Choosing Decision Making Techniques

Presenter: Santosh Shet
01JST16CS132

Overview Type Of Classifiers

- Neural Network
- Logistic Regression (Predictive Learning Model)
- Nearest Neighbor
- Decision Trees

What Influences Choosing A Classifier?

Choosing a classifier

The first step in choosing a classifier is to closely study the training data!

Why studying the data is important?

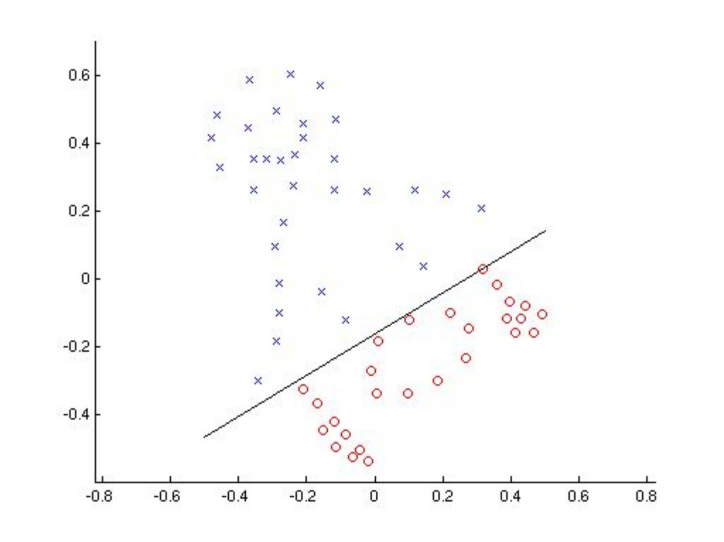
1. Check for various type of features that many be required to be considered
2. Dataset maybe taken from different sources/distribution
3. Check for class imbalance
4. Check for unusual range values that affects training

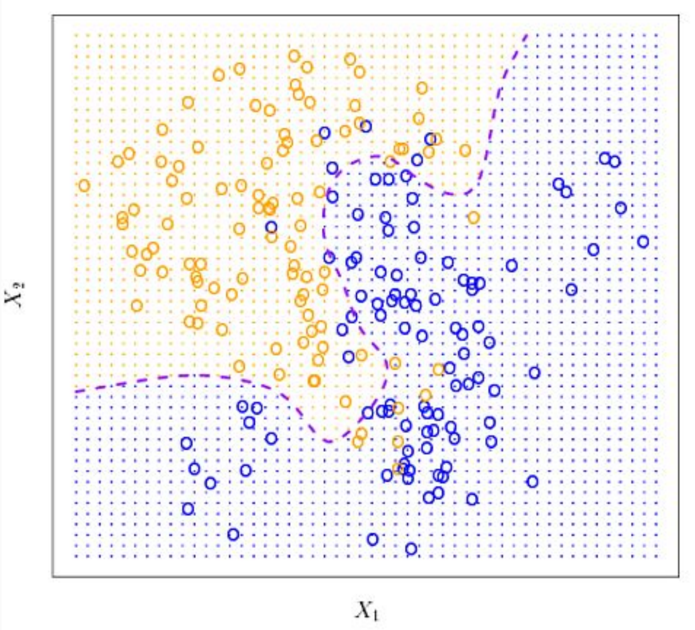
Choosing a classifier

Check for various type of features that many be required to be considered

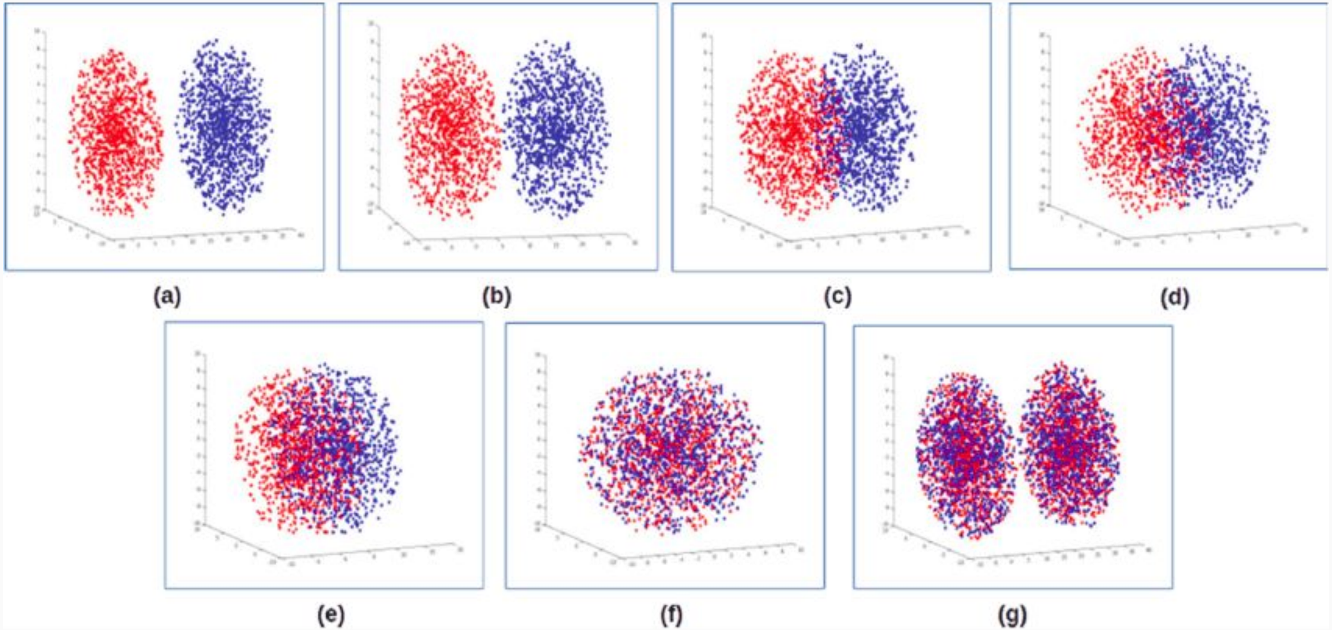
- Create individual class histograms of each available features.
- A two dimensional scatterplots for pairs of the best single features can be formed to study the shape and locations of the classes and their degree of overlap.

Decision Boundary



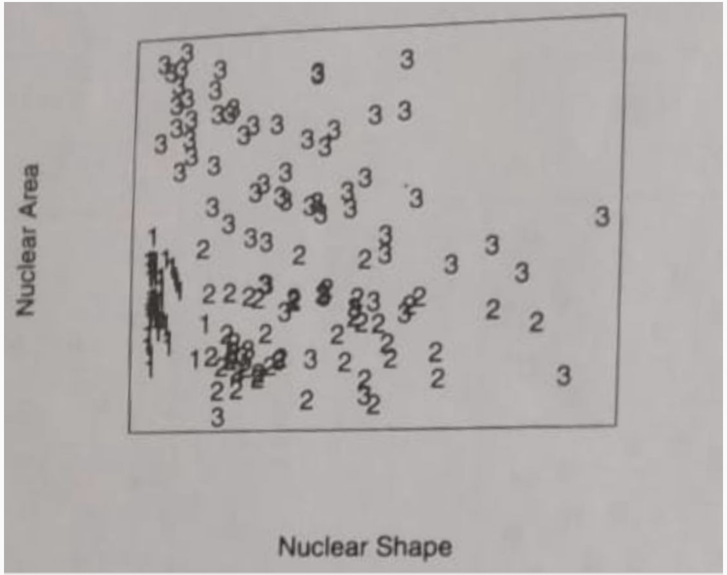


Continued..



Scatter Plots showing 3D Data: (a) Non-overlapping; (b) Barely touching; Overlap of: (c) <25%; (d) 50%; (e) 75%; (f) Fully Overlapping; (g) Random class labels.

An Example



Scatter plot for real data on three classes of leukocytes based on nucleus area and nuclear shape

Considerable overlap between the classes and too many features to easily visualize d-dimensional space, 2d scatterplots can be used to see if features are normally distributed.

We can consider non linear transformation so that it is normally distributed. Ex: Using log or square root

This may not work for multivariate normal for each class.

What Is The Optimal Number Of Features?

Optimal Number of Features

An ideal model should do justice to both: good prediction yet not overly complex to interpret & use

We can add other features that are that are not correlated with the ones we already have. A precaution should be taken not to reduce the performance by adding such “**noisy features**”

One way to do is to select the best features:

- Subset Selection
- Forward Stepwise Selection
- Backward Stepwise Selection

Subset Selection

Requires massive computation power!

Fit models with each possible combinations of n features.

Total number of models 2^p

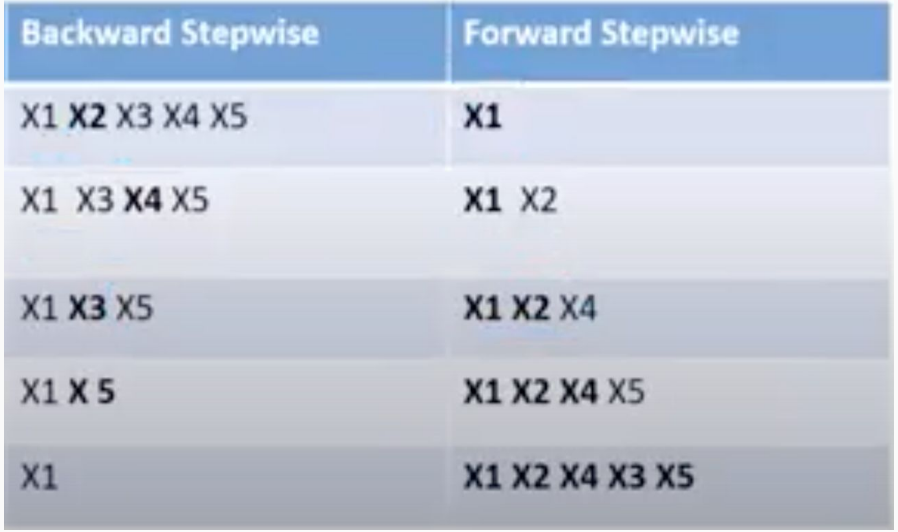
This technique can be broken in two stages:

Stage 1: Fit all combinations of models that has only k features out of n. Pick the best model from the set of all k predictions models (call this Model(k))

Stage 2: Select the one that is best from Model(1), Model(2) ,... , Model(n)

Forward & Backward Stepwise Selection

Feature set = {X1,X2,X3,X4,X5}



Backward Stepwise	Forward Stepwise
X1 **X2** X3 X4 X5	**X1**
X1 X3 **X4** X5	**X1** X2
X1 **X3** X5	**X1 X2** X4
X1 **X 5**	**X1 X2 X4** X5
X1	**X1 X2 X4 X3 X5**

Thank You