

Artificial Intelligence Application in Algorithmic Trading

Arvin Sahni
Tianyi Zhang
Avirath Kakkar

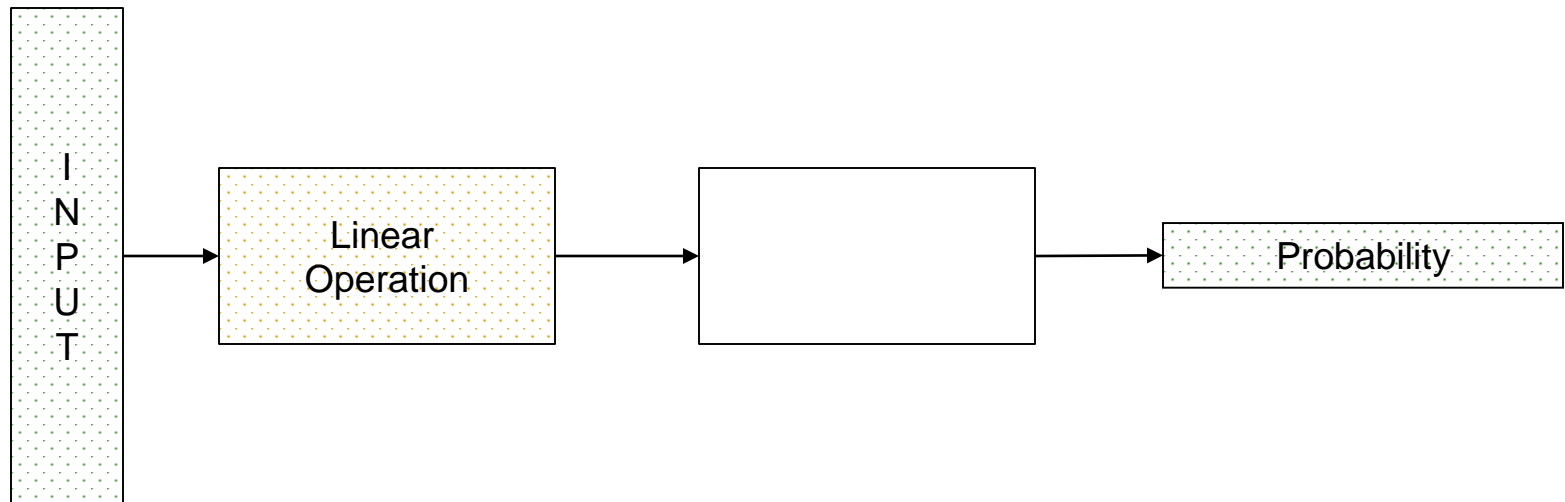
Introduction

- Supervised Learning
 - Classification
 - Regression
 - Unsupervised Learning
 - Reinforcement Learning
-

Introduction

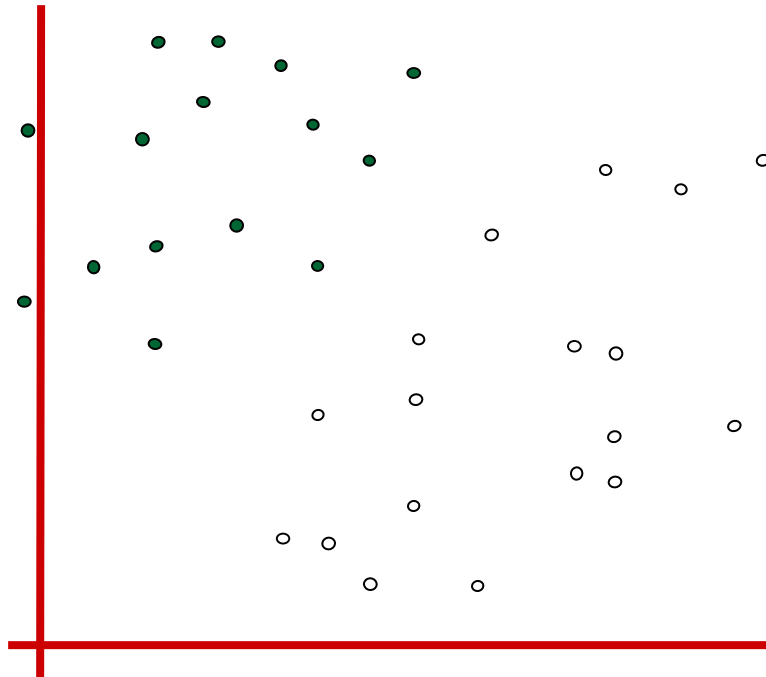
- Supervised Learning Algorithms
 - Logistic Regression
 - Decision Trees
 - Support Vector Machines
 - Neural Networks
 - Convolutional Neural Networks
 - Recurrent Neural Networks
-

Logistic Regression

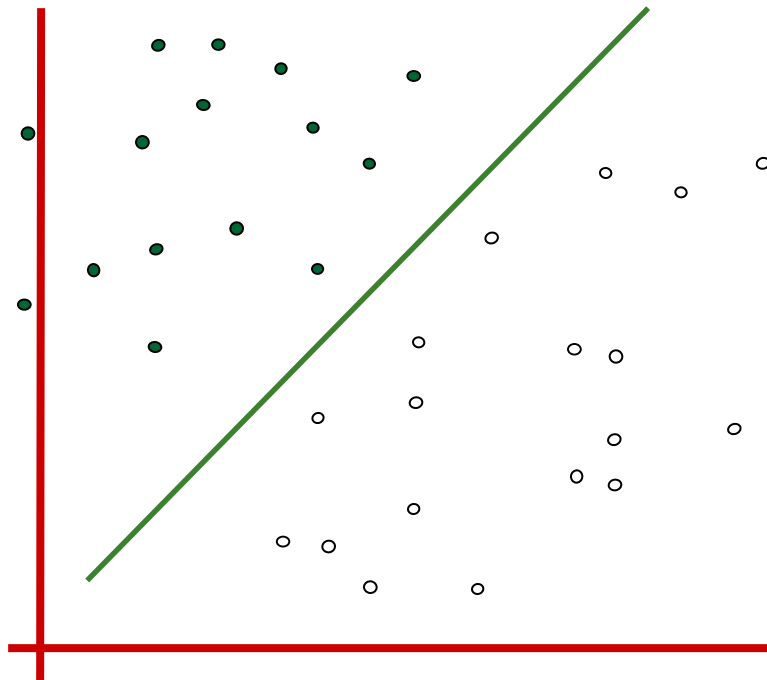


Support Vector Machines

Draw a straight Line to separate the filled dots from the empty dots

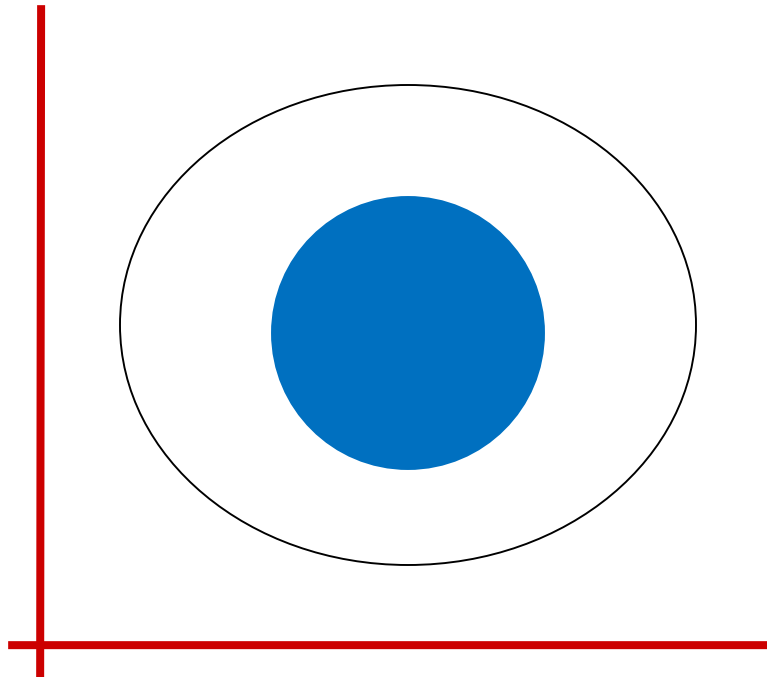


Support Vector Machines



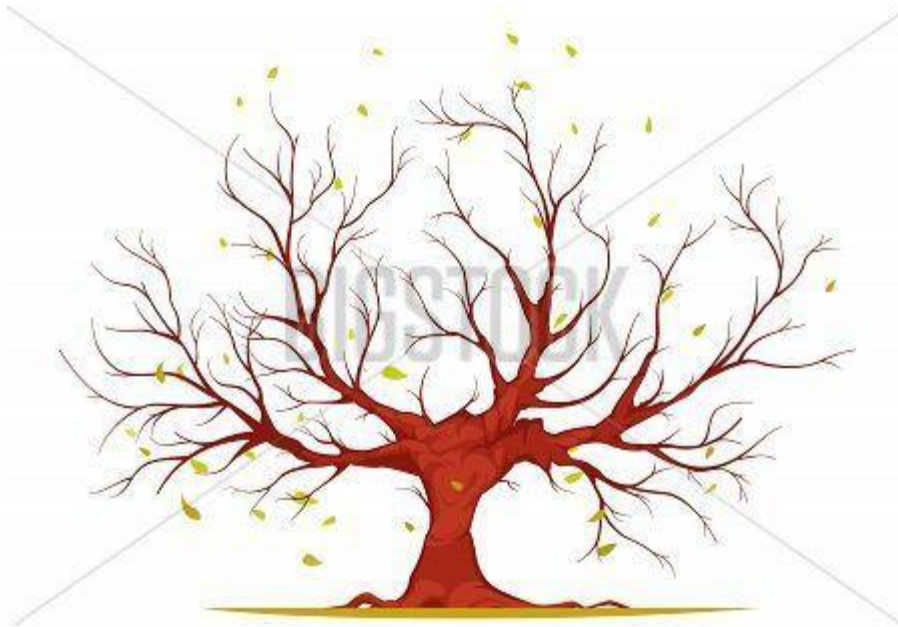
Support Vector Machines

Draw a straight Line to separate the two regions!!!!!!!

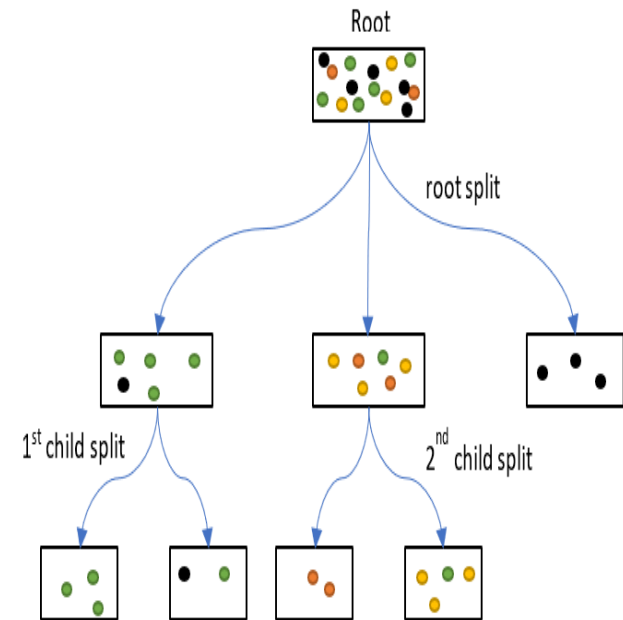


Decision Trees

Defy Gravity

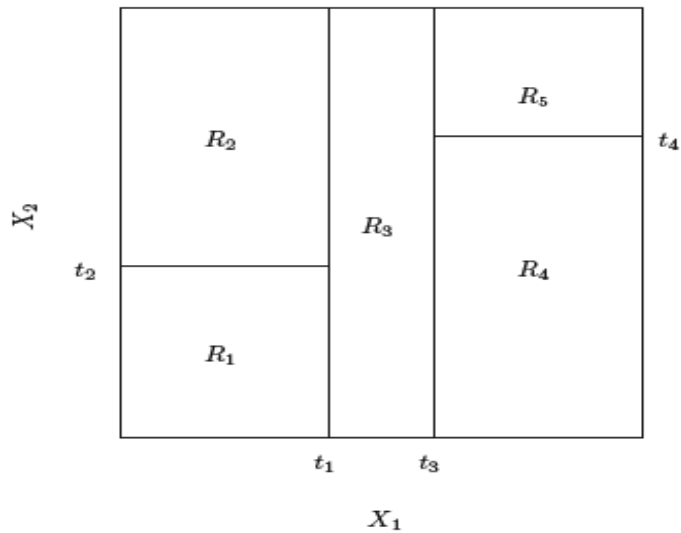


www.bigstock.com · 237498532



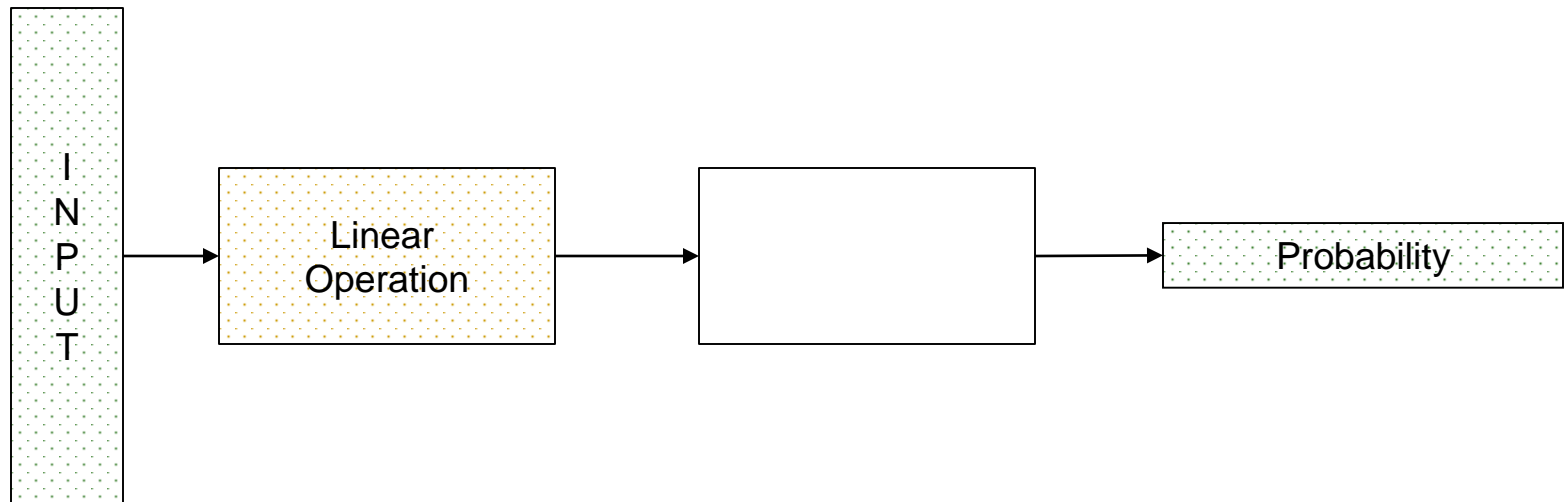
Decision Trees

Divide the input space into rectangular regions



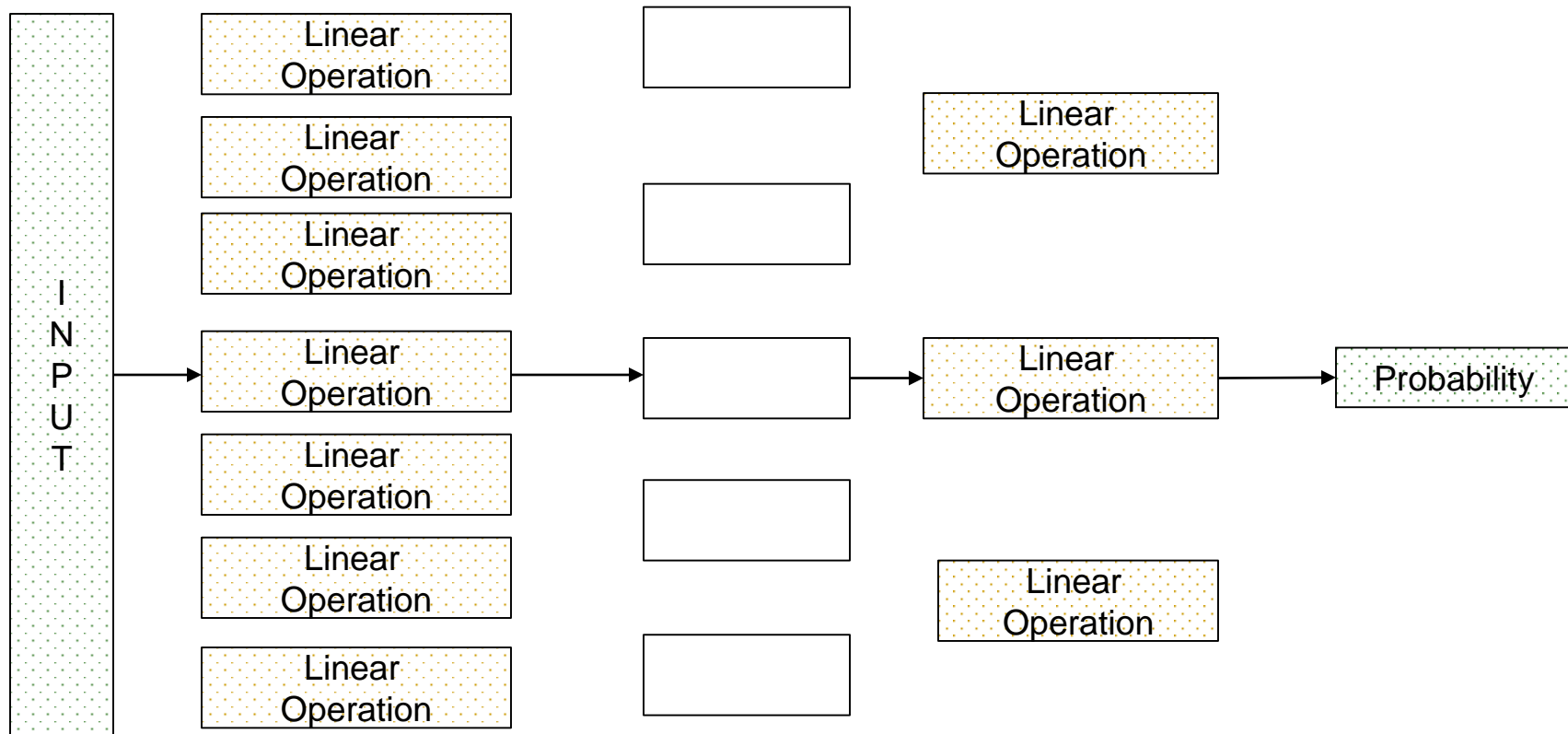
Neural Network

Similar to Logistic Regression, But....



Neural Network

A lot more

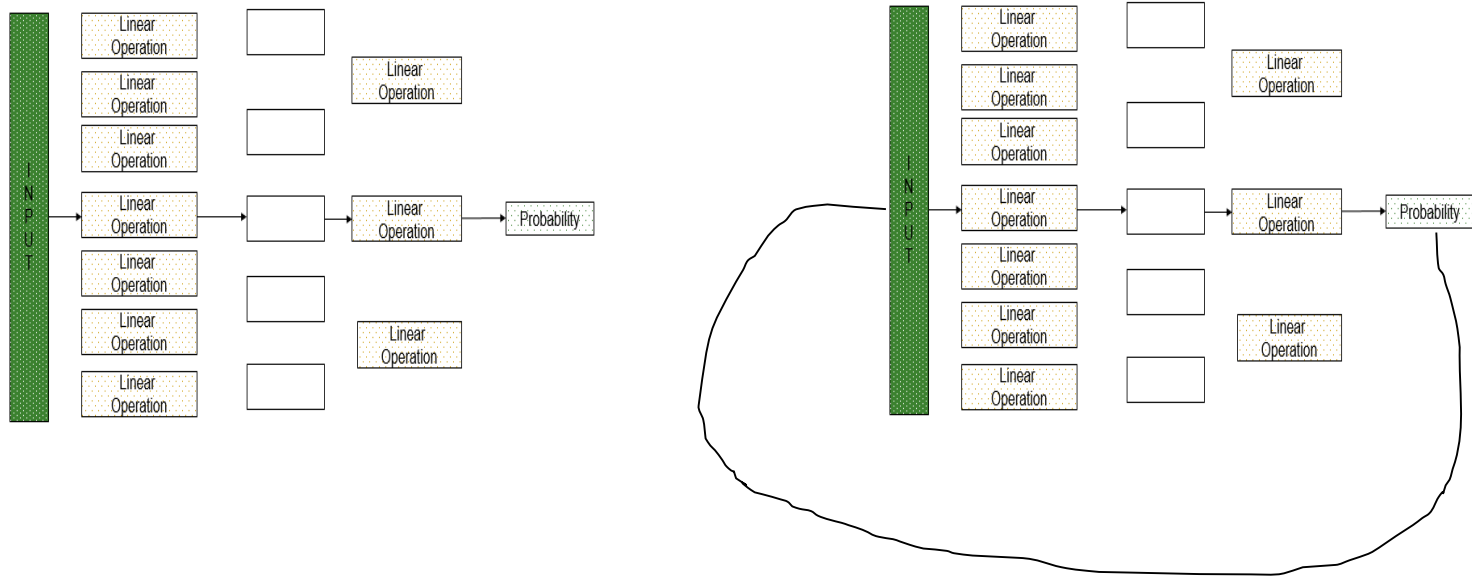


Convolutional Neural Networks

- Ever wondered which moving average to use?
 - 20 dma , 200 dma , exponential moving average
 - Wavelets?

Recurrent Neural Networks

- How Big can the input layer really be
- Obvious impact on efficiency and cost issues of the neural network
- Can yesterdays prediction be used to mitigate some of the above issues?



Typical Dataset

FEATURES

S
A
M
P
L
E
S

A	B	C	D	E	F	G	H	I	J	K	L	
Date	Open	High	Low	Close	R3	R2	R1	Pivot	S1	S2	S3	s
8/4/2014	1295.4	1296.4	1287	1289.2	1326.3	1308.9	1302.1	1291.5	1284.7	1274.1	1256.7	
8/5/2014	1289.2	1295	1283.3	1289.6	1309.7	1300.3	1294.7	1290.9	1285.3	1281.5	1272.1	
8/6/2014	1288.9	1311	1288.5	1306.9	1312.7	1301	1295.3	1289.3	1283.6	1277.6	1265.9	
8/7/2014	1306.6	1316.4	1303	1314.5	1347.1	1324.6	1315.8	1302.1	1293.3	1279.6	1257.1	
8/8/2014	1314.5	1324.3	1305.7	1310.6	1338.1	1324.7	1319.6	1311.3	1306.2	1297.9	1284.5	
8/11/2014	1310.4	1312.9	1306.4	1309.4	1350.7	1332.1	1321.4	1313.5	1302.8	1294.9	1276.3	
8/12/2014	1309.2	1319.3	1306.8	1310	1322.6	1316.1	1312.7	1309.6	1306.2	1303.1	1296.6	
8/13/2014	1309.6	1316.4	1306	1313.7	1337	1324.5	1317.3	1312	1304.8	1299.5	1287	
8/14/2014	1313.4	1321.8	1310	1313.9	1332.8	1322.4	1318.1	1312	1307.7	1301.6	1291.2	
8/15/2014	1313.4	1316.5	1293	1305.5	1338.8	1327	1320.5	1315.2	1308.7	1303.4	1291.6	
8/18/2014	1304.2	1304.9	1296.5	1298.2	1352	1328.5	1317	1305	1293.5	1281.5	1258	
8/19/2014	1298.6	1303.7	1294.7	1296.2	1316.7	1308.3	1303.2	1299.9	1294.8	1291.5	1283.1	
8/20/2014	1297.2	1299.3	1288.7	1292.5	1316.2	1307.2	1301.7	1298.2	1292.7	1289.2	1280.2	
8/21/2014	1291.9	1292	1273.4	1277.3	1314.7	1304.1	1298.3	1293.5	1287.7	1282.9	1272.3	
8/22/2014	1278.2	1283.9	1274.6	1281.8	1318.1	1299.5	1288.4	1280.9	1269.8	1262.3	1243.7	
8/25/2014	1280.8	1281.6	1275	1277	1298.7	1289.4	1285.6	1280.1	1276.3	1270.8	1261.5	
8/26/2014	1277.4	1291.9	1275.7	1281.4	1291.1	1284.5	1280.7	1277.9	1274.1	1271.3	1264.7	
8/27/2014	1281.6	1288.2	1280.9	1283.5	1315.4	1299.2	1290.3	1283	1274.1	1266.8	1250.6	
8/28/2014	1283.7	1297.6	1283	1290	1298.8	1291.5	1287.5	1284.2	1280.2	1276.9	1269.6	
8/29/2014	1290.4	1292.5	1284.1	1288	1319.4	1304.8	1297.4	1290.2	1282.8	1275.6	1261	
9/2/2014	1288.5	1288.8	1263.1	1266.5	1305	1296.6	1292.3	1288.2	1283.9	1279.8	1271.4	
9/3/2014	1266.7	1272.4	1261.9	1270	1324.2	1298.5	1282.5	1272.8	1256.8	1247.1	1221.4	
9/4/2014	1269.8	1279.2	1261.3	1261.7	1289.1	1278.6	1274.3	1268.1	1263.8	1257.6	1247.1	
9/5/2014	1261.7	1274.8	1258	1269.2	1303.2	1285.3	1273.5	1267.4	1255.6	1249.5	1231.6	
9/8/2014	1269.5	1272.6	1252.1	1256.3	1300.9	1284.1	1276.7	1267.3	1259.9	1250.5	1233.7	
9/9/2014	1257.1	1258.9	1248.1	1256.6	1301.3	1280.8	1268.6	1260.3	1248.1	1239.8	1219.3	
9/10/2014	1256.2	1258.5	1244.5	1249.8	1276.1	1265.3	1261	1254.5	1250.2	1243.7	1232.9	
9/11/2014	1249.8	1251	1235.3	1240.9	1278.9	1264.9	1257.4	1250.9	1243.4	1236.9	1222.9	
9/12/2014	1240.9	1242.3	1228.1	1229	1273.8	1258.1	1249.5	1242.4	1233.8	1226.7	1211	

Typical Dataset

	A	B	C	D	E	F	G	H	I	J	K	L	s
	Date	Open	High	Low	Close	R3	R2	R1	Pivot	S1	S2	S3	
Training Data	8/4/2014	1295.4	1296.4	1287	1289.2	1326.3	1308.9	1302.1	1291.5	1284.7	1274.1	1256.7	
	8/5/2014	1289.2	1295	1283.3	1289.6	1309.7	1300.3	1294.7	1290.9	1285.3	1281.5	1272.1	
	8/6/2014	1288.9	1311	1288.5	1306.9	1312.7	1301	1295.3	1289.3	1283.6	1277.6	1265.9	
	8/7/2014	1306.6	1316.4	1303	1314.5	1347.1	1324.6	1315.8	1302.1	1293.3	1279.6	1257.1	
	8/8/2014	1314.5	1324.3	1305.7	1310.6	1338.1	1324.7	1319.6	1311.3	1306.2	1297.9	1284.5	
	8/11/2014	1310.4	1312.9	1306.4	1309.4	1350.7	1332.1	1321.4	1313.5	1302.8	1294.9	1276.3	
	8/12/2014	1309.2	1319.3	1306.8	1310	1322.6	1316.1	1312.7	1309.6	1306.2	1303.1	1296.6	
	8/13/2014	1309.6	1316.4	1306	1313.7	1337	1324.5	1317.3	1312	1304.8	1299.5	1287	
	8/14/2014	1313.4	1321.8	1310	1313.9	1332.8	1322.4	1318.1	1312	1307.7	1301.6	1291.2	
	8/15/2014	1313.4	1316.5	1293	1305.5	1338.8	1327	1320.5	1315.2	1308.7	1303.4	1291.6	
	8/18/2014	1304.2	1304.9	1296.5	1298.2	1352	1328.5	1317	1305	1293.5	1281.5	1258	
	8/19/2014	1298.6	1303.7	1294.7	1296.2	1316.7	1308.3	1303.2	1299.9	1294.8	1291.5	1283.1	
	8/20/2014	1297.2	1299.3	1288.7	1292.5	1316.2	1307.2	1301.7	1298.2	1292.7	1289.2	1280.2	
	8/21/2014	1291.9	1292	1273.4	1277.3	1314.7	1304.1	1298.3	1293.5	1287.7	1282.9	1272.3	
	8/22/2014	1278.2	1283.9	1274.6	1281.8	1318.1	1299.5	1288.4	1280.9	1269.8	1262.3	1243.7	
	8/25/2014	1280.8	1281.6	1275	1277	1298.7	1289.4	1285.6	1280.1	1276.3	1270.8	1261.5	
Test Data	8/26/2014	1277.4	1291.9	1275.7	1281.4	1291.1	1284.5	1280.7	1277.9	1274.1	1271.3	1264.7	
	8/27/2014	1281.6	1288.2	1280.9	1283.5	1315.4	1299.2	1290.3	1283	1274.1	1266.8	1250.6	
	8/28/2014	1283.7	1297.6	1283	1290	1298.8	1291.5	1287.5	1284.2	1280.2	1276.9	1269.6	
	8/29/2014	1290.4	1292.5	1284.1	1288	1319.4	1304.8	1297.4	1290.2	1282.8	1275.6	1261	
	9/2/2014	1288.5	1288.8	1263.1	1266.5	1305	1296.6	1292.3	1288.2	1283.9	1279.8	1271.4	
	9/3/2014	1266.7	1272.4	1261.9	1270	1324.2	1298.5	1282.5	1272.8	1256.8	1247.1	1221.4	
	9/4/2014	1269.8	1279.2	1261.3	1261.7	1289.1	1278.6	1274.3	1268.1	1263.8	1257.6	1247.1	
	9/5/2014	1261.7	1274.8	1258	1269.2	1303.2	1285.3	1273.5	1267.4	1255.6	1249.5	1231.6	
	9/8/2014	1269.5	1272.6	1252.1	1256.3	1300.9	1284.1	1276.7	1267.3	1259.9	1250.5	1233.7	
	9/9/2014	1257.1	1258.9	1248.1	1256.6	1301.3	1280.8	1268.6	1260.3	1248.1	1239.8	1219.3	
	9/10/2014	1256.2	1258.5	1244.5	1249.8	1276.1	1265.3	1261	1254.5	1250.2	1243.7	1232.9	
	9/11/2014	1249.8	1251	1235.3	1240.9	1278.9	1264.9	1257.4	1250.9	1243.4	1236.9	1222.9	
	9/12/2014	1240.9	1242.3	1228.1	1229	1273.8	1258.1	1249.5	1242.4	1233.8	1226.7	1211	


Bias Variance Tradeoff

- Bias : Machine refuses to learn anything from training
 - Variance : Learns only the training data set
-

Model Validation

- Why do we need to split the data into training and test
 - Some of the machines suffer from high variance problem
 - Train -Test-Train -Test-Train - Test-Train -Test-Train -Test-Train-Train-Train-Train.....
 - Various Types Of Validation Datasets
 - Validation set
 - K fold cv
 - LOOCV
-

Validation Set



	A	B	C	D	E	F	G	H	I	J	K	L	s
	Date	Open	High	Low	Close	R3	R2	R1	Pivot	S1	S2	S3	
	8/4/2014	1295.4	1296.4	1287	1289.2	1326.3	1308.9	1302.1	1291.5	1284.7	1274.1	1256.7	
	8/5/2014	1289.2	1295	1283.3	1289.6	1309.7	1300.3	1294.7	1290.9	1285.3	1281.5	1272.1	
	8/6/2014	1288.9	1311	1288.5	1306.9	1312.7	1301	1295.3	1289.3	1283.6	1277.6	1265.9	
	8/7/2014	1306.6	1316.4	1303	1314.5	1347.1	1324.6	1315.8	1302.1	1293.3	1279.6	1257.1	
	8/8/2014	1314.5	1324.3	1305.7	1310.6	1338.1	1324.7	1319.6	1311.3	1306.2	1297.9	1284.5	
	8/11/2014	1310.4	1312.9	1306.4	1309.4	1350.7	1332.1	1321.4	1313.5	1302.8	1294.9	1276.3	
	8/12/2014	1309.2	1319.3	1306.8	1310	1322.6	1316.1	1312.7	1309.6	1306.2	1303.1	1296.6	
	8/13/2014	1309.6	1316.4	1306	1313.7	1337	1324.5	1317.3	1312	1304.8	1299.5	1287	
	8/14/2014	1313.4	1321.8	1310	1313.9	1332.8	1322.4	1318.1	1312	1307.7	1301.6	1291.2	
	8/15/2014	1313.4	1316.5	1293	1305.5	1338.8	1327	1320.5	1315.2	1308.7	1303.4	1291.6	
	8/18/2014	1304.2	1304.9	1296.5	1298.2	1352	1328.5	1317	1305	1293.5	1281.5	1258	
	8/19/2014	1298.6	1303.7	1294.7	1296.2	1316.7	1308.3	1303.2	1299.9	1294.8	1291.5	1283.1	
	8/20/2014	1297.2	1299.3	1288.7	1292.5	1316.2	1307.2	1301.7	1298.2	1292.7	1289.2	1280.2	
	8/21/2014	1291.9	1292	1273.4	1277.3	1314.7	1304.1	1298.3	1293.5	1287.7	1282.9	1272.3	
	8/22/2014	1278.2	1283.9	1274.6	1281.8	1318.1	1299.5	1288.4	1280.9	1269.8	1262.3	1243.7	
	8/25/2014	1280.8	1281.6	1275	1277	1298.7	1289.4	1285.6	1280.1	1276.3	1270.8	1261.5	
	8/26/2014	1277.4	1291.9	1275.7	1281.4	1291.1	1284.5	1280.7	1277.9	1274.1	1271.3	1264.7	
	8/27/2014	1281.6	1288.2	1280.9	1283.5	1315.4	1299.2	1290.3	1283	1274.1	1266.8	1250.6	
	8/28/2014	1283.7	1297.6	1283	1290	1298.8	1291.5	1287.5	1284.2	1280.2	1276.9	1269.6	
	8/29/2014	1290.4	1292.5	1284.1	1288	1319.4	1304.8	1297.4	1290.2	1282.8	1275.6	1261	
	9/2/2014	1288.5	1288.8	1263.1	1266.5	1305	1296.6	1292.3	1288.2	1283.9	1279.8	1271.4	
	9/3/2014	1266.7	1272.4	1261.9	1270	1324.2	1298.5	1282.5	1272.8	1256.8	1247.1	1221.4	
	9/4/2014	1269.8	1279.2	1261.3	1261.7	1289.1	1278.6	1274.3	1268.1	1263.8	1257.6	1247.1	
	9/5/2014	1261.7	1274.8	1258	1269.2	1303.2	1285.3	1273.5	1267.4	1255.6	1249.5	1231.6	
	9/8/2014	1269.5	1272.6	1252.1	1256.3	1300.9	1284.1	1276.7	1267.3	1259.9	1250.5	1233.7	
	9/9/2014	1257.1	1258.9	1248.1	1256.6	1301.3	1280.8	1268.6	1260.3	1248.1	1239.8	1219.3	
	9/10/2014	1256.2	1258.5	1244.5	1249.8	1276.1	1265.3	1261	1254.5	1250.2	1243.7	1232.9	
	9/11/2014	1249.8	1251	1235.3	1240.9	1278.9	1264.9	1257.4	1250.9	1243.4	1236.9	1222.9	
	9/12/2014	1240.9	1242.3	1228.1	1229	1273.8	1258.1	1249.5	1242.4	1233.8	1226.7	1211	

K-Fold Cross Validation

- $K=3$, Divide Training data into 3 equal parts through random sampling



- Validate on each and train on other two
- Finally average over 3

Classification Performance

- Confusion Matrix

PREDICTED	
ACTUAL	00
	01
10	11

Classification Performance

- Accuracy : Winners and Losers predicted correctly
 - Sensitivity : Percentage of winners identified
 - Specificity : Percentage of losers Identified
 - Precision : predicted Winners that actually won
 - ROC : sensitivity plotted on y –axis and specificity on x axis .
 - $AUC > .5$ implies predictive power
-

-
- LAB : Build a random classifier on gold . Calculate :
 - Accuracy
 - Precision
 - AUC of the ROC curve
-

Feature Selection

Overview

- Why ?
 - Filters Methods
 - Correlation
 - Variance threshold
 - T-test
 - Wrapper Methods
 - Forward Selection
 - Backward Elimination
 - Stepwise Selection
 - Recursive Feature Elimination
-

Filter Methods

- Statistical methods for feature elimination
 - Rank each feature according to some univariate metric and select the highest ranking features or eliminate based on a threshold
 - These methods are independent of the predictive model adapted
-

Examples of Filter Methods

- Correlation
- Variance Threshold

Predictors with high variance have higher data spread and more conducive for prediction. So, we eliminate predictors with variance lesser than a particular value.

- T Test

Statistical hypothesis tests used to determine the validity of the null hypothesis that the predictor and category are independent.

Filter Methods

- Advantages:
 - Very efficient and fast to compute
- Disadvantages:
 - A feature that is not useful by itself can provide a significant performance improvement when taken with others. Filter methods can miss it

Wrapper Methods

- Assess the quality of a set of features using a specific algorithm by internal cross-validation
- In essence, wrapper methods are search algorithms that treat the predictors as the inputs and utilize model performance as the output to be optimized.

Wrapper Methods

- RFE
 - Forward Selection
 - Backward Elimination
-

Recursive Feature Elimination

- Cross validate to obtain best set of hyper parameters to build a model
 - Rank features by importance
 - Take a predetermined subset of most important features
 - Repeat above steps predetermined number of times
 - Feature importance can be ranked by shuffling each feature and measuring impact on accuracy.
-

Logistic Regression

Overview

- Introduction
 - Guidelines for using the model
 - Modelling Approach
 - Fitting Regression Model
 - Linear Discriminant Analysis
 - Ridge and Lasso
-

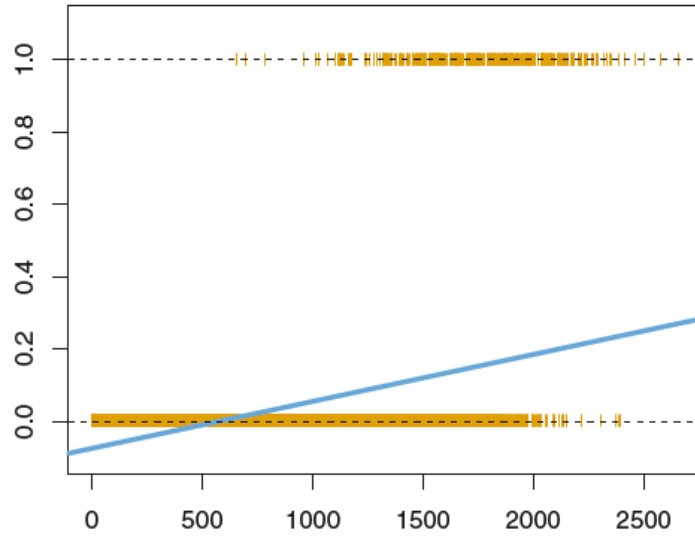
Introduction

- Models Conditional Probability
 - Problems Of modelling probabilities as a linear regression ?
 - Mathematical Trick to overcome above
-

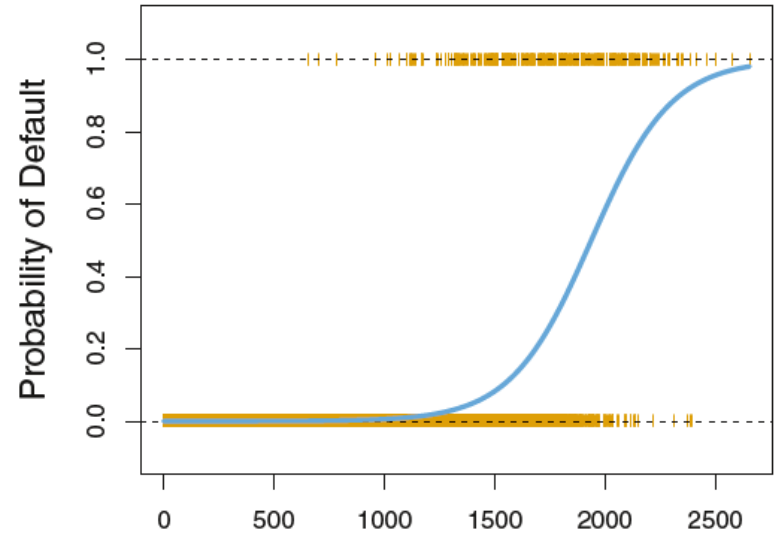
Introduction

- Model $\text{Log}(\text{Odds})$ as a linear regression
- $\text{Odds} = p/(1-p)$
- $\text{Logit}(p) = \text{Log}(\text{odds}) = \beta * X$
- $P = \exp(\beta * X) / (1 + \exp(\beta * X))$

Introduction



Linear Model



Logistic Model

Guidelines for using the model

- Outcome must be binary
 - Outliers in the data should be removed, i.e. samples with z values below -3.29 or greater than 3.29 should be removed
 - High inter correlations (multicollinearity) among the predictors is not preferable.
-

Modelling Outline

- Maximize Logarithm of the likelihood function
- Likelihood

$$L(\beta_0, \beta) = \prod_{i=1}^n p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}$$

- Logarithm of the likelihood

$$\ell(\beta_0, \beta) = \sum_{i=1}^n y_i \log p(x_i) + (1 - y_i) \log 1 - p(x_i)$$

- Differential Calculus to maximize the above expression

Fitting Regression Model

- To maximize the log-likelihood, we set its derivatives to zero

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^N x_i (y_i - p(x_i; \beta)) = 0,$$

Modelling Approach

$$\Pr(G = k|X = x) = \frac{\exp(\beta_{k0} + \beta_k^T x)}{1 + \sum_{\ell=1}^{K-1} \exp(\beta_{\ell 0} + \beta_{\ell}^T x)},$$

$$\Pr(G = K|X = x) = \frac{1}{1 + \sum_{\ell=1}^{K-1} \exp(\beta_{\ell 0} + \beta_{\ell}^T x)},$$

- The entire parameter set $\theta = \{\beta_{10}, \beta_1^T, \dots, \beta_{(K-1)0}, \beta_{K-1}^T\}$,
- we denote the probabilities $\Pr(G = k|X = x) = p_k(x; \theta)$.
- When $K = 2$, then it is said as binary class logistic regression.

Results

- After solving the MLE (maximum likelihood estimate) we get
 - Coefficient Values
 - Std errors of the coefficients
 - Z stat
 - P value
 - **Null Hypothesis** : Coefficient value is Zero
 - **Summary** : If coefficients have a low p value we have faith in the regression model
-

Lab2

- Logistic regression

Linear Discriminant Analysis

- LDA is better suited than LR when
 - Number of samples is small
 - Multi class classification problem

Linear Discriminant Analysis

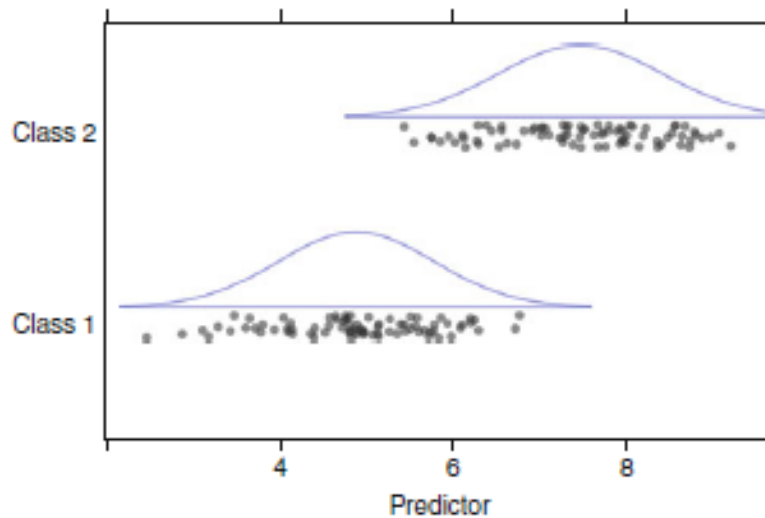
- Application of Bayes theorem

$$\Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^K \pi_l f_l(x)}$$

- Conditional Probability : $\Pr(Y = k|X = x)$
- Posterior probability : $\Pr(X = x|Y = k)$
- Probability Of sample belonging to a class k : π_k

Linear Discriminant Analysis

$$\Pr(X = x | Y = k)$$



Linear Discriminant Analysis

- Number of predictors= 1
- To calculate the posterior probability , we assume the distribution to be Normal within each class
- To estimate the mean and variance we assume
 - Mean equals class mean of the sample
 - Variance is same across all classes

π_k

Linear Discriminant Analysis

- Class mean and class variance is given by

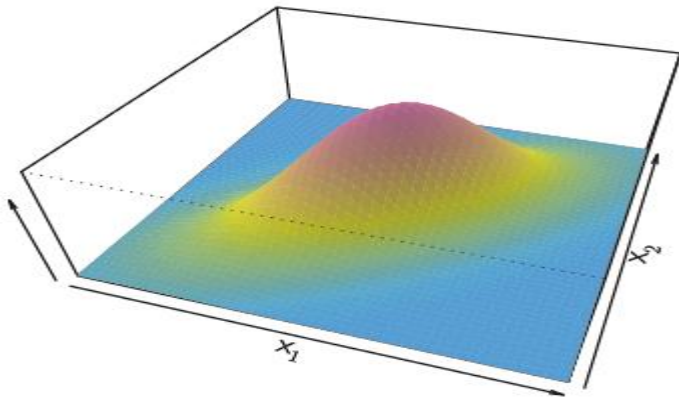
$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n - K} \sum_{k=1}^K \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2$$

π_k

Linear Discriminant Analysis

- Number of predictors > 1
- To calculate the posterior probability , we assume the distribution to be **multivariate Normal within each class**



Linear Discriminant Analysis

- We need to estimate the class mean for each class which will have dimension $p \times 1$ for each class
- A variance covariance matrix , shared across all classes with a dimension $p \times p$.

Quadratic Discriminant Analysis

- **A variance covariance matrix , is built for Each class**
with a dimension $p \times p$.

Linear Discriminant Analysis

- LDA can be used with both single and multiple predictors
 - In case of multiple predictors the density function is assumed to be drawn from a Multinomial Gaussian distribution rather than a simple Gaussian Distribution
 - Why use LDA? : Better performance in case of multiple response classes and stable even in case of well-separated classes
-

Lab 2

- Linear Discriminant Analysis

Ridge and Lasso

- Logistic regression parameters were solved using MLE , by minimizing a certain objective function (Entropy)
- Ridge regression
 - Original function + $\lambda \sum_{j=1}^p \beta_j^2$,
- Lasso Regression
 - Original Function + $\lambda \sum_{j=1}^p |\beta_j|$

Lab 2

- Regularization

Tree Based Methods

Overview

- Introduction
 - CART
 - Bagging
 - Random Forest
 - Gradient Boosted Trees
-

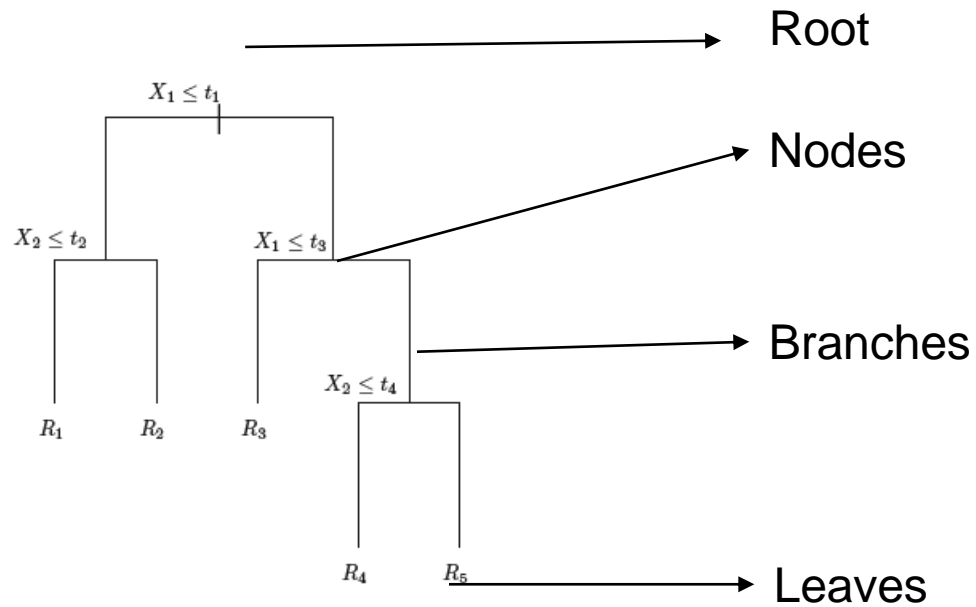
Introduction

- Supervised learning algorithm for classification and regression problems
 - Divide datasets like binary trees in data structures
 - Based on the principle that a group of weak learners can come together to form a strong learner
-

Introduction

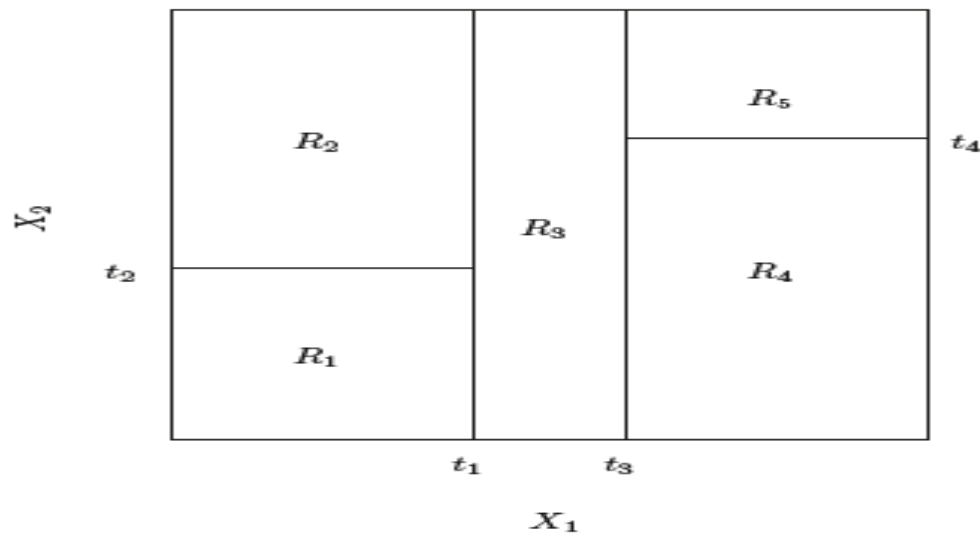
- Extends the concept of binary trees to split the data of a dataset in two parts, right and left
- Each node in the tree represents the input variable based on which the dataset should be split and the split value of that variable
- The leaf nodes of the trees are the values of output variable

Decision Trees



Decision Trees

Alternative representation



Decision Trees

- **Task is two fold**
 - Divide predictor space into certain number of rectangular regions
 - Every observation belonging to a particular region has the same predicted value/class
-

CART

- **Minimise the below for a regression problem**

$$\sum_{i: x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i: x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2,$$

CART

- Loop over all predictors, and for each predictor find ideal candidate points to split the data set into two halves .
 - From all the potential predictor and “candidate point” combination we select that pair which minimizes a certain objective function.
 - Repeat the above search for each predictor but **limit the observation space to the newly found regions.**
 - Do you see how overfitting happens as a result of this .
-

CART

- Minimise the below for a **classification** problem : Gini Index

$$p(1-p) + q(1-q) + r(1-r)$$

p : proportion of samples belonging to class 1

q : proportion of samples belonging to class 0

r : proportion of samples belonging to class 2

- The above rewards node purity
- **Alternatively** minimize :

$$-p \ln(p) - q \ln(q) - r \ln(r)$$

Pruning

- If RSS stops changing by a minimum amount we stop pruning
---- Short-sighted
 - Cost complexity pruning
-

Cost complexity pruning

- Make a list of various sub trees
- For a particular value of alpha calculate the below expression for all the subtrees

$$\sum_{m=1}^{|T|} \sum_{i: x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$$

SubTrees\alpha	0	0.2	0.4	0.6	0.8	1	100
T0	0.005	0.706528	0.901173	0.300269	0.526376	0.833438	0.74279
T1	0.173048	0.163957	0.187586	0.873345	0.637597	0.212069	0.222974
T2	0.615329	0.531109	0.40785	0.048201	0.833519	0.93335	0.286031
T3	0.781338	0.92636	0.467605	0.07	0.589382	0.07	0.819952
T4	0.723903	0.490303	0.41486	0.927928	0.05	0.14571	0.232553

Cost complexity pruning

- Find the minimum value of the above expression among all the subtrees for that value of alpha
- Map various alpha values to various subtrees

Subtree	alpha
T0	0
T1	0.2
T2	0.6
T3	0.07
T4	0.05

Cost complexity pruning

- Calculate validation set accuracy/mse for each subtree/alpha

Subtree	alpha	Validation accuracy
T0	0	56%
T1	0.2	67%
T2	0.6	45%
T3	0.07	58%
T4	0.05	80%

- Take alpha that maximises accuracy, hence you have a subtree which yields higher validation data set performance

Ensemble Methods

- Use multiple models to obtain better predictive performance (bias variance tradeoff)
 - Each learning method is a hypothesis and ensembles combine multiple hypotheses in the hope to form a better hypothesis
 - Typically more computation is required since it requires training multiple learners
-

Ensemble Methods

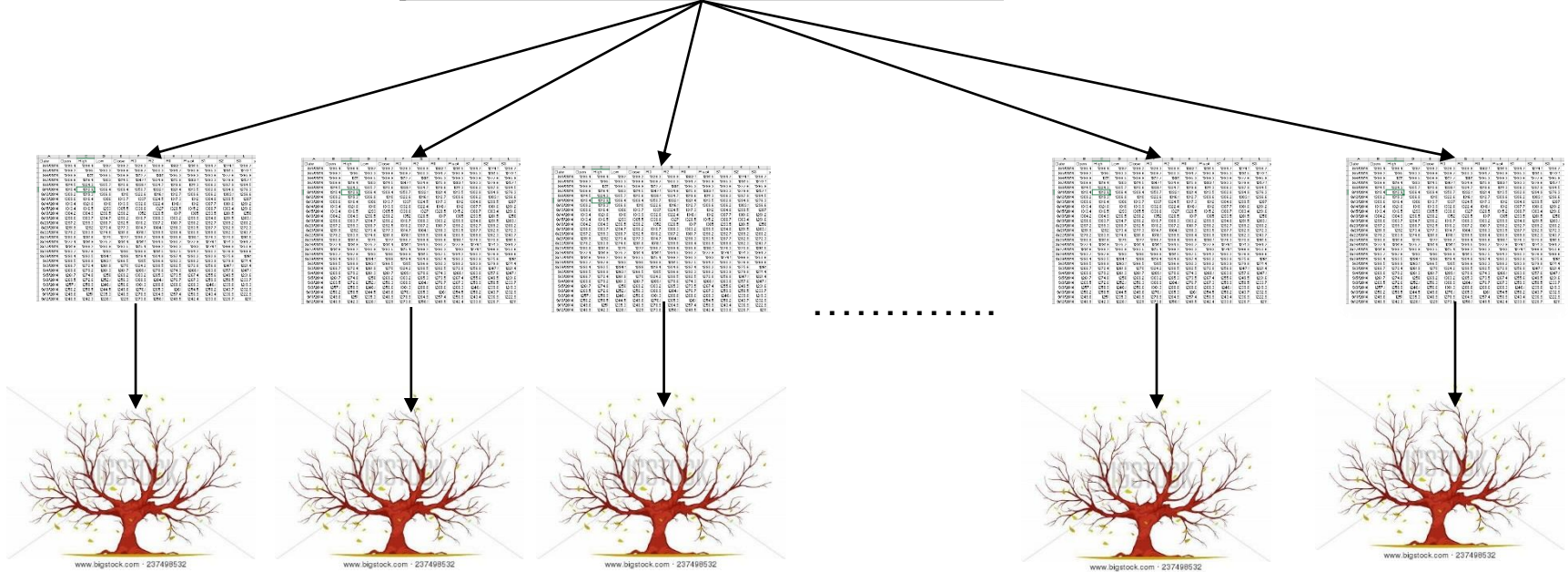
- Bagging
 - Random Forest
 - Boosting
-

B(Bootstrap)AGGING(Aggregating)

- Instead of training a tree on the entire dataset, we can train a bunch of them on bootstrapped samples of the dataset
 - Each model in the ensemble is given an equal weight
 - These are collectively called bagging trees
-

B(Bootstrap)AGGING(Aggregating)

A	B	C	D	E	F	G	H	I	J	K	L	S
Date	Open	High	Low	Close	R3	R2	R1	Pivot	S1	S2	S3	s
8/4/2014	1295.4	1296.4	1287	1289.2	1326.3	1308.9	1302.1	1291.5	1284.7	1274.1	1256.7	
8/5/2014	1289.2	1295	1283.3	1289.6	1309.7	1300.3	1294.7	1290.9	1285.3	1281.5	1272.1	
8/6/2014	1288.9	1311	1288.5	1306.9	1312.7	1301	1295.3	1289.3	1283.6	1277.6	1265.9	
8/7/2014	1306.6	1316.4	1303	1314.5	1347.1	1324.6	1315.8	1302.1	1293.3	1279.6	1257.1	
8/8/2014	1314.5	1324.3	1305.7	1310.6	1358.1	1324.7	1319.6	1311.3	1306.2	1297.9	1284.5	
8/11/2014	1310.4	1312.9	1306.4	1309.4	1350.7	1332.1	1321.4	1313.5	1302.8	1294.9	1276.3	
8/12/2014	1309.2	1319.3	1306.8	1310	1322.6	1316.1	1312.7	1309.6	1306.2	1303.1	1296.6	
8/13/2014	1309.6	1316.4	1306	1313.7	1337	1324.5	1317.3	1312	1304.8	1299.5	1287	
8/14/2014	1313.4	1321.8	1310	1313.9	1332.8	1322.4	1318.1	1312	1307.7	1301.6	1291.2	
8/15/2014	1313.4	1316.5	1293	1305.5	1338.8	1327	1320.5	1315.2	1308.7	1303.4	1291.6	
8/18/2014	1304.2	1304.9	1296.5	1298.2	1352	1328.5	1317	1305	1293.5	1281.5	1258	
8/19/2014	1298.6	1303.7	1294.7	1296.2	1316.7	1308.3	1303.2	1299.9	1294.8	1291.5	1283.1	
8/20/2014	1297.2	1299.3	1286.7	1292.5	1316.2	1307.2	1301.7	1298.2	1292.7	1289.2	1280.2	
8/21/2014	1291.9	1292	1273.4	1277.3	1314.7	1304.1	1298.3	1293.5	1287.7	1282.9	1272.3	
8/22/2014	1278.2	1283.9	1274.6	1281.8	1318.1	1299.5	1288.4	1280.9	1269.8	1262.3	1243.7	
8/25/2014	1280.8	1281.6	1275	1277	1298.7	1289.4	1285.6	1280.1	1276.3	1270.8	1261.5	
8/26/2014	1277.4	1291.9	1275.7	1281.4	1291.1	1284.5	1280.7	1277.9	1274.1	1271.3	1264.7	
8/27/2014	1281.6	1288.2	1280.9	1283.5	1315.4	1299.2	1290.3	1283	1274.1	1266.8	1250.6	
8/28/2014	1283.7	1297.6	1283	1290	1298.8	1291.5	1287.5	1284.2	1280.2	1276.9	1269.6	
8/29/2014	1290.4	1292.5	1284.1	1288	1319.4	1304.8	1297.4	1290.2	1282.8	1275.6	1261	
9/2/2014	1288.5	1288.8	1263.1	1266.5	1305	1296.6	1292.3	1288.2	1283.9	1279.8	1271.4	
9/3/2014	1266.7	1272.4	1261.9	1270	1324.2	1298.5	1292.5	1272.8	1266.8	1247.1	1221.4	
9/4/2014	1269.8	1279.2	1261.3	1261.7	1289.1	1278.6	1274.3	1268.1	1263.8	1257.6	1247.1	
9/5/2014	1261.7	1274.8	1258	1269.2	1303.2	1285.3	1273.5	1267.4	1255.6	1249.5	1231.6	
9/8/2014	1269.5	1272.6	1252.1	1256.3	1300.9	1284.1	1276.7	1267.3	1259.9	1250.5	1233.7	
9/9/2014	1257.1	1258.9	1248.1	1256.6	1301.3	1280.8	1268.6	1260.3	1248.1	1239.8	1219.3	
9/10/2014	1256.2	1258.5	1244.5	1249.8	1276.1	1265.3	1261	1254.5	1250.2	1243.7	1232.9	
9/11/2014	1249.8	1251	1235.3	1240.9	1278.9	1264.9	1257.4	1250.9	1243.4	1236.9	1222.9	
9/12/2014	1240.9	1242.3	1228.1	1229	1273.8	1258.1	1249.5	1242.4	1233.8	1226.7	1211	



Methodology of Bagging

- For $i = 1 \dots M$
 - - Draw $n' < n$ (bootstrap) samples from the dataset
 - - Learn classifier C_i on each of the bootstrapped sample
- Final classifier is a vote of $C_1 \dots C_M$
- Increases classifier stability/ reduces variance
- Inbuilt validation data set : OOB

Methodology of Bagging

- Why bagging works
 - Averaging reduces variance
 - Each bag forces the tree to focus on a few samples. Not just on the prominent ones
-

Random Forests

- Random Forests are an improvement over the Bagging Trees
 - Bagging trees use greedy criteria at each node split to minimize the error and this might result in higher correlation among the different learners
 - Random Forests changes the way the trees are learned by removing this greedy criteria and randomly selecting the features to be used at each node
-

Random Forests

A	B	C	D	E	F	G	H	I	J	K	L	M
Date	Open	High	Low	Close	R3	R2	R1	Pivot	S1	S2	S3	S
8/4/2014	1295.4	1296.4	1287	1289.2	1286.3	1308.9	1302.1	1291.5	1284.7	1274.1	1256.7	
8/5/2014	1289.2	1295	1283.3	1289.6	1309.7	1300.3	1294.7	1290.9	1285.3	1281.5	1272.1	
8/6/2014	1288.9	1311	1288.5	1306.9	1312.7	1301	1295.3	1289.3	1283.6	1277.6	1265.9	
8/7/2014	1306.6	1316.4	1303	1314.5	1347.1	1324.6	1315.8	1302.1	1293.3	1279.6	1257.1	
8/8/2014	1314.5	1324.3	1305.7	1310.6	1338.1	1324.7	1319.6	1311.3	1306.2	1297.9	1284.5	
8/11/2014	1310.4	1312.9	1306.4	1309.4	1350.7	1332.1	1321.4	1313.5	1302.8	1294.9	1276.3	
8/12/2014	1309.2	1319.3	1306.8	1310	1322.6	1316.1	1312.7	1309.6	1306.2	1303.1	1296.6	
8/13/2014	1309.6	1316.4	1306	1313.7	1337	1324.5	1317.3	1312	1304.8	1299.5	1287	
8/14/2014	1313.4	1321.8	1310	1313.9	1332.8	1322.4	1318.1	1312	1307.7	1301.6	1291.2	
8/15/2014	1313.4	1316.5	1293	1305.5	1338.8	1327	1320.5	1315.2	1308.7	1303.4	1291.6	
8/18/2014	1304.2	1304.9	1296.5	1298.2	1352	1328.5	1317	1305	1293.5	1281.5	1258	
8/19/2014	1298.6	1303.7	1294.7	1296.2	1316.7	1308.3	1303.2	1299.9	1294.8	1291.5	1283.1	
8/20/2014	1297.2	1299.3	1286.7	1292.5	1316.2	1307.2	1301.7	1298.2	1293.2	1289.2	1280.2	
8/21/2014	1291.9	1292	1273.4	1277.3	1314.7	1304.1	1298.3	1293.5	1287.7	1282.9	1272.3	
8/22/2014	1278.2	1283.9	1274.6	1281.8	1318.1	1293.5	1288.4	1280.9	1263.8	1262.3	1243.7	
8/25/2014	1280.8	1281.6	1275	1277	1298.7	1289.4	1285.6	1280.1	1276.3	1270.8	1261.5	
8/26/2014	1277.4	1291.9	1275.7	1281.4	1291.1	1284.5	1280.7	1277.9	1274.1	1271.3	1264.7	
8/27/2014	1281.6	1286.2	1280.9	1283.5	1315.4	1293.2	1290.3	1283	1274.1	1266.8	1250.6	
8/28/2014	1283.7	1297.6	1283	1290	1298.8	1291.5	1287.5	1284.2	1280.2	1276.9	1269.6	
8/29/2014	1290.4	1292.5	1284.1	1288	1319.4	1304.8	1297.4	1290.2	1282.8	1275.6	1261	
9/2/2014	1288.5	1288.8	1263.1	1266.5	1305	1296.6	1292.3	1288.2	1283.9	1279.8	1271.4	
9/3/2014	1266.7	1272.4	1261.9	1270	1324.2	1298.5	1282.5	1272.8	1266.8	1247.1	1221.4	
9/4/2014	1269.8	1279.2	1261.3	1261.7	1289.1	1278.6	1274.3	1268.1	1263.8	1257.6	1247.1	
9/5/2014	1261.7	1274.8	1258	1269.2	1303.2	1285.3	1273.5	1267.4	1255.6	1243.5	1231.6	
9/8/2014	1269.5	1272.6	1252.1	1256.3	1300.9	1284.1	1276.7	1267.3	1259.9	1250.5	1233.7	
9/9/2014	1257.1	1258.9	1248.1	1256.6	1301.3	1280.8	1268.6	1260.3	1248.1	1239.8	1219.3	
9/10/2014	1256.2	1258.5	1244.5	1249.8	1276.1	1265.3	1261	1254.5	1250.2	1243.7	1232.9	
9/11/2014	1249.8	1251	1235.3	1240.9	1278.9	1264.9	1257.4	1250.9	1243.4	1236.9	1222.9	
9/12/2014	1240.9	1242.3	1228.1	1229	1273.8	1258.1	1249.5	1242.4	1233.8	1226.7	1211	

Bag

A	B	C	D	E	F	G	H	I	J	K	L	M
Date	Open	High	Low	Close	R3	R2	R1	Pivot	S1	S2	S3	S
8/4/2014	1295.4	1296.4	1287	1289.2	1286.3	1308.9	1302.1	1291.5	1284.7	1274.1	1256.7	
8/5/2014	1289.2	1295	1283.3	1289.6	1309.7	1300.3	1294.7	1290.9	1285.3	1281.5	1272.1	
8/6/2014	1288.9	1311	1288.5	1306.9	1312.7	1301	1295.3	1289.3	1283.6	1277.6	1265.9	
8/7/2014	1306.6	1316.4	1303	1314.5	1347.1	1324.6	1315.8	1302.1	1293.3	1279.6	1257.1	
8/8/2014	1314.5	1324.3	1305.7	1310.6	1338.1	1324.7	1319.6	1311.3	1306.2	1297.9	1284.5	
8/11/2014	1310.4	1312.9	1306.4	1309.4	1350.7	1332.1	1321.4	1313.5	1302.8	1294.9	1276.3	
8/12/2014	1309.2	1319.3	1306.8	1310	1322.6	1316.1	1312.7	1309.6	1306.2	1303.1	1296.6	
8/13/2014	1309.6	1316.4	1306	1313.7	1337	1324.5	1317.3	1312	1304.8	1299.5	1287	
8/14/2014	1313.4	1321.8	1310	1313.9	1332.8	1322.4	1318.1	1312	1307.7	1301.6	1291.2	
8/15/2014	1313.4	1316.5	1293	1305.5	1338.8	1327	1320.5	1315.2	1308.7	1303.4	1291.6	
8/18/2014	1304.2	1304.9	1296.5	1298.2	1352	1328.5	1317	1305	1293.5	1281.5	1258	
8/19/2014	1298.6	1303.7	1294.7	1296.2	1316.7	1308.3	1303.2	1299.9	1294.8	1291.5	1283.1	
8/20/2014	1297.2	1299.3	1286.7	1292.5	1316.2	1307.2	1301.7	1298.2	1293.2	1289.2	1280.2	
8/21/2014	1291.9	1292	1273.4	1277.3	1314.7	1304.1	1298.3	1293.5	1287.7	1282.9	1272.3	
8/22/2014	1278.2	1283.9	1274.6	1281.8	1318.1	1293.5	1288.4	1280.9	1263.8	1262.3	1243.7	
8/25/2014	1280.8	1281.6	1275	1277	1298.7	1289.4	1285.6	1280.1	1276.3	1270.8	1261.5	
8/26/2014	1277.4	1291.9	1275.7	1281.4	1291.1	1284.5	1280.7	1277.9	1274.1	1271.3	1264.7	
8/27/2014	1281.6	1286.2	1280.9	1283.5	1315.4	1293.2	1290.3	1283	1274.1	1266.8	1250.6	
8/28/2014	1283.7	1297.6	1283	1290	1298.8	1291.5	1287.5	1284.2	1280.2	1276.9	1269.6	
8/29/2014	1290.4	1292.5	1284.1	1288	1319.4	1304.8	1297.4	1290.2	1282.8	1275.6	1261	
9/2/2014	1288.5	1288.8	1263.1	1266.5	1305	1296.6	1292.3	1288.2	1283.9	1279.8	1271.4	
9/3/2014	1266.7	1272.4	1261.9	1270	1324.2	1298.5	1282.5	1272.8	1266.8	1247.1	1221.4	
9/4/2014	1269.8	1279.2	1261.3	1261.7	1289.1	1278.6	1274.3	1268.1	1263.8	1257.6	1247.1	
9/5/2014	1261.7	1274.8	1258	1269.2	1303.2	1285.3	1273.5	1267.4	1255.6	1243.5	1231.6	
9/8/2014	1269.5	1272.6	1252.1	1256.3	1300.9	1284.1	1276.7	1267.3	1259.9	1250.5	1233.7	
9/9/2014	1257.1	1258.9	1248.1	1256.6	1301.3	1280.8	1268.6	1260.3	1248.1	1239.8	1219.3	
9/10/2014	1256.2	1258.5	1244.5	1249.8	1276.1	1265.3	1261	1254.5	1250.2	1243.7	1232.9	
9/11/2014	1249.8	1251	1235.3	1240.9	1278.9	1264.9	1257.4	1250.9	1243.4	1236.9	1222.9	
9/12/2014	1240.9	1242.3	1228.1	1229	1273.8	1258.1	1249.5	1242.4	1233.8	1226.7	1211	

Random Forests

A	B	C	D	E	F	G	H	I	J	K	L	M
Date	Open	High	Low	Close	R3	R2	R1	Pivot	S1	S2	S3	S
8/4/2014	1295.4	1296.4	1287	1289.2	1286.3	1308.9	1302.1	1291.5	1284.7	1274.1	1256.7	
8/5/2014	1289.2	1295	1283.3	1289.6	1309.7	1300.3	1294.7	1290.9	1285.3	1281.5	1272.1	
8/6/2014	1288.9	1311	1288.5	1306.9	1312.7	1301	1295.3	1289.3	1283.6	1277.6	1265.9	
8/7/2014	1306.6	1316.4	1303	1314.5	1347.1	1324.6	1315.8	1302.1	1293.3	1279.6	1257.1	
8/8/2014	1314.5	1324.3	1305.7	1310.6	1338.1	1324.7	1319.6	1311.3	1306.2	1297.9	1284.5	
8/11/2014	1310.4	1312.9	1306.4	1309.4	1350.7	1332.1	1321.4	1313.5	1302.8	1294.9	1276.3	
8/12/2014	1309.2	1319.3	1306.8	1310	1322.6	1316.1	1312.7	1309.6	1306.2	1303.1	1296.6	
8/13/2014	1309.6	1316.4	1306	1313.7	1337	1324.5	1317.3	1312	1304.8	1299.5	1287	
8/14/2014	1313.4	1321.8	1310	1313.9	1332.8	1322.4	1318.1	1312	1307.7	1301.6	1291.2	
8/15/2014	1313.4	1316.5	1293	1305.5	1338.8	1327	1320.5	1315.2	1308.7	1303.4	1291.6	
8/18/2014	1304.2	1304.9	1296.5	1298.2	1352	1328.5	1317	1305	1293.5	1281.5	1258	
8/19/2014	1298.6	1303.7	1294.7	1296.2	1316.7	1308.3	1303.2	1299.9	1294.8	1291.5	1283.1	
8/20/2014	1297.2	1299.3	1286.7	1292.5	1316.2	1307.2	1301.7	1298.2	1292.7	1289.2	1280.2	
8/21/2014	1291.9	1292	1273.4	1277.3	1314.7	1304.1	1298.3	1293.5	1287.7	1282.9	1272.3	
8/22/2014	1278.2	1283.9	1274.6	1281.8	1318.1	1293.5	1288.4	1280.9	1263.8	1262.3	1243.7	
8/25/2014	1280.8	1281.6	1275	1277	1298.7	1289.4	1285.6	1280.1	1276.3	1270.8	1261.5	
8/26/2014	1277.4	1291.9	1275.7	1281.4	1291.1	1284.5	1280.7	1277.9	1274.1	1271.3	1264.7	
8/27/2014	1281.6	1286.2	1280.9	1283.5	1315.4	1293.2	1290.3	1283	1274.1	1266.8	1250.6	
8/28/2014	1283.7	1297.6	1283	1290	1298.8	1291.5	1287.5	1284.2	1280.2	1276.9	1269.6	
8/29/2014	1290.4	1292.5	1284.1	1288	1319.4	1304.8	1297.4	1290.2	1282.8	1275.6	1261	
9/2/2014	1288.5	1288.8	1263.1	1266.5	1305	1296.6	1292.3	1288.2	1283.9	1279.8	1271.4	
9/3/2014	1266.7	1272.4	1261.9	1270	1324.2	1298.5	1282.5	1272.8	1266.8	1247.1	1221.4	
9/4/2014	1269.8	1279.2	1261.3	1261.7	1289.1	1278.6	1274.3	1268.1	1263.8	1257.6	1247.1	
9/5/2014	1261.7	1274.8	1258	1269.2	1303.2	1285.3	1273.5	1267.4	1255.6	1243.5	1231.6	
9/8/2014	1269.5	1272.6	1252.1	1256.3	1300.9	1284.1	1276.7	1267.3	1259.9	1250.5	1233.7	
9/9/2014	1257.1	1258.9	1248.1	1256.6	1301.3	1280.8	1268.6	1260.3	1248.1	1239.8	1219.3	
9/10/2014	1256.2	1258.5	1244.5	1249.8	1276.1	1265.3	1261	1254.5	1250.2	1243.7	1232.9	
9/11/2014	1249.8	1251	1235.3	1240.9	1278.9	1264.9	1257.4	1250.9	1243.4	1236.9	1222.9	
9/12/2014	1240.9	1242.3	1228.1	1229	1273.8	1258.1	1249.5	1242.4	1233.8	1226.7	1211	

Bag

A	B	C	D	E	F	G	H	I	J	K	L	M
Date	Open	High	Low	Close	R3	R2	R1	Pivot	S1	S2	S3	S
8/4/2014	1295.4	1296.4	1287	1289.2	1286.3	1308.9	1302.1	1291.5	1284.7	1274.1	1256.7	
8/5/2014	1289.2	1295	1283.3	1289.6	1309.7	1300.3	1294.7	1290.9	1285.3	1281.5	1272.1	
8/6/2014	1288.9	1311	1288.5	1306.9	1312.7	1301	1295.3	1289.3	1283.6	1277.6	1265.9	
8/7/2014	1306.6	1316.4	1303	1314.5	1347.1	1324.6	1315.8	1302.1	1293.3	1279.6	1257.1	
8/8/2014	1314.5	1324.3	1305.7	1310.6	1338.1	1324.7	1319.6	1311.3	1306.2	1297.9	1284.5	
8/11/2014	1310.4	1312.9	1306.4	1309.4	1350.7	1332.1	1321.4	1313.5	1302.8	1294.9	1276.3	
8/12/2014	1309.2	1319.3	1306.8	1310	1322.6	1316.1	1312.7	1309.6	1306.2	1303.1	1296.6	
8/13/2014	1309.6	1316.4	1306	1313.7	1337	1324.5	1317.3	1312	1304.8	1299.5	1287	
8/14/2014	1313.4	1321.8	1310	1313.9	1332.8	1322.4	1318.1	1312	1307.7	1301.6	1291.2	
8/15/2014	1313.4	1316.5	1293	1305.5	1338.8	1327	1320.5	1315.2	1308.7	1303.4	1291.6	
8/18/2014	1304.2	1304.9	1296.5	1298.2	1352	1328.5	1317	1305	1293.5	1281.5	1258	
8/19/2014	1298.6	1303.7	1294.7	1296.2	1316.7	1308.3	1303.2	1299.9	1294.8	1291.5	1283.1	
8/20/2014	1297.2	1299.3	1286.7	1292.5	1316.2	1307.2	1301.7	1298.2	1292.7	1289.2	1280.2	
8/21/2014	1291.9	1292	1273.4	1277.3	1314.7	1304.1	1298.3	1293.5	1287.7	1282.9	1272.3	
8/22/2014	1278.2	1283.9	1274.6	1281.8	1318.1	1293.5	1288.4	1280.9	1263.8	1262.3	1243.7	
8/25/2014	1280.8	1281.6	1275	1277	1298.7	1289.4	1285.6	1280.1	1276.3	1270.8	1261.5	
8/26/2014	1277.4	1291.9	1275.7	1281.4	1291.1	1284.5	1280.7	1277.9	1274.1	1271.3	1264.7	
8/27/2014	1281.6	1286.2	1280.9	1283.5	1315.4	1293.2	1290.3	1283	1274.1	1266.8	1250.6	
8/28/2014	1283.7	1297.6	1283	1290	1298.8	1291.5	1287.5	1284.2	1280.2	1276.9	1269.6	
8/29/2014	1290.4	1292.5	1284.1	1288	1319.4	1304.8	1297.4	1290.2	1282.8	1275.6	1261	
9/2/2014	1288.5	1288.8	1263.1	1266.5	1305	1296.6	1292.3	1288.2	1283.9	1279.8	1271.4	
9/3/2014	1266.7	1272.4	1261.9	1270	1324.2	1298.5	1282.5	1272.8	1266.8	1247.1	1221.4	
9/4/2014	1269.8	1279.2	1261.3	1261.7	1289.1	1278.6	1274.3	1268.1	1263.8	1257.6	1247.1	
9/5/2014	1261.7	1274.8	1258	1269.2	1303.2	1285.3	1273.5	1267.4	1255.6	1243.5	1231.6	
9/8/2014	1269.5	1272.6	1252.1	1256.3	1300.9	1284.1	1276.7	1267.3	1259.9	1250.5	1233.7	
9/9/2014	1257.1	1258.9	1248.1	1256.6	1301.3	1280.8	1268.6	1260.3	1248.1	1239.8	1219.3	
9/10/2014	1256.2	1258.5	1244.5	1249.8	1276.1	1265.3	1261	1254.5	1250.2	1243.7	1232.9	
9/11/2014	1249.8	1251	1235.3	1240.9	1278.9	1264.9	1257.4	1250.9	1243.4	1236.9	1222.9	
9/12/2014	1240.9	1242.3	1228.1	1229	1273.8	1258.1	1249.5	1242.4	1233.8	1226.7	1211	

Features

Date Open High Low Close R3 R2 R1
Pivot S1 S2 S3 sma5 sma10 sma20
sma30 sma40 sma50 sma100 sma200
mom5 mom6 mom7 mom8 mom9

Random Forests

A	B	C	D	E	F	G	H	I	J	K	L	M
Date	Open	High	Low	Close	R3	R2	R1	Pivot	S1	S2	S3	S
8/4/2014	1295.4	1296.4	1287	1289.2	1286.3	1308.9	1302.1	1291.5	1284.7	1274.1	1256.7	
8/5/2014	1289.2	1295	1283.3	1289.6	1309.7	1300.3	1294.7	1290.9	1285.3	1281.5	1272.1	
8/6/2014	1288.9	1311	1288.5	1306.9	1312.7	1301	1295.3	1289.3	1283.6	1277.6	1265.9	
8/7/2014	1306.6	1316.4	1303	1314.5	1347.1	1324.6	1315.8	1302.1	1293.3	1279.6	1257.1	
8/8/2014	1314.5	1324.3	1305.7	1310.6	1338.1	1324.7	1319.6	1311.3	1306.2	1297.9	1284.5	
8/11/2014	1310.4	1312.9	1306.4	1309.4	1350.7	1332.1	1321.4	1313.5	1302.8	1294.9	1276.3	
8/12/2014	1309.2	1319.3	1306.8	1310	1322.6	1316.1	1312.7	1309.6	1306.2	1303.1	1296.6	
8/13/2014	1309.6	1316.4	1306	1313.7	1337	1324.5	1317.3	1312	1304.8	1299.5	1287	
8/14/2014	1313.4	1321.8	1310	1313.9	1332.8	1322.4	1318.1	1312	1307.7	1301.6	1291.2	
8/15/2014	1313.4	1316.5	1293	1305.5	1338.8	1327	1320.5	1315.2	1308.7	1303.4	1291.6	
8/18/2014	1304.2	1304.9	1296.5	1298.2	1352	1328.5	1317	1305	1293.5	1281.5	1258	
8/19/2014	1298.6	1303.7	1294.7	1296.2	1316.7	1308.3	1303.2	1299.9	1294.8	1291.5	1283.1	
8/20/2014	1297.2	1299.3	1288.7	1292.5	1316.2	1307.2	1301.7	1298.2	1292.7	1289.2	1280.2	
8/21/2014	1291.9	1292	1273.4	1277.3	1314.7	1304.1	1298.3	1293.5	1287.7	1282.9	1272.3	
8/22/2014	1278.2	1283.9	1274.6	1281.8	1318.1	1293.5	1288.4	1280.9	1263.8	1262.3	1243.7	
8/25/2014	1280.8	1281.6	1275	1277	1298.7	1289.4	1285.6	1280.1	1276.3	1270.8	1261.5	
8/26/2014	1277.4	1291.9	1275.7	1281.4	1291.1	1284.5	1280.7	1277.9	1274.1	1271.3	1264.7	
8/27/2014	1281.6	1286.2	1260.9	1283.5	1315.4	1293.2	1290.3	1283	1274.1	1266.8	1250.6	
8/28/2014	1283.7	1297.6	1283	1290	1298.8	1291.5	1287.5	1284.2	1280.2	1276.9	1269.6	
8/29/2014	1290.4	1292.5	1284.1	1288	1319.4	1304.8	1297.4	1290.2	1282.8	1275.6	1261	
9/2/2014	1288.5	1288.8	1263.1	1266.5	1305	1296.6	1292.3	1288.2	1283.9	1279.8	1271.4	
9/3/2014	1266.7	1272.4	1261.9	1270	1324.2	1298.5	1282.5	1272.8	1266.8	1247.1	1221.4	
9/4/2014	1269.8	1279.2	1261.3	1261.7	1289.1	1278.6	1274.3	1268.1	1263.8	1257.6	1247.1	
9/5/2014	1261.7	1274.8	1258	1269.2	1303.2	1285.3	1273.5	1267.4	1255.6	1243.5	1231.6	
9/8/2014	1269.5	1272.6	1252.1	1256.3	1300.9	1284.1	1276.7	1267.3	1259.9	1250.5	1233.7	
9/9/2014	1257.1	1258.9	1248.1	1256.6	1301.3	1280.8	1268.6	1260.3	1248.1	1239.8	1219.3	
9/10/2014	1256.2	1258.5	1244.5	1249.8	1276.1	1265.3	1261	1254.5	1250.2	1243.7	1232.9	
9/11/2014	1249.8	1251	1235.3	1240.9	1278.9	1264.9	1257.4	1250.9	1243.4	1236.9	1222.9	
9/12/2014	1240.9	1242.3	1228.1	1229	1273.8	1258.1	1249.5	1242.4	1233.8	1226.7	1211	

Bag

A	B	C	D	E	F	G	H	I	J	K	L	M
Date	Open	High	Low	Close	R3	R2	R1	Pivot	S1	S2	S3	S
8/4/2014	1295.4	1296.4	1287	1289.2	1286.3	1308.9	1302.1	1291.5	1284.7	1274.1	1256.7	
8/5/2014	1289.2	1295	1283.3	1289.6	1309.7	1300.3	1294.7	1290.9	1285.3	1281.5	1272.1	
8/6/2014	1288.9	1311	1288.5	1306.9	1312.7	1301	1295.3	1289.3	1283.6	1277.6	1265.9	
8/7/2014	1306.6	1316.4	1303	1314.5	1347.1	1324.6	1315.8	1302.1	1293.3	1279.6	1257.1	
8/8/2014	1314.5	1324.3	1305.7	1310.6	1338.1	1324.7	1319.6	1311.3	1306.2	1297.9	1284.5	
8/11/2014	1310.4	1312.9	1306.4	1309.4	1350.7	1332.1	1321.4	1313.5	1302.8	1294.9	1276.3	
8/12/2014	1309.2	1319.3	1306.8	1310	1322.6	1316.1	1312.7	1309.6	1306.2	1303.1	1296.6	
8/13/2014	1309.6	1316.4	1306	1313.7	1337	1324.5	1317.3	1312	1304.8	1299.5	1287	
8/14/2014	1313.4	1321.8	1310	1313.9	1332.8	1322.4	1318.1	1312	1307.7	1301.6	1291.2	
8/15/2014	1313.4	1316.5	1293	1305.5	1338.8	1327	1320.5	1315.2	1308.7	1303.4	1291.6	
8/18/2014	1304.2	1304.9	1296.5	1298.2	1352	1328.5	1317	1305	1293.5	1281.5	1258	
8/19/2014	1298.6	1303.7	1294.7	1296.2	1316.7	1308.3	1303.2	1299.9	1294.8	1291.5	1283.1	
8/20/2014	1297.2	1299.3	1288.7	1292.5	1316.2	1307.2	1301.7	1298.2	1292.7	1289.2	1280.2	
8/21/2014	1291.9	1292	1273.4	1277.3	1314.7	1304.1	1298.3	1293.5	1287.7	1282.9	1272.3	
8/22/2014	1278.2	1283.9	1274.6	1281.8	1318.1	1293.5	1288.4	1280.9	1263.8	1262.3	1243.7	
8/25/2014	1280.8	1281.6	1275	1277	1298.7	1289.4	1285.6	1280.1	1276.3	1270.8	1261.5	
8/26/2014	1277.4	1291.9	1275.7	1281.4	1291.1	1284.5	1280.7	1277.9	1274.1	1271.3	1264.7	
8/27/2014	1281.6	1286.2	1260.9	1283.5	1315.4	1293.2	1290.3	1283	1274.1	1266.8	1250.6	
8/28/2014	1283.7	1297.6	1283	1290	1298.8	1291.5	1287.5	1284.2	1280.2	1276.9	1269.6	
8/29/2014	1290.4	1292.5	1284.1	1288	1319.4	1304.8	1297.4	1290.2	1282.8	1275.6	1261	
9/2/2014	1288.5	1288.8	1263.1	1266.5	1305	1296.6	1292.3	1288.2	1283.9	1279.8	1271.4	
9/3/2014	1266.7	1272.4	1261.9	1270	1324.2	1298.5	1282.5	1272.8	1266.8	1247.1	1221.4	
9/4/2014	1269.8	1279.2	1261.3	1261.7	1289.1	1278.6	1274.3	1268.1	1263.8	1257.6	1247.1	
9/5/2014	1261.7	1274.8	1258	1269.2	1303.2	1285.3	1273.5	1267.4	1255.6	1243.5	1231.6	
9/8/2014	1269.5	1272.6	1252.1	1256.3	1300.9	1284.1	1276.7	1267.3	1259.9	1250.5	1233.7	
9/9/2014	1257.1	1258.9	1248.1	1256.6	1301.3	1280.8	1268.6	1260.3	1248.1	1239.8	1219.3	
9/10/2014	1256.2	1258.5	1244.5	1249.8	1276.1	1265.3	1261	1254.5	1250.2	1243.7	1232.9	
9/11/2014	1249.8	1251	1235.3	1240.9	1278.9	1264.9	1257.4	1250.9	1243.4	1236.9	1222.9	
9/12/2014	1240.9	1242.3	1228.1	1229	1273.8	1258.1	1249.5	1242.4	1233.8	1226.7	1211	

Features

R3 R2 R1 Pivot S1 S2

Randomly select
m features

Date Open High Low Close R3 R2 R1
Pivot S1 S2 S3 sma5 sma10 sma20
sma30 sma40 sma50 sma100 sma200
mom5 mom6 mom7 mom8 mom9

Random Forests

A	B	C	D	E	F	G	H	I	J	K	L	M
Date	Open	High	Low	Close	R3	R2	R1	Pivot	S1	S2	S3	S
8/4/2014	1295.4	1296.4	1287	1289.2	1326.3	1308.9	1302.1	1291.5	1284.7	1274.1	1256.7	
8/5/2014	1289.2	1295	1283.3	1289.6	1309.7	1300.3	1294.7	1290.9	1285.3	1281.5	1272.1	
8/6/2014	1288.9	1311	1288.5	1306.9	1312.7	1301	1295.3	1289.3	1283.6	1277.6	1265.9	
8/7/2014	1306.6	1316.4	1303	1314.5	1347.1	1324.6	1315.8	1302.1	1293.3	1279.6	1257.1	
8/8/2014	1314.5	1324.3	1305.7	1310.6	1338.1	1324.7	1319.6	1311.3	1306.2	1297.9	1284.5	
8/11/2014	1310.4	1312.9	1306.4	1309.4	1350.7	1332.1	1321.4	1313.5	1302.8	1294.9	1276.3	
8/12/2014	1309.2	1319.3	1306.8	1310	1322.6	1316.1	1312.7	1309.6	1306.2	1303.1	1296.6	
8/13/2014	1309.6	1316.4	1306	1313.7	1337	1324.5	1317.3	1312	1304.8	1299.5	1287	
8/14/2014	1313.4	1321.8	1310	1313.9	1332.8	1322.4	1318.1	1312	1307.7	1301.6	1291.2	
8/15/2014	1313.4	1316.5	1293	1305.5	1338.8	1327	1320.5	1315.2	1308.7	1303.4	1291.6	
8/18/2014	1304.2	1304.9	1296.5	1298.2	1352	1328.5	1317	1305	1293.5	1281.5	1258	
8/19/2014	1298.6	1303.7	1294.7	1296.2	1316.7	1308.3	1303.2	1299.9	1294.8	1291.5	1283.1	
8/20/2014	1297.2	1299.3	1286.7	1292.5	1316.2	1307.2	1301.7	1298.2	1292.7	1289.2	1280.2	
8/21/2014	1291.9	1292	1273.4	1277.3	1314.7	1304.1	1298.3	1293.5	1287.7	1282.9	1272.3	
8/22/2014	1278.2	1283.9	1274.6	1281.8	1318.1	1293.5	1288.4	1280.9	1263.8	1262.3	1243.7	
8/25/2014	1280.8	1281.6	1275	1277	1298.7	1289.4	1285.6	1280.1	1276.3	1270.8	1261.5	
8/26/2014	1277.4	1291.9	1275.7	1281.4	1291.1	1284.5	1280.7	1277.9	1274.1	1271.3	1264.7	
8/27/2014	1281.6	1286.2	1260.9	1283.5	1315.4	1293.2	1290.3	1283	1274.1	1266.8	1250.6	
8/28/2014	1283.7	1297.6	1283	1290	1298.8	1291.5	1287.5	1284.2	1280.2	1276.9	1269.6	
8/29/2014	1290.4	1292.5	1284.1	1288	1319.4	1304.8	1297.4	1290.2	1282.8	1275.6	1261	
9/2/2014	1288.5	1288.8	1263.1	1266.5	1305	1296.6	1292.3	1288.2	1283.9	1279.8	1271.4	
9/3/2014	1266.7	1272.4	1261.9	1270	1324.2	1298.5	1292.5	1272.8	1266.8	1247.1	1221.4	
9/4/2014	1269.8	1279.2	1261.3	1261.7	1289.1	1278.6	1274.3	1268.1	1263.8	1257.6	1247.1	
9/5/2014	1261.7	1274.8	1258	1269.2	1303.2	1285.3	1273.5	1267.4	1255.6	1243.5	1231.6	
9/8/2014	1269.5	1272.6	1252.1	1256.3	1300.9	1284.1	1276.7	1267.3	1259.9	1250.5	1233.7	
9/9/2014	1257.1	1258.9	1248.1	1256.6	1301.3	1280.8	1268.6	1260.3	1248.1	1239.8	1219.3	
9/10/2014	1256.2	1258.5	1244.5	1249.8	1276.1	1265.3	1261	1254.5	1250.2	1243.7	1232.9	
9/11/2014	1249.8	1251	1235.3	1240.9	1278.9	1264.9	1257.4	1250.9	1243.4	1236.9	1222.9	
9/12/2014	1240.9	1242.3	1228.1	1229	1273.8	1258.1	1249.5	1242.4	1233.8	1226.7	1211	

Bag

A	B	C	D	E	F	G	H	I	J	K	L	M
Date	Open	High	Low	Close	R3	R2	R1	Pivot	S1	S2	S3	S
8/4/2014	1295.4	1296.4	1287	1289.2	1326.3	1308.9	1302.1	1291.5	1284.7	1274.1	1256.7	
8/5/2014	1289.2	1295	1283.3	1289.6	1309.7	1300.3	1294.7	1290.9	1285.3	1281.5	1272.1	
8/6/2014	1288.9	1311	1288.5	1306.9	1312.7	1301	1295.3	1289.3	1283.6	1277.6	1265.9	
8/7/2014	1306.6	1316.4	1303	1314.5	1347.1	1324.6	1315.8	1302.1	1293.3	1279.6	1257.1	
8/8/2014	1314.5	1324.3	1305.7	1310.6	1338.1	1324.7	1319.6	1311.3	1306.2	1297.9	1284.5	
8/11/2014	1310.4	1312.9	1306.4	1309.4	1350.7	1332.1	1321.4	1313.5	1302.8	1294.9	1276.3	
8/12/2014	1309.2	1319.3	1306.8	1310	1322.6	1316.1	1312.7	1309.6	1306.2	1303.1	1296.6	
8/13/2014	1309.6	1316.4	1306	1313.7	1337	1324.5	1317.3	1312	1304.8	1299.5	1287	
8/14/2014	1313.4	1321.8	1310	1313.9	1332.8	1322.4	1318.1	1312	1307.7	1301.6	1291.2	
8/15/2014	1313.4	1316.5	1293	1305.5	1338.8	1327	1320.5	1315.2	1308.7	1303.4	1291.6	
8/18/2014	1304.2	1304.9	1296.5	1298.2	1352	1328.5	1317	1305	1293.5	1281.5	1258	
8/19/2014	1298.6	1303.7	1294.7	1296.2	1316.7	1308.3	1303.2	1299.9	1294.8	1291.5	1283.1	
8/20/2014	1297.2	1299.3	1286.7	1292.5	1316.2	1307.2	1301.7	1298.2	1292.7	1289.2	1280.2	
8/21/2014	1291.9	1292	1273.4	1277.3	1314.7	1304.1	1298.3	1293.5	1287.7	1282.9	1272.3	
8/22/2014	1278.2	1283.9	1274.6	1281.8	1318.1	1293.5	1288.4	1280.9	1263.8	1262.3	1243.7	
8/25/2014	1280.8	1281.6	1275	1277	1298.7	1289.4	1285.6	1280.1	1276.3	1270.8	1261.5	
8/26/2014	1277.4	1291.9	1275.7	1281.4	1291.1	1284.5	1280.7	1277.9	1274.1	1271.3	1264.7	
8/27/2014	1281.6	1286.2	1260.9	1283.5	1315.4	1293.2	1290.3	1283	1274.1	1266.8	1250.6	
8/28/2014	1283.7	1297.6	1283	1290	1298.8	1291.5	1287.5	1284.2	1280.2	1276.9	1269.6	
8/29/2014	1290.4	1292.5	1284.1	1288	1319.4	1304.8	1297.4	1290.2	1282.8	1275.6	1261	
9/2/2014	1288.5	1288.8	1263.1	1266.5	1305	1296.6	1292.3	1288.2	1283.9	1279.8	1271.4	
9/3/2014	1266.7	1272.4	1261.9	1270	1324.2	1298.5	1292.5	1272.8	1266.8	1247.1	1221.4	
9/4/2014	1269.8	1279.2	1261.3	1261.7	1289.1	1278.6	1274.3	1268.1	1263.8	1257.6	1247.1	
9/5/2014	1261.7	1274.8	1258	1269.2	1303.2	1285.3	1273.5	1267.4	1255.6	1243.5	1231.6	
9/8/2014	1269.5	1272.6	1252.1	1256.3	1300.9	1284.1	1276.7	1267.3	1259.9	1250.5	1233.7	
9/9/2014	1257.1	1258.9	1248.1	1256.6	1301.3	1280.8	1268.6	1260.3	1248.1	1239.8	1219.3	
9/10/2014	1256.2	1258.5	1244.5	1249.8	1276.1	1265.3	1261	1254.5	1250.2	1243.7	1232.9	
9/11/2014	1249.8	1251	1235.3	1240.9	1278.9	1264.9	1257.4	1250.9	1243.4	1236.9	1222.9	
9/12/2014	1240.9	1242.3	1228.1	1229	1273.8	1258.1	1249.5	1242.4	1233.8	1226.7	1211	

Features

Identify Best Split

R3 R2 R1 Pivot S1 S2

Randomly select
m features

Date Open High Low Close R3 R2 R1
Pivot S1 S2 S3 sma5 sma10 sma20
sma30 sma40 sma50 sma100 sma200
mom5 mom6 mom7 mom8 mom9

Random Forests

A	B	C	D	E	F	G	H	I	J	K	L	M
Date	Open	High	Low	Close	R3	R2	R1	Pivot	S1	S2	S3	S
8/4/2014	1295.4	1296.4	1287	1289.2	1326.3	1308.9	1302.1	1291.5	1284.7	1274.1	1256.7	
8/5/2014	1289.2	1295	1283.3	1289.6	1309.7	1300.3	1294.7	1290.9	1285.3	1281.5	1272.1	
8/6/2014	1288.9	1311	1288.5	1306.9	1312.7	1301	1295.3	1289.3	1283.6	1277.6	1265.9	
8/7/2014	1306.6	1316.4	1303	1314.5	1347.1	1324.6	1315.8	1302.1	1293.3	1279.6	1257.1	
8/8/2014	1314.5	1324.3	1305.7	1310.6	1338.1	1324.7	1319.6	1311.3	1306.2	1297.9	1284.5	
8/11/2014	1310.4	1312.9	1306.4	1309.4	1350.7	1332.1	1321.4	1313.5	1302.8	1294.9	1276.3	
8/12/2014	1309.2	1319.3	1306.8	1310	1322.6	1316.1	1312.7	1309.6	1306.2	1303.1	1296.6	
8/13/2014	1309.6	1316.4	1306	1313.7	1337	1324.5	1317.3	1312	1304.8	1299.5	1287	
8/14/2014	1313.4	1321.8	1310	1313.9	1332.8	1322.4	1318.1	1312	1307.7	1301.6	1291.2	
8/15/2014	1313.4	1316.5	1293	1305.5	1338.8	1327	1320.5	1315.2	1308.7	1303.4	1291.6	
8/18/2014	1304.2	1304.9	1296.5	1298.2	1352	1328.5	1317	1305	1293.5	1281.5	1258	
8/19/2014	1298.6	1303.7	1294.7	1296.2	1316.7	1308.3	1303.2	1299.9	1294.8	1291.5	1283.1	
8/20/2014	1297.2	1299.3	1286.7	1292.5	1316.2	1307.2	1301.7	1298.2	1292.7	1289.2	1280.2	
8/21/2014	1291.9	1292	1273.4	1277.3	1314.7	1304.1	1298.3	1293.5	1287.7	1282.9	1272.3	
8/22/2014	1278.2	1283.9	1274.6	1281.8	1318.1	1293.5	1288.4	1280.9	1263.8	1262.3	1243.7	
8/25/2014	1280.8	1281.6	1275	1277	1298.7	1289.4	1285.6	1280.1	1276.3	1270.8	1261.5	
8/26/2014	1277.4	1291.9	1275.7	1281.4	1291.1	1284.5	1280.7	1277.9	1274.1	1271.3	1264.7	
8/27/2014	1281.6	1286.2	1280.9	1283.5	1315.4	1293.2	1290.3	1283	1274.1	1266.8	1250.6	
8/28/2014	1283.7	1297.6	1283	1290	1298.8	1291.5	1287.5	1284.2	1280.2	1276.9	1269.6	
8/29/2014	1290.4	1292.5	1284.1	1288	1319.4	1304.8	1297.4	1290.2	1282.8	1275.6	1261	
9/2/2014	1288.5	1288.8	1263.1	1266.5	1305	1296.6	1292.3	1288.2	1283.9	1279.8	1271.4	
9/3/2014	1266.7	1272.4	1261.9	1270	1324.2	1298.5	1282.5	1272.8	1266.8	1247.1	1221.4	
9/4/2014	1269.8	1279.2	1261.3	1261.7	1289.1	1278.6	1274.3	1268.1	1263.8	1257.6	1247.1	
9/5/2014	1261.7	1274.8	1258	1269.2	1303.2	1285.3	1273.5	1267.4	1255.6	1243.5	1231.6	
9/8/2014	1269.5	1272.6	1252.1	1256.3	1300.9	1284.1	1276.7	1267.3	1259.9	1250.5	1233.7	
9/9/2014	1257.1	1258.9	1248.1	1256.6	1301.3	1280.8	1268.6	1260.3	1248.1	1239.8	1219.3	
9/10/2014	1256.2	1258.5	1244.5	1249.8	1276.1	1265.3	1261	1254.5	1250.2	1243.7	1232.9	
9/11/2014	1249.8	1251	1235.3	1240.9	1278.9	1264.9	1257.4	1250.9	1243.4	1236.9	1222.9	
9/12/2014	1240.9	1242.3	1228.1	1229	1273.8	1258.1	1249.5	1242.4	1233.8	1226.7	1211	

Bag

Date	Open	High	Low	Close	R3	R2	R1	Pivot	S1	S2	S3	S
8/4/2014	1295.4	1296.4	1287	1289.2	1326.3	1308.9	1302.1	1291.5	1284.7	1274.1	1256.7	
8/5/2014	1289.2	1295	1283.3	1289.6	1309.7	1300.3	1294.7	1290.9	1285.3	1281.5	1272.1	
8/6/2014	1288.9	1311	1288.5	1306.9	1312.7	1301	1295.3	1289.3	1283.6	1277.6	1265.9	
8/7/2014	1306.6	1316.4	1303	1314.5	1347.1	1324.6	1315.8	1302.1	1293.3	1279.6	1257.1	
8/8/2014	1314.5	1324.3	1305.7	1310.6	1338.1	1324.7	1319.6	1311.3	1306.2	1297.9	1284.5	
8/11/2014	1310.4	1312.9	1306.4	1309.4	1350.7	1332.1	1321.4	1313.5	1302.8	1294.9	1276.3	
8/12/2014	1309.2	1319.3	1306.8	1310	1322.6	1316.1	1312.7	1309.6	1306.2	1303.1	1296.6	
8/13/2014	1309.6	1316.4	1306	1313.7	1337	1324.5	1317.3	1312	1304.8	1299.5	1287	
8/14/2014	1313.4	1321.8	1310	1313.9	1332.8	1322.4	1318.1	1312	1307.7	1301.6	1291.2	
8/15/2014	1313.4	1316.5	1293	1305.5	1338.8	1327	1320.5	1315.2	1308.7	1303.4	1291.6	
8/18/2014	1304.2	1304.9	1296.5	1298.2	1352	1328.5	1317	1305	1293.5	1281.5	1258	
8/19/2014	1298.6	1303.7	1294.7	1296.2	1316.7	1308.3	1303.2	1299.9	1294.8	1291.5	1283.1	
8/20/2014	1297.2	1299.3	1286.7	1292.5	1316.2	1307.2	1301.7	1298.2	1292.7	1289.2	1280.2	
8/21/2014	1291.9	1292	1273.4	1277.3	1314.7	1304.1	1298.3	1293.5	1287.7	1282.9	1272.3	
8/22/2014	1278.2	1283.9	1274.6	1281.8	1318.1	1293.5	1288.4	1280.9	1263.8	1262.3	1243.7	
8/25/2014	1280.8	1281.6	1275	1277	1298.7	1289.4	1285.6	1280.1	1276.3	1270.8	1261.5	
8/26/2014	1277.4	1291.9	1275.7	1281.4	1291.1	1284.5	1280.7	1277.9	1274.1	1271.3	1264.7	
8/27/2014	1281.6	1286.2	1280.9	1283.5	1315.4	1293.2	1290.3	1283	1274.1	1266.8	1250.6	
8/28/2014	1283.7	1297.6	1283	1290	1298.8	1291.5	1287.5	1284.2	1280.2	1276.9	1269.6	
8/29/2014	1290.4	1292.5	1284.1	1288	1319.4	1304.8	1297.4	1290.2	1282.8	1275.6	1261	
9/2/2014	1288.5	1288.8	1263.1	1266.5	1305	1296.6	1292.3	1288.2	1283.9	1279.8	1271.4	
9/3/2014	1266.7	1272.4	1261.9	1270	1324.2	1298.5	1282.5	1272.8	1266.8	1247.1	1221.4	
9/4/2014	1269.8	1279.2	1261.3	1261.7	1289.1	1278.6	1274.3	1268.1	1263.8	1257.6	1247.1	
9/5/2014	1261.7	1274.8	1258	1269.2	1303.2	1285.3	1273.5	1267.4	1255.6	1243.5	1231.6	
9/8/2014	1269.5	1272.6	1252.1	1256.3	1300.9	1284.1	1276.7	1267.3	1259.9	1250.5	1233.7	
9/9/2014	1257.1	1258.9	1248.1	1256.6	1301.3	1280.8	1268.6	1260.3	1248.1	1239.8	1219.3	
9/10/2014	1256.2	1258.5	1244.5	1249.8	1276.1	1265.3	1261	1254.5	1250.2	1243.7	1232.9	
9/11/2014	1249.8	1251	1235.3	1240.9	1278.9	1264.9	1257.4	1250.9	1243.4	1236.9	1222.9	
9/12/2014	1240.9	1242.3	1228.1	1229	1273.8	1258.1	1249.5	1242.4	1233.8	1226.7	1211	

Features

Identify Best Split

R3 R2 R1 Pivot S1 S2

Randomly select
m features

Date Open High Low Close R3 R2 R1
Pivot S1 S2 S3 sma5 sma10 sma20
sma30 sma40 sma50 sma100 sma200
mom5 mom6 mom7 mom8 mom9

Add to relevant Node

Methodology Of Random Forests

- Create 'BAGS' of samples
 - Build the first tree
 - Randomly select m predictors $<$ total number of predictors
 - Find best j,s
 - Take fresh sample of m predictors
 - Find best j,s
 - Repeat till stopping criteria met
 - Build N such trees
-

Random Forests

- Forcing trees to learn from seemingly less important predictors reduces over fit
- Aggregating Over uncorrelated learners reduces variance
- Guidelines for selecting m
 - Classification = \sqrt{p}
 - Regression = $p/3$
- OOB error reliable estimate of test error!!!
- Variable Importance “drop in the error function “ for each predictor averaged across all trees and then Ranked

Gradient Boosted Trees

- Boost the performance of the collective ensemble based on the previous models' misclassification
- It is an incremental model, so parallel computation cannot be done

Boosting

- Forward stage wise additive modelling process

Set $f_m(x) = f_{m-1}(x) + \beta_m b(x; \gamma_m)$.

$$\min_{\{\beta_m, \gamma_m\}_1^M} \sum_{i=1}^N L \left(y_i, \sum_{m=1}^M \beta_m b(x_i; \gamma_m) \right).$$

$$L(y, f(x)) = (y - f(x))^2,$$

$$\begin{aligned} L(y_i, f_{m-1}(x_i) + \beta b(x_i; \gamma)) &= (y_i - f_{m-1}(x_i) - \beta b(x_i; \gamma))^2 \\ &= (r_{im} - \beta b(x_i; \gamma))^2, \end{aligned}$$

Methodology of Gradient Boosted Trees

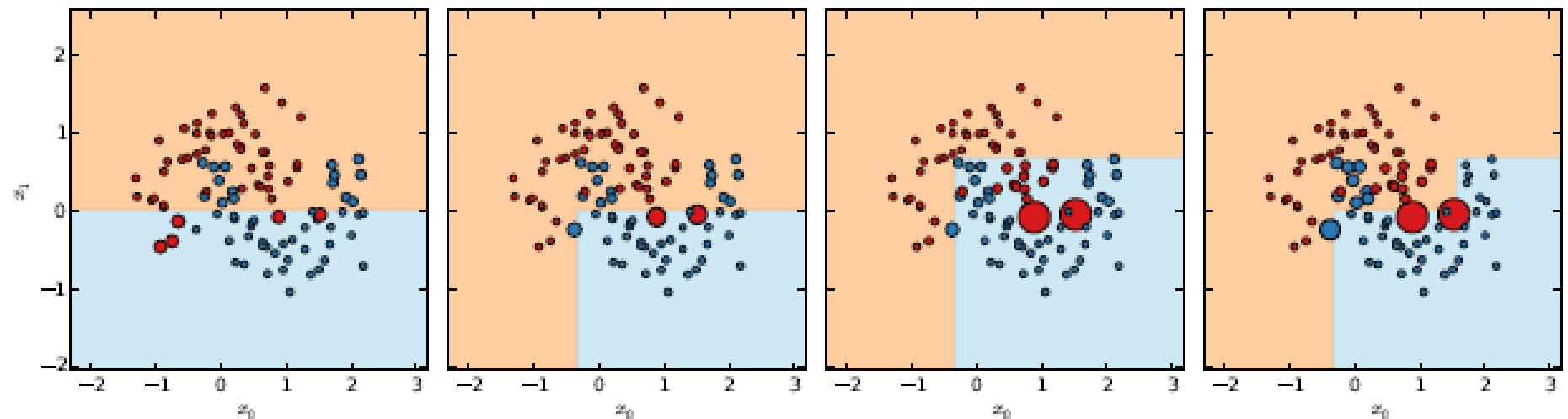
- Important Parameters
 - Number Of trees
 - Learning rate
 - Depth / Number of terminal nodes
 - Regularization Parameters : Gamma , L2, L1
-

Methodology of Gradient Boosted Trees

- Fit a tree (T1) to the training data set
- Calculate the residuals
- Fit a tree T2 to the residuals of T1
- Update Model as $T1 + \text{shrinkage} * T2$
- Fit a model T3 to the residuals of T2
- Update Model as $T1 + \text{shrinkage} * T2 + \text{shrinkage} * T3$
- .
- .
- Fit a model TN to the residuals of TN-1

Gradient Boosted Trees

- Iterative learning of training samples based on errors



Lab 3a

- Implement XGBoost and Random Forest algorithms on the given data set

Lab 3b

- Implement XGBoost and Random Forest algorithms on the given data set

Support Vector Machines

Overview

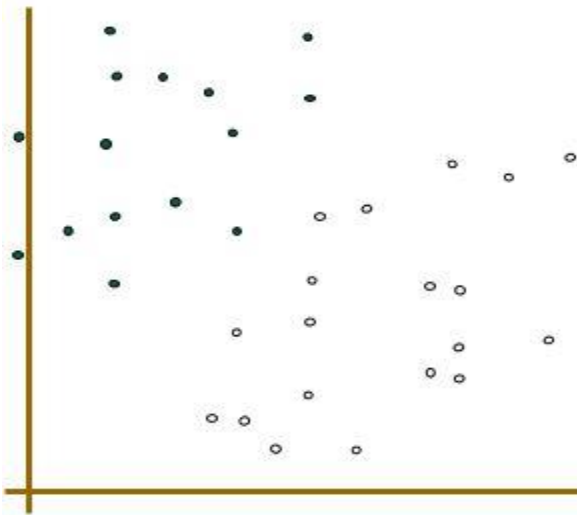
- Introduction
 - Hyperplane
 - Maximum Marginal Hyperplane
 - Soft Margin Hyperplane
 - Non-Linear SVMs
-

Introduction

- Supervised learning algorithm for classification and regression problems
 - Exploits the linear or non-linear **separability** of data points using a hyperplane
 - Gives us the optimal hyperplane which can be used for categorizing new examples
-

Hyperplane

- Mathematically, it is defined as a ' $(p-1)$ ' dimensional flat affine that can separate ' p ' dimensional data
- Taking an example of 2-dimensional data



- Class 1
- Class 2

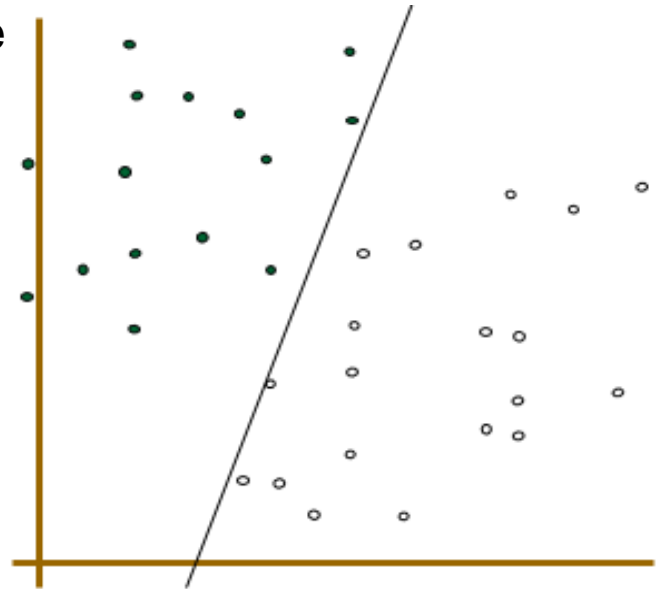
Hyperplane

- Here $p=2$, so a hyperplane of $p-1=1$ dimensions can be drawn through the dataset
- A subspace affine of 1 dimension is a line

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$$

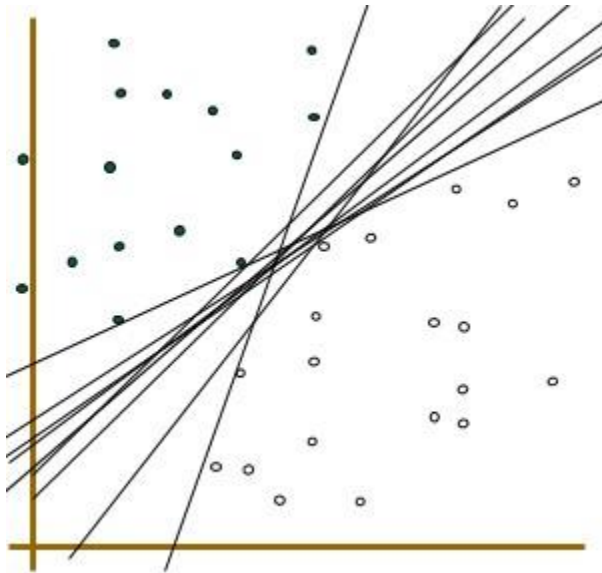
$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p = 0$$

- Sign Of the above tells us class



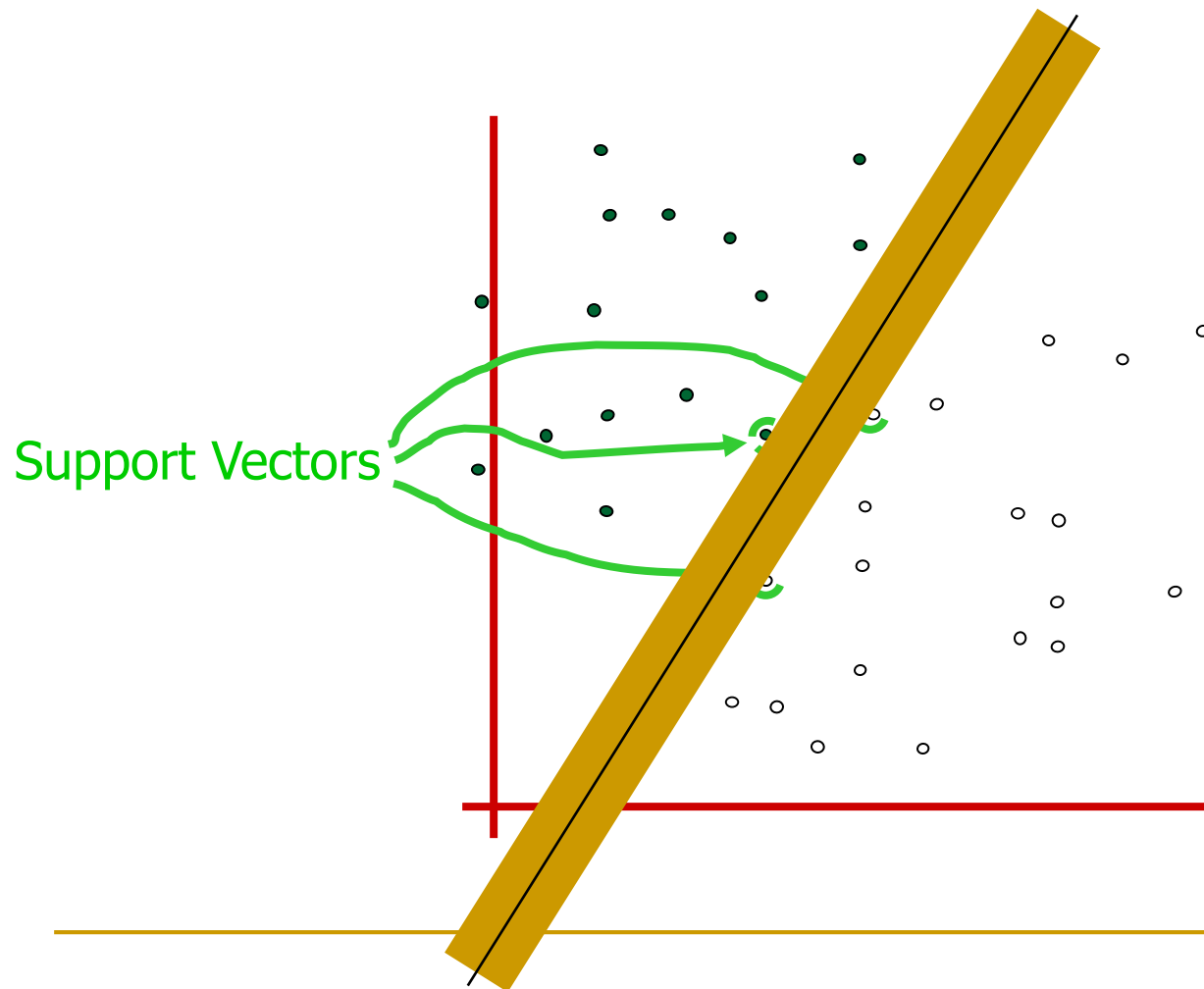
Hyperplane

- Similarly, the concept can be extended to higher dimensional data
- But the question arises, which line to choose



- Any of these would be fine..
- ..but which is best?

Maximal Margin Hyperplane



- This is the simplest kind of SVM (Called an LSVM)

Maximal Margin Hyperplane

- From the set of all possible hyperplanes, we select the hyperplane whose margin is the highest, where margin is defined as the width that the boundary could be increased by before hitting a data point.
 - Intuitively, the maximal margin hyperplane represents the mid-line of the widest “slab” that we can insert between the two classes
 - The points which are at the boundary of the margin are called ‘support vectors’ because these points are the ones which ‘support’ the margin
-

Maximal Margin Hyperplane

$$\underset{\beta_0, \beta_1, \dots, \beta_p}{\text{maximize}} \ M$$

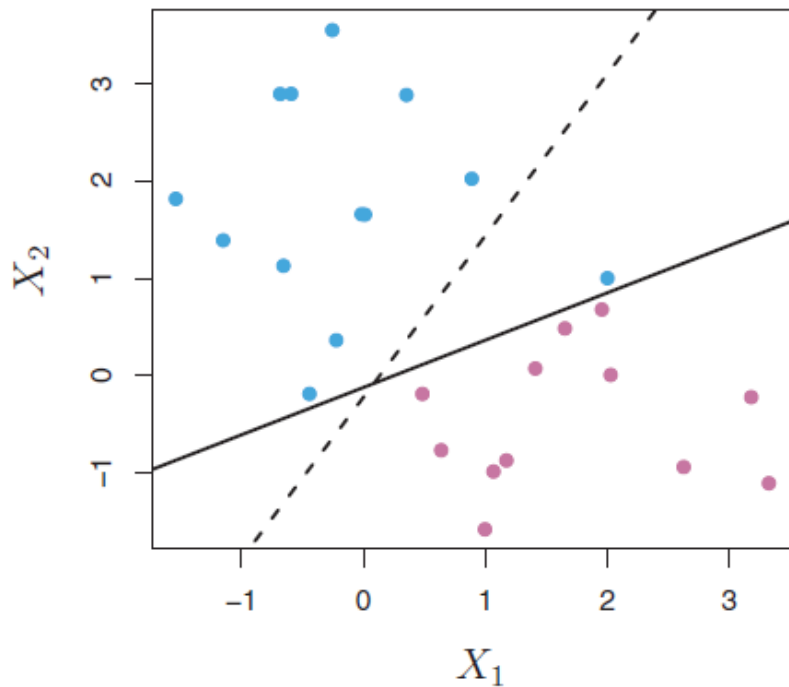
$$\text{subject to } \sum_{j=1}^p \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}) \geq M \quad \forall i = 1, \dots, n.$$

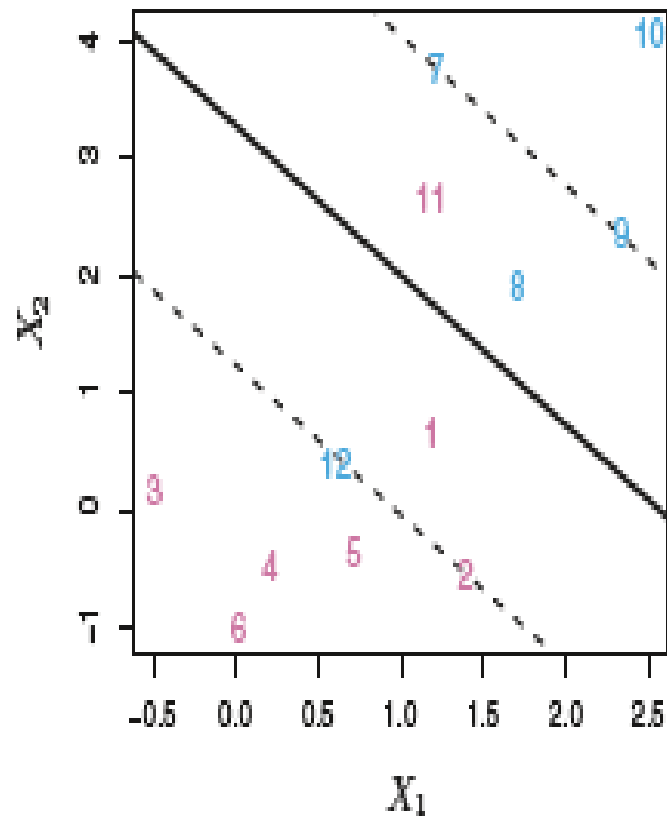
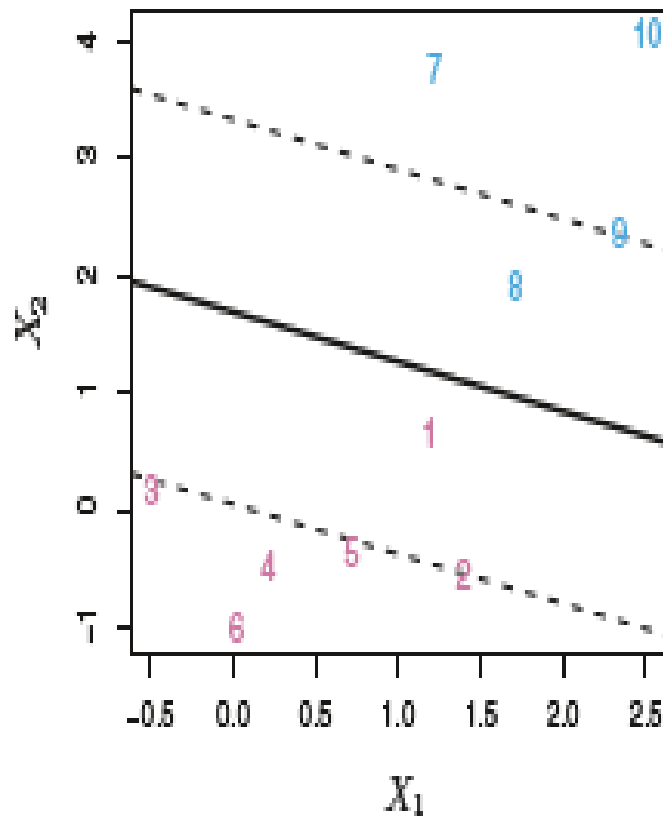
$$f(x) = \beta_0 + \sum_{i=1}^n \alpha_i \langle x, x_i \rangle,$$

Datasets with Noise

- A little bit noise can cause a significant change in the classifier



Soft Margin Classifier



Soft Margin Classifier

$$\underset{\beta_0, \beta_1, \dots, \beta_p, \epsilon_1, \dots, \epsilon_n}{\text{maximize}} \quad M$$

$$\text{subject to} \quad \sum_{j=1}^p \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

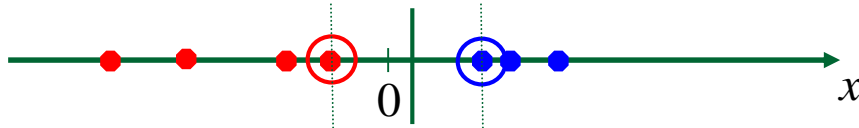
$$\epsilon_i \geq 0, \quad \sum_{i=1}^n \epsilon_i \leq C,$$

Soft Margin Classifier

- To capture the underlying nature of the data, we allow misclassification at the training stage and allow for this rate to be less than a budget 'C'
- In practise we control the C as a tuning parameter and determine its value by cross validation tests
- Only support vectors affect the solutions. Samples really far away from the classifier do not affect the solution
- If these support vectors are quite few in number what does it tell you about the classification ?

Non-linear SVMs

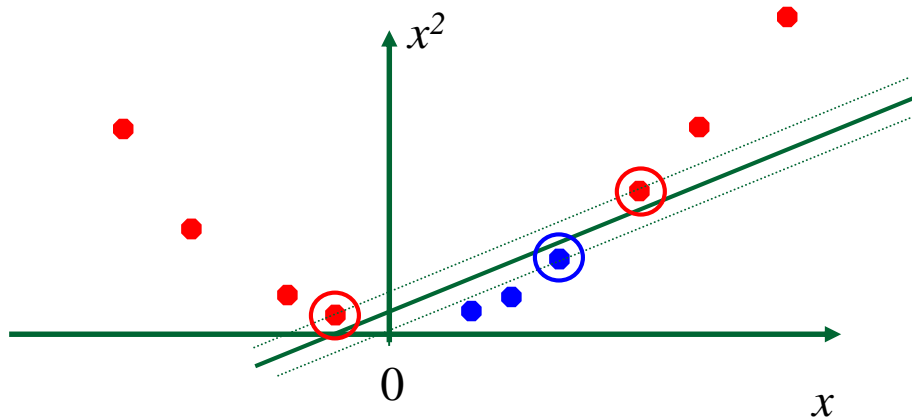
- Datasets that are linearly separable with some noise work out great:



- But what are we going to do if the dataset is just too hard?

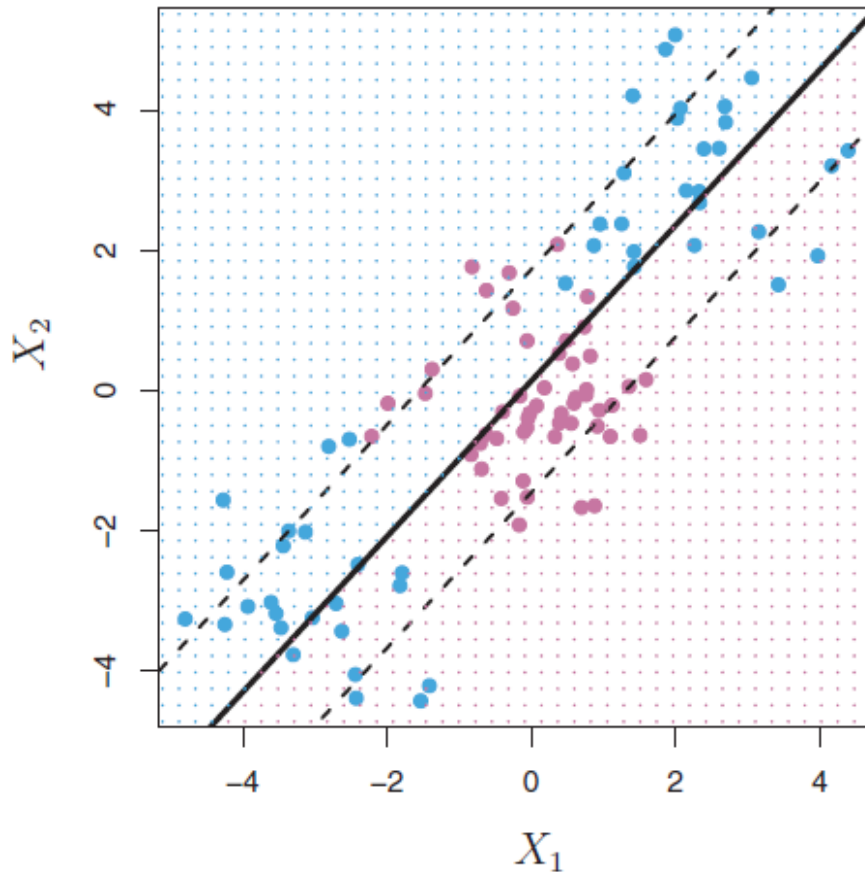


- The solution is mapping to a higher dimensional space



Non-Linear SVMs

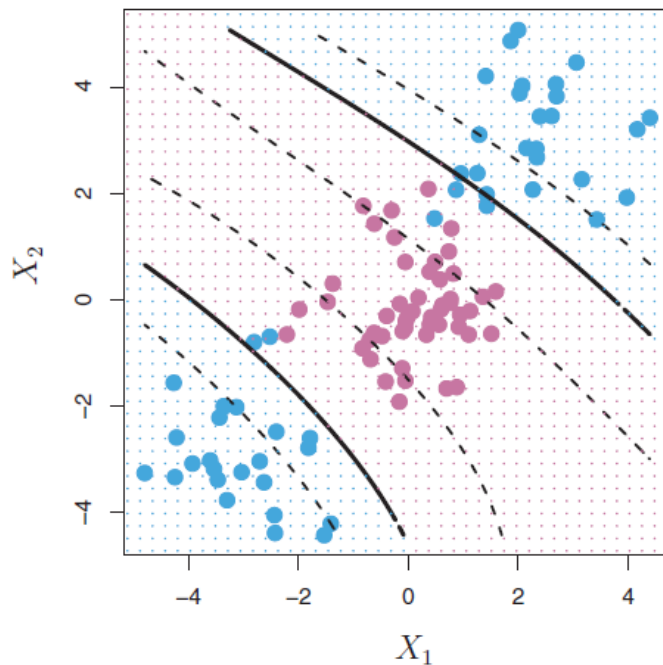
■ Looking at the data below



- The data clearly has non linear separability but the linear SVM is unable to capture the separation, mapping to higher dimensional feature-spaces is essential

Non-Linear SVMs

- So, instead of building SVMs with just X_1 and X_2 we include higher dimensional data i.e. X_{12} and X_{22} , resulting in a better separability of data



- Even higher degrees can be used for the capturing the non-linearity

Non-Linear SVMs

$$\begin{aligned} & \underset{\beta_0, \beta_{11}, \beta_{12}, \dots, \beta_{p1}, \beta_{p2}, \epsilon_1, \dots, \epsilon_n}{\text{maximize}} && M \\ & \text{subject to} && y_i \left(\beta_0 + \sum_{j=1}^p \beta_{j1} x_{ij} + \sum_{j=1}^p \beta_{j2} x_{ij}^2 \right) \geq M(1 - \epsilon_i), \\ & && \sum_{i=1}^n \epsilon_i \leq C, \quad \epsilon_i \geq 0, \quad \sum_{j=1}^p \sum_{k=1}^2 \beta_{jk}^2 = 1. \end{aligned}$$

Non-Linear SVMs

- The idea of non-linear SVMs is not to increase the features in the dataset itself but to map the low dimensional feature sets to higher dimensional feature sets using the kernel trick.
 - The kernel function plays the role of the dot product in the feature space.
-

The Solution

$$f(x) = \beta_0 + \sum_{i=1}^n \alpha_i \langle x, x_i \rangle,$$

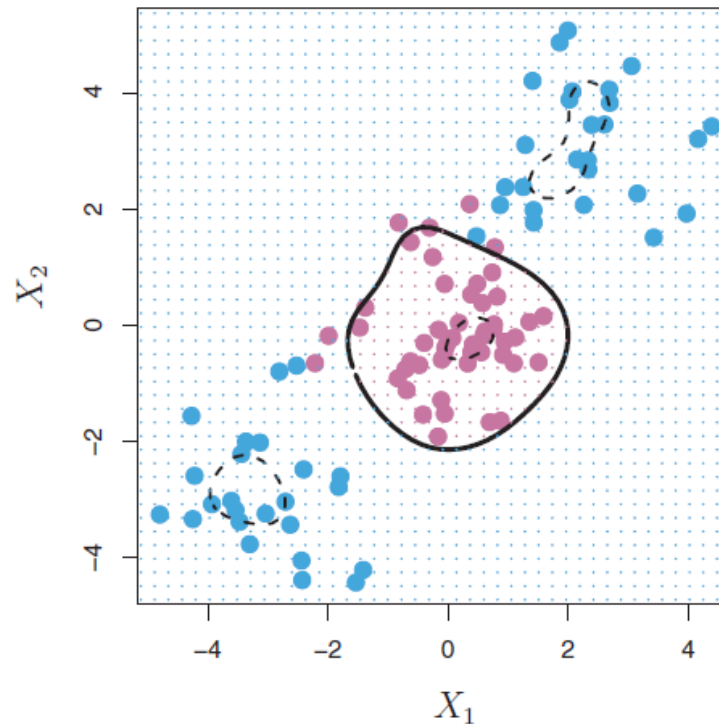
- Coefficients are non zero only for support vectors i.e samples that lie ON the margin or on the wrong side

Examples of Kernel Functions

- Linear: $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j$
- Polynomial of power p : $K(\mathbf{x}_i, \mathbf{x}_j) = (1 + \mathbf{x}_i^T \mathbf{x}_j)^p$
- Gaussian (radial-basis function network):
$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$$
- Sigmoid: $K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\beta_0 \mathbf{x}_i^T \mathbf{x}_j + \beta_1)$

Radial Kernel

- Using radial basis function for the above data set



Lab 4

- Implement Support vector machines on a radial kernel and comment on the results