1. Describe briefly the steps in a typical data science project.

Understand -> collect -> explore viz -> clean n transform -> model -> validate -> communicate viz and deploy!

1. Why is exploratory data analysis a crucial step in any data analysis?

Because without a primal exploration of the data you can’t understand it, and find its limits, and characteristics, and you can’t see its potential, neither its *form*

1. Why is data visualization useful for data analysis?

As humans, we tend to understands things better if they’re visual.

It is to SEE the data, and understand its relations in a graphical way. This can also make you SEE the limits, characteristics, and possible ideas on what to clean and what to transform.

1. Why are tools developed for reproducible research useful in business data analysis workflows?

So that we can reutilize the same models in different environments; apply it for different production sets.

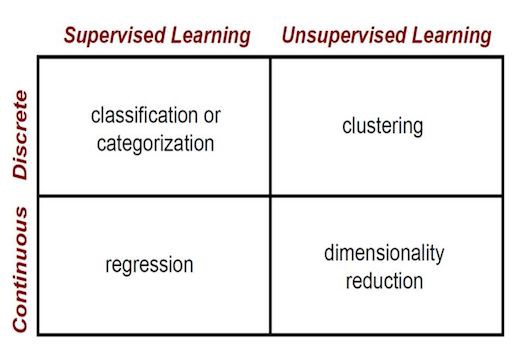
1. Describe the supervised learning problem in general.

It is done in the context of **classification**, when we want to map input to output labels (x to y), or **regression** when we want to map input to a continuous output.

it includes: logistic regression, naïve bayes, SVM, neural networks and random forests.

the **goal** is to find *specific relationships* or structure in the input data that allow us to effectively produce correct output data. Note that “correct” output is determined entirely from the training data, so while we do have a ground truth that our model will assume is true, it is not to say that data labels are always correct in real-world situations. Noisy, or incorrect, data labels will clearly reduce the effectiveness of your model.

main considerations are model complexity, and the bias-variance tradeoff



1. Why do we need a test set?

We need a test set to determine if the model works well with outside-train-set data. It is a way of trying if the model generalizes well and is not too specific on our train data

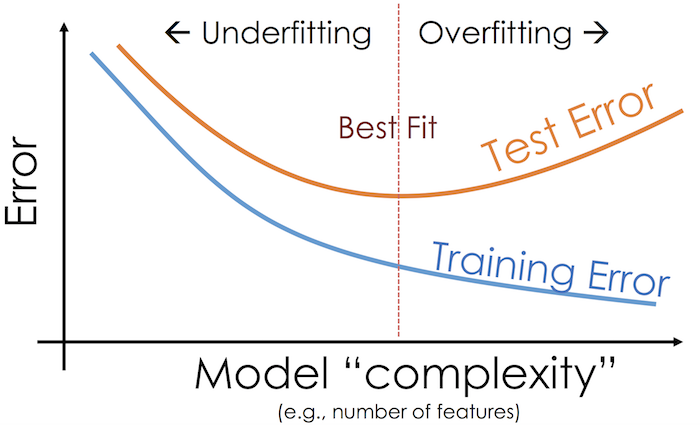
1. Why are evaluation metrics calculated from the training set misleading?

Because they don’t have to be true in the live data. The model must be tried outside the set to determine other kinds of metrics

1. How do we do k-fold cross validation for model evaluation?

We separate the whole data in k folds and we evaluate for each of those particular folds and compare performance. Cross validation uses all k possibilities and summarizes them all, keeping track of performance.

1. How does the train and test error change vs model complexity?

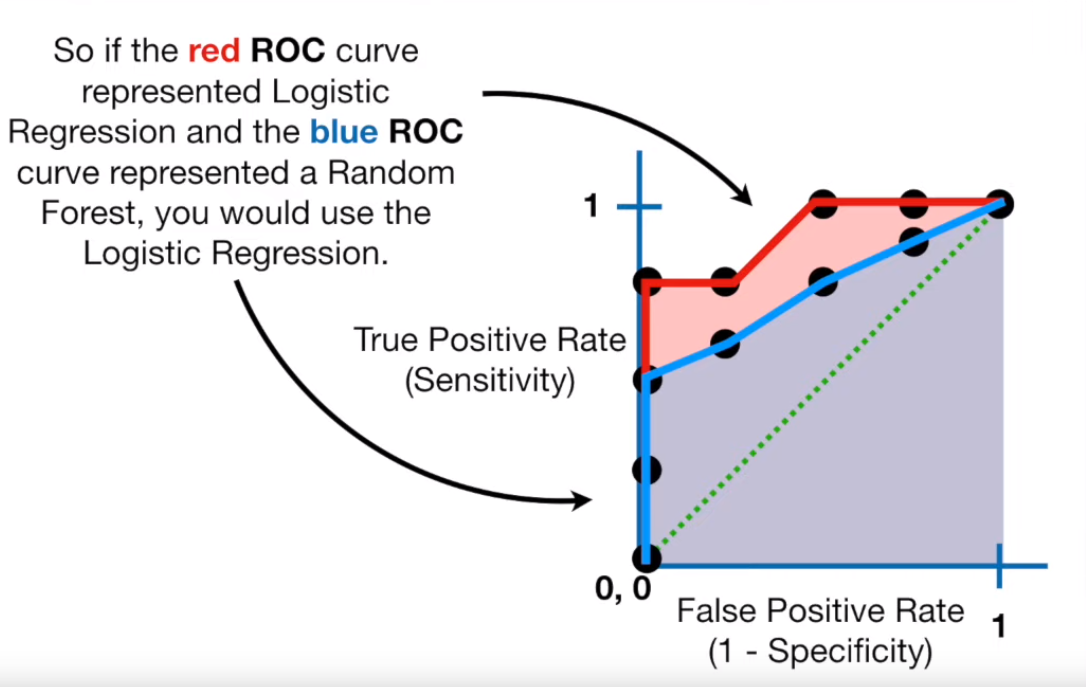


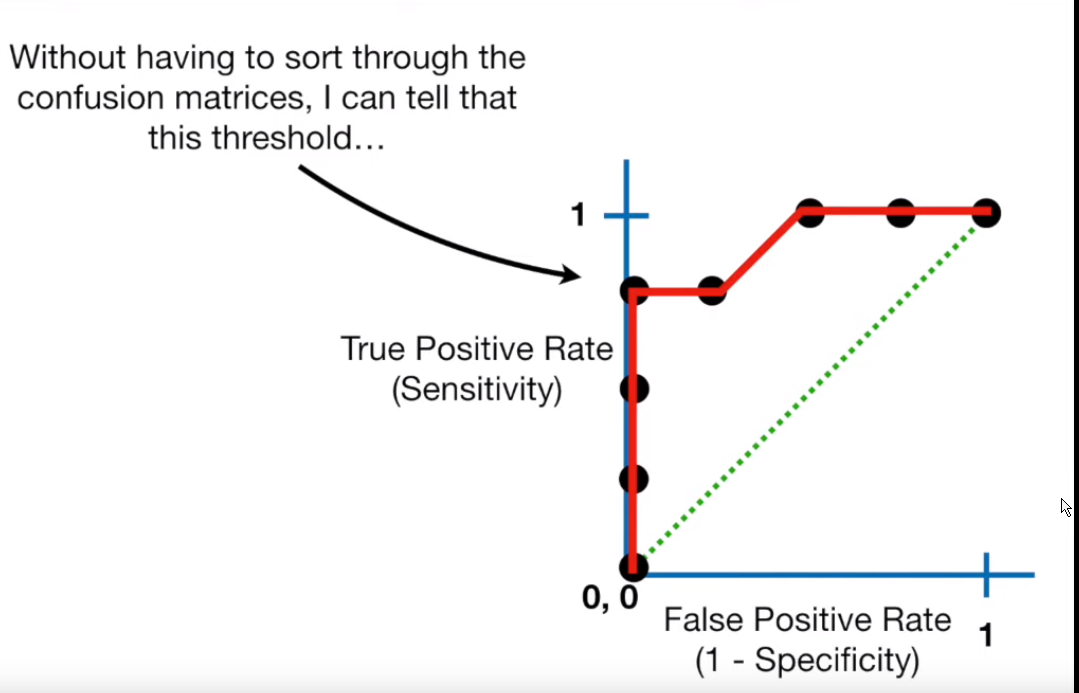


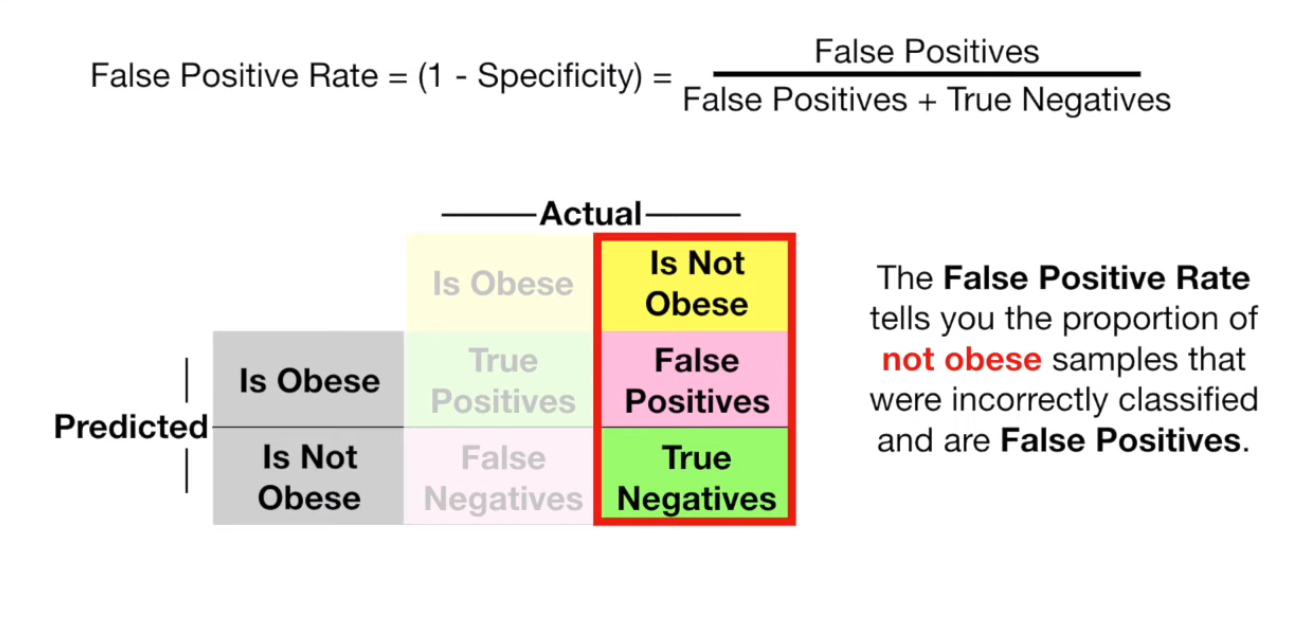
1. Describe overfitting.

"the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably"

Contains more parameters that can be justified by the data. It is too precise to the training data, that it cannot find a generalization and use it for outer data. It has a poor predictive performance.

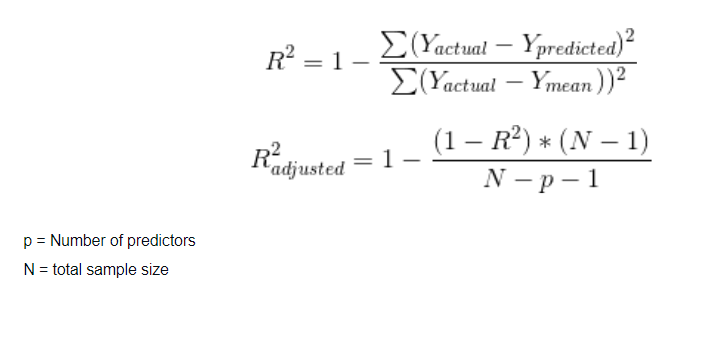
1. What is an ROC curve? 





THE AUC CURVE helps you see which method is better depending on how bigger is the area that’s the better method

The bias error is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).  
The variance is an error from sensitivity to small fluctuations in the training set. High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs (overfitting).



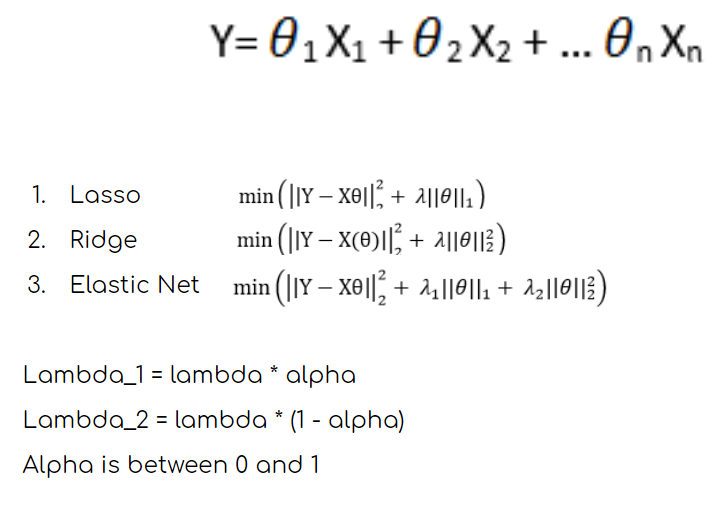
Lambda 0 to INF with cross validation

Ridge: taking square of the slope, better when most variables are useful

Lasso: take absolute value of the slope. Lasso can shrink all the way to 0   
it is better for excluding useless variables, reduces variance (overfitting)

elasticnet, hybrid result. is especially good at dealing with situations where there are correlations between parameters.

changing parameters for linear model complexity change?



Sum of square residuals +

decision trees, hyp params raise complexity:

Maxdepth: maximum depth of the tree. The default maxdepth = 0 means that no restrictions are applied to tree sizes.

Minsplit: the minimum sum of weights in a node in order to be considered for splitting.

Minbucket: the minimum sum of weights in a terminal node.

Mincriterion: the value of the test statistic (for testtype == "Teststatistic"), or 1 - p-value (for other values of testtype) that must be exceeded in order to implement a split.

1 Take a set of training data

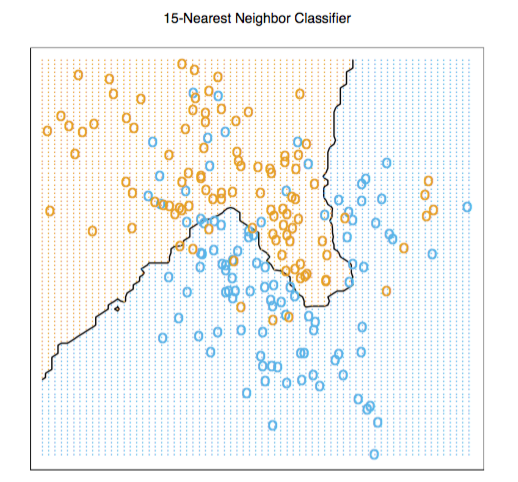
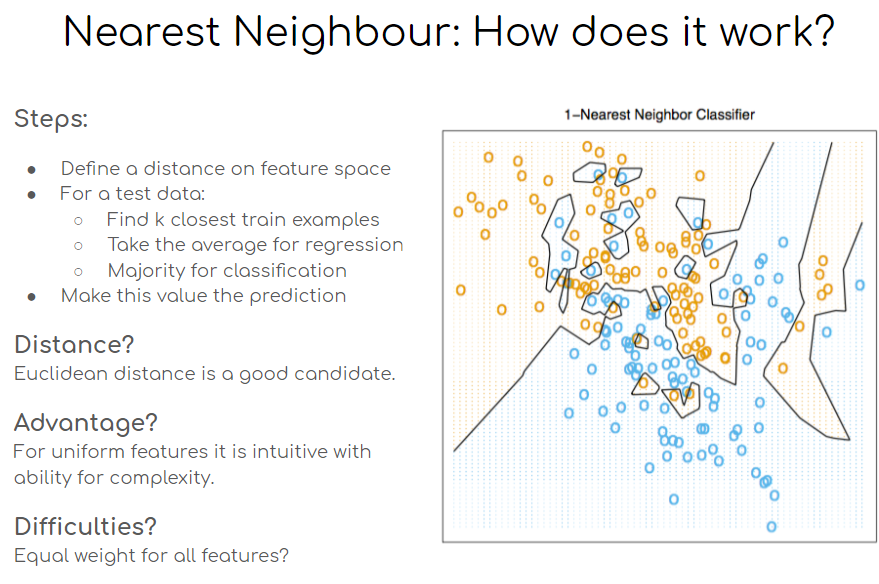
2 Look at the available features

3 Find the feature and a split defined on this feature that maximizes information gain (reduces the most entropy)

4 Repeat this step on the child nodes

To stop growing, set a minimum number of training inputs to use on each leaf.

k neearest neighbor,



unsupervised – identify patterns

clustering,

When do we stop? When no data point changes cluster

Do clusters depend on initial cluster points? Yes

What can we do about this exposure? Repeat the algorithm multiple times

Which cluster to choose in the end? The one with smallest Within-cluster point jitter

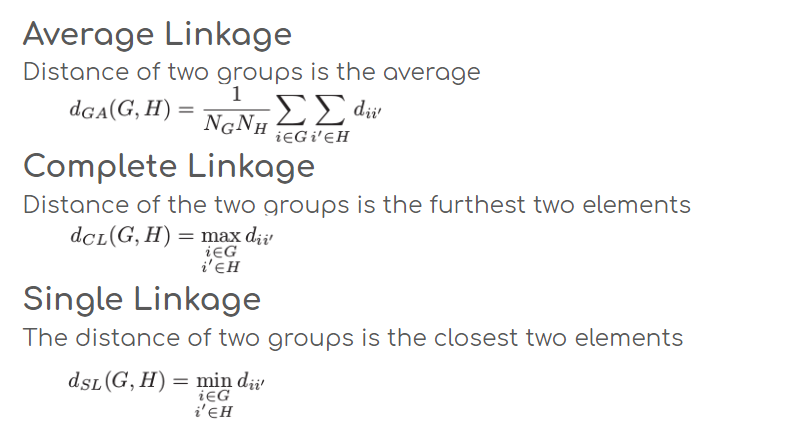
k- means: it includes Within, Total, Between cluster point scatter

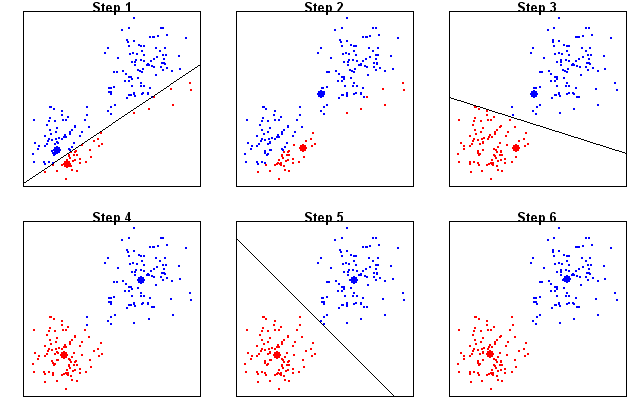
Take randomly or using a heuristic K data points. Let these be the initial cluster centers

Assign each datapoint from the dataset to the closest cluster

Update cluster center by calculating the new center of each cluster determined by its new members

Repeat from point 2, or stop if no changes were done

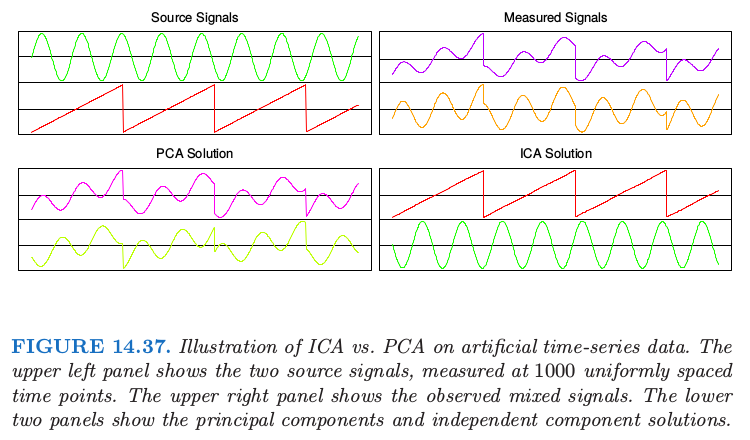
Hierarchical clustering:



PCA, principal component analysis. Helps us reduce dimensions in order to visualize and discard what’s not too informative. We remain with the most important aspects of our data. Usually you can work with 2 PCAs , you have to see how much importance each PCA has, and visualize how much information you are losing by dropping the least importants.

provides a reduced-rank representation of data

helps to compress data

ICA, INDEPENDENT  


provides a representation of the data as independent sub-elements

helps to separate data Hyperparameter search: go for random search

