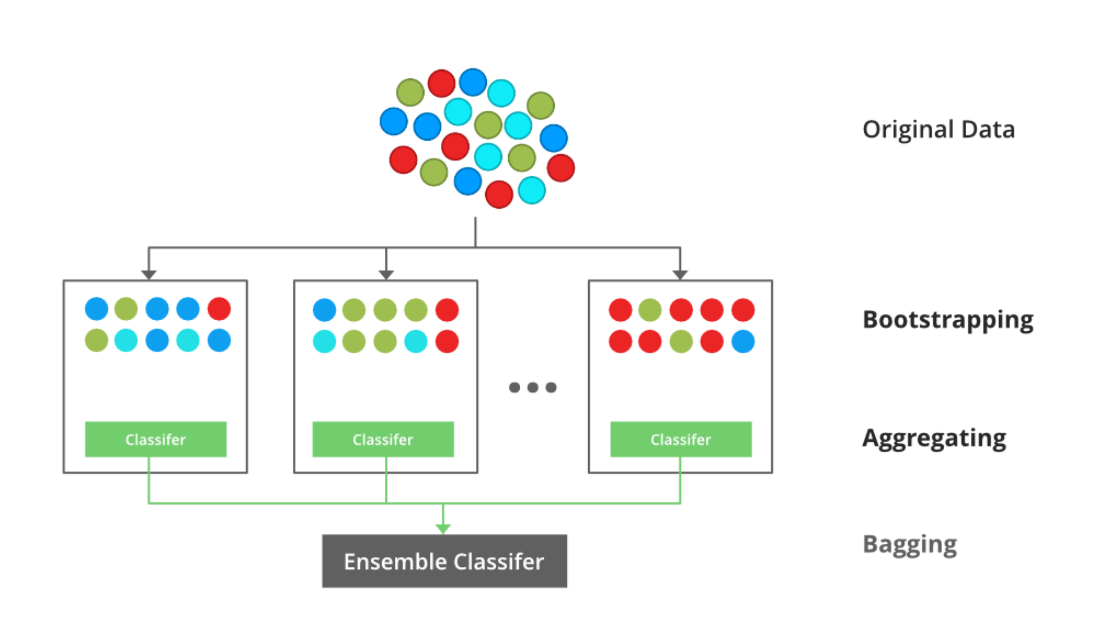
SGD REGRESSION

- The general idea is to start with a random point and find a way to update this point with each iteration such that we descend the slope.
- The steps of the algorithm are
  - Find the slope of the objective function with respect to each parameter/feature. In other words, compute the gradient of the function.
  - Pick a random initial value for the parameters. (To clarify, in the parabola example, differentiate "y" with respect to "x". If we had more features like x1, x2 etc., we take the partial derivative of "y" with respect to each of the features.)
  - Update the gradient function by plugging in the parameter values.
  - Calculate the step sizes for each feature as : step size = gradient * learning rate.
  - Calculate the new parameters as : new params = old params -step size
  - Repeat steps 3 to 5 until gradient is almost 0.
- The "learning rate" mentioned above is a flexible parameter which heavily influences the convergence of the algorithm. Larger learning rates make the algorithm take huge steps down the slope and it might jump across the minimum point thereby missing it. So, it is always good to stick to low learning rate such as 0.01

XGBREGRESS

XGB REGRESSION

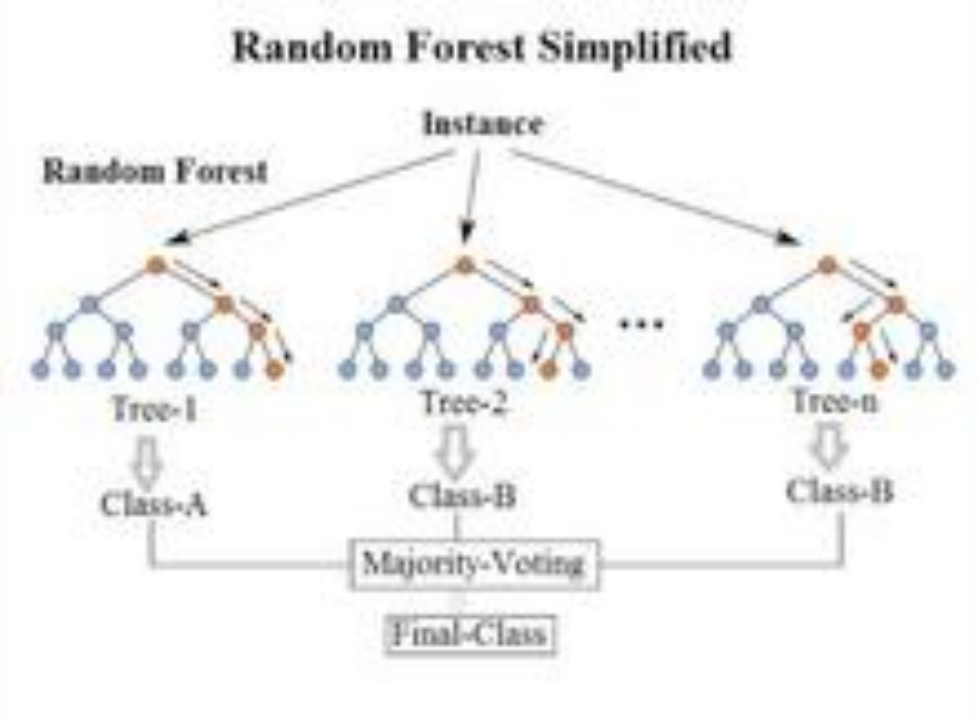
- XGBoost is a short of eXtreme Gradient Boosting
- Boosting is an ensemble method
  - Each tree boosts attributes that led to mis-classifications of previous tree
- Tree Pruning
  - Generally, results in deeper, but optimized trees
- Easy to use:
  - Easy to install
  - Highly developed R/python interfaces for users
- Efficiency
  - Parallel computation
  - Can handle missing values automatically
- Accuracy
  - Good results for most datasets



RANDOMFORES

RANDOM FOREST

- Used to solve regression of classification problems
- Algorithm consists of "decision trees"
- Each tree is data sample from training set
- Combines the output of multiple trees to come to one decision



RESULTS

Best Performing Model

Classifier:

XG Boost Classifier
Random Forest
Logistic Regression

Model	PCA	Kernel PCA
Logistic Regression	99.8	94.43
Linear Regression	94.83	94.33
XGB Classifier	99.95	94.43
SGD Classifier	58.05	94.43
Random Forest	99.93	94.00
K Neighbours Classifier	94.25	94.43
SVMs	94.25	94.35
ANN	94.25	94.43

CONCLUSION

CONCLUSION

- Reduced 50% of dimension and trained with 50-50 split on train and test set.
- Data is Linear because it performs better with PCA
- Perform best with XG Boost Classifier
- Data is better fitted for Ensemble and boosting model

REFERENCES

REFERENCES

- https://towardsdatascience.com/multi-class-classification-one-vs-all-one-vs-one-94daed32a87b
- https://www.wallstreetmojo.com/linear-regression/
- https://scikit-learn.org/stable/modules/svm.html
- https://medium.com/@zxr.nju/what-is-the-kernel-trick-why-is-it-important-98a98db0961d
- https://scikit-learn.org/stable/modules/sgd.html
- https://www.geeksforgeeks.org/xgboost/
- Principal Component Regression — Clearly Explained and Implemented | by Kenneth Leung | Towards Data Science
- https://www.kaggle.com/competitions/prudential-life-insurance-assessment/data
- Bacteremia detection from complete blood count and differential leukocyte count with machine learning: complementary and competitive with C-reactive protein and procalcitonin tests | BMC Infectious Diseases | Full Text (biomedcentral.com)
- https://people.eecs.berkeley.edu/~wainwrig/stat241b/scholkopf_kernel.pdf
- http://localhost:8888/notebooks/Desktop/Kernal%20Paper.ipynb
- Sec_15PCA-1.pdf (He Wang Slides)

TEAM

ABHILASHA JAIN

AILIN DOLSON-FAZIO