Classification

Adeeb Mahmood

September 4, 2020

Contents

In this presentation I will discuss:

- Introduction To Classification.
- Different Methods Of Classification
- Batch Learning Vs Incremental Learning
- Class Imbalance

Introduction

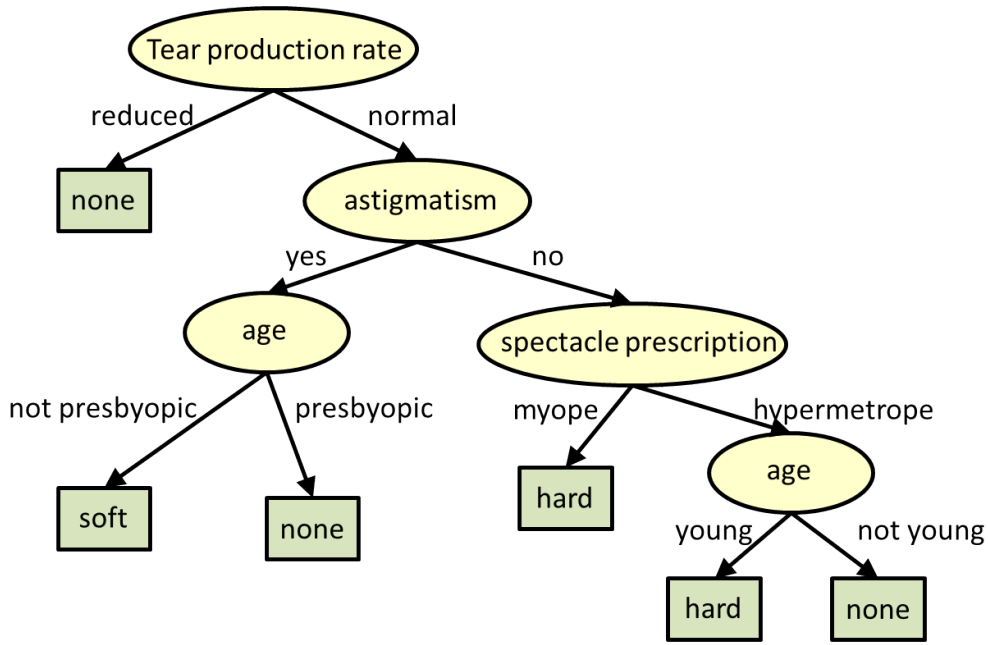
- To start lets talk about what classification is.
- Classification is a means of taking some input data and assigning a label to them.
- We have training data to make an algorithm then we input the data we actually care about.
- Here's an example revolving around irises a type of flower.

Sepal Length(cm)	Sepal Width(cm)
5.1	3.5
4.9	3.0
4.7	3.2
4.6	3.1
5.0	3.6

- They could be a 'setosa', 'versicolor' or a 'virginica'

Decision Tree Classifier

So the basic idea is that we ask a question and depending on the answer we split the data up into smaller subsets. So we start with the entire data set and it is partitioned into smaller and smaller data sets as we ask more and more questions



A decision tree to determine what kind of contact lens a person may wear.

Decision Tree Classifier

Entropy

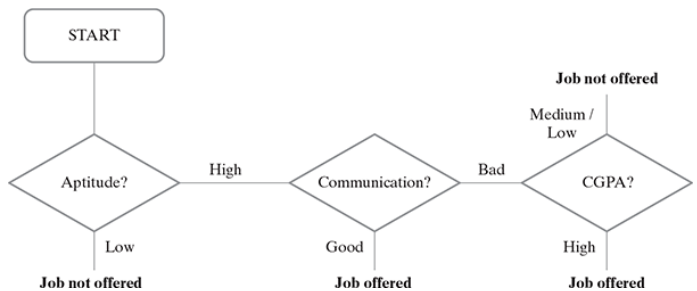
$$E(S) = \sum_{i=1}^{c} -p_i log_2 p_i$$

- Now we need to introduce information gain
  Information Gain(S, A) = Entropy(S_{bs}) − Entropy(S_{as})
- Finally all we need is: Entropy$(S_{as}) = \sum_{i=1}^{n} w_i$Entropy$(p_i)$

key: $E(S)$ - Entropy
p - refers to the proportion of values falling into the i-th class label.
c - the number of different class labels
w_i - proportion of examples falling into each partition

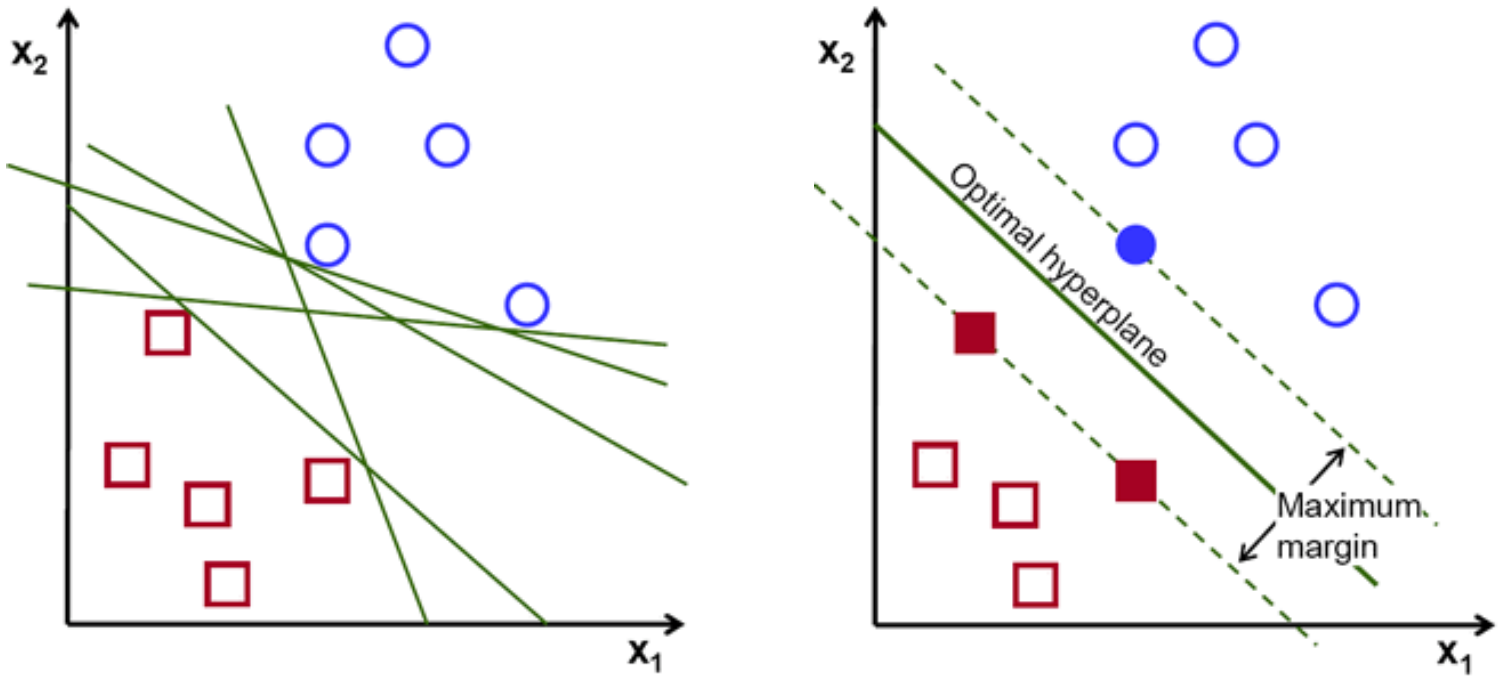
Decision Tree Classifier

CGPA	Communication	Aptitude	Programming Skill	Job offered?
High	Good	High	Good	Yes
Medium	Good	High	Good	Yes
Low	Bad	Low	Good	No
Low	Good	Low	Bad	No
High	Good	High	Bad	Yes
High	Good	High	Good	Yes
Medium	Bad	Low	Bad	No
Medium	Bad	Low	Good	No
High	Bad	High	Good	Yes
Medium	Good	High	Good	Yes
Low	Bad	High	Bad	No
Low	Bad	High	Bad	No
Medium	Good	High	Bad	Yes
Low	Good	Low	Good	No
High	Bad	Low	Bad	No
Medium	Bad	High	Good	No
High	Bad	Low	Bad	No
Medium	Good	High	Bad	Yes



Support Vector Machine

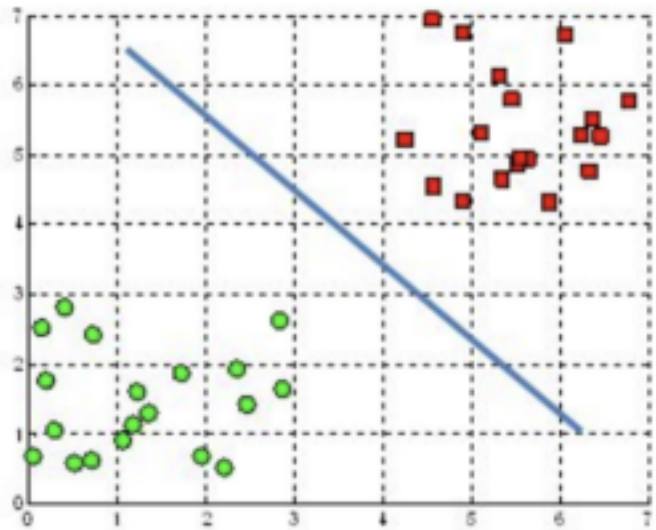
- Plot a graph of the training data using two different colours
- Separate the different classes with a line
- Then we plot the actual data



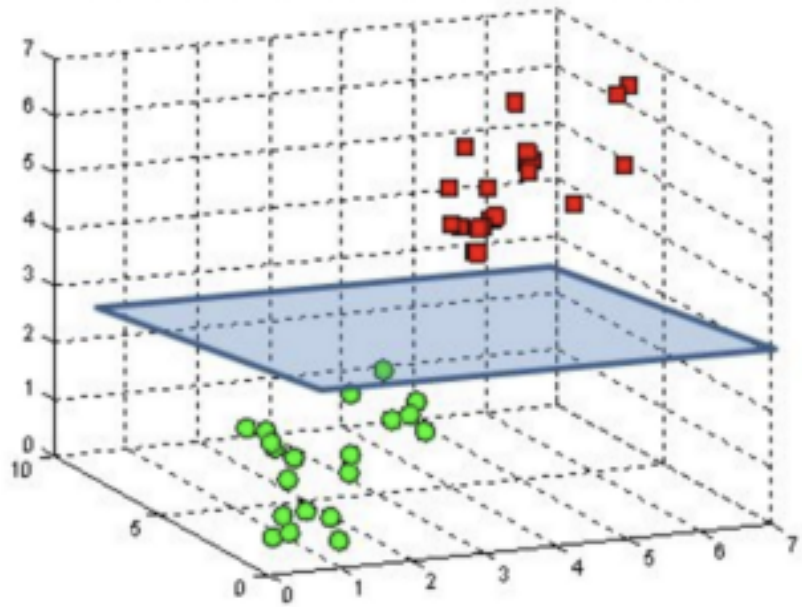
Support Vector Machine

The dimension of the hyperplane depends on the number of features. If the number of input features is 2 then the hyperplane is a line. If the number of input features is 3 then the hyperplane becomes a two-dimensional plane

A hyperplane in $\mathbb{R}^2$ is a line



A hyperplane in $\mathbb{R}^3$ is a plane



Incremental Learning and Batch Learning

- So far we have only looked at Batch learning, but what is batch learning and what is the alternative?
- Our alternative is Incremental learning

Hoeffding Trees

It revolves around the idea of the Hoeffding bound which goes as follows:

Hoeffding Bound

Given a random variable r with range R, if it has n observations which has a mean of $\bar{r}$. Then the true mean is at least $\bar{r} - \epsilon$ with probability $1 - \delta$ where $\epsilon = \sqrt{\frac{R^2 ln(1/\delta)}{2n}}$

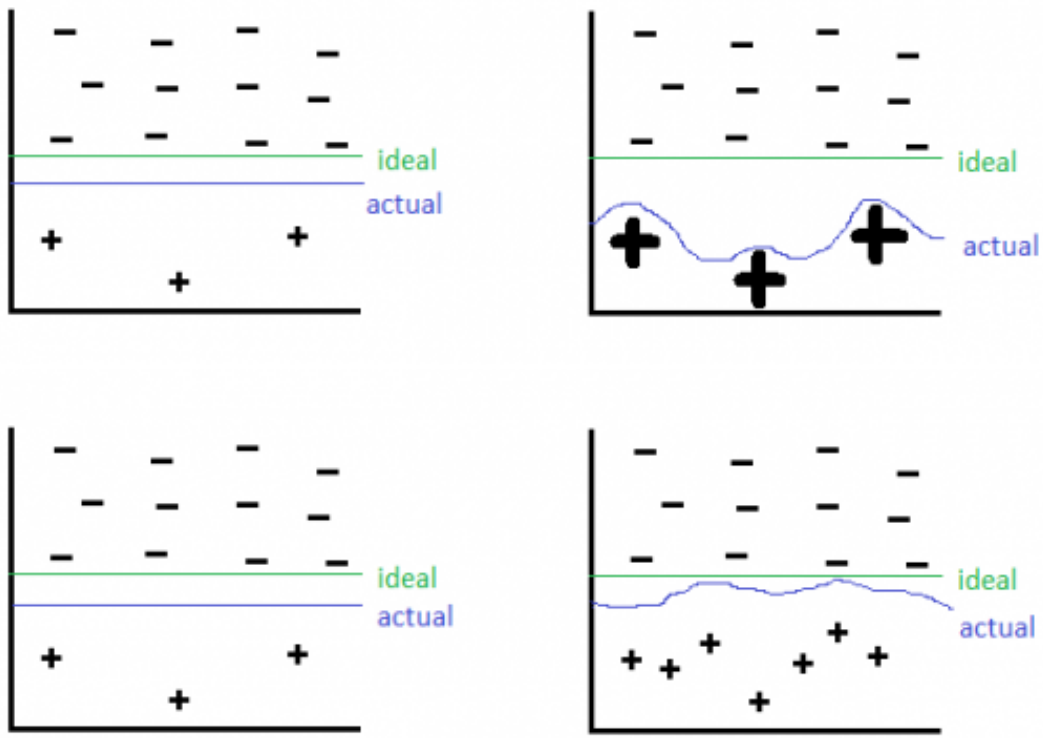
So let $G(\cdot)$ be a function that measures information gain for an attribute X. So $G(X_a)$ is the attribute with the highest information gain and $G(X_b)$ is the attribute with the second highest information gain. So $\Delta\bar{G} = \bar{G}(X_a) - \bar{G}(X_b)$. Then we split the tree on the n observaions as soon as $\epsilon < \Delta\bar{G}$ with the attribute X_a Made the correct choice with probability $1 - \delta$

Class Imbalance

- Ideally we have an equal amount of data for all the classes.
- Class imbalance plays an important role we see this if we look at the accuracy for a machine learning algorithm.
- So we have to introduce a different metric for measuring how good a machine learning algorithm is. So we introduce False Positive, False negative, True Positive and True negative rates
- So now we can compare different machine learning algorithms effectiveness when working with class imbalance.

Class Imbalance

- But how exactly do we reduce the effect of a data set with class imbalance?
- Introduce a cost function
- Under sampling and Over sampling
- SMOTE



End

Thank you for Listening