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.

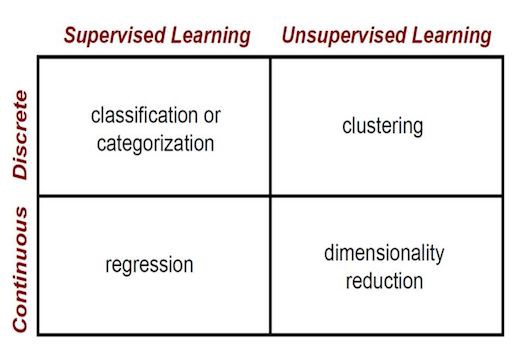
1. Describe the supervised learning problem in general.

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

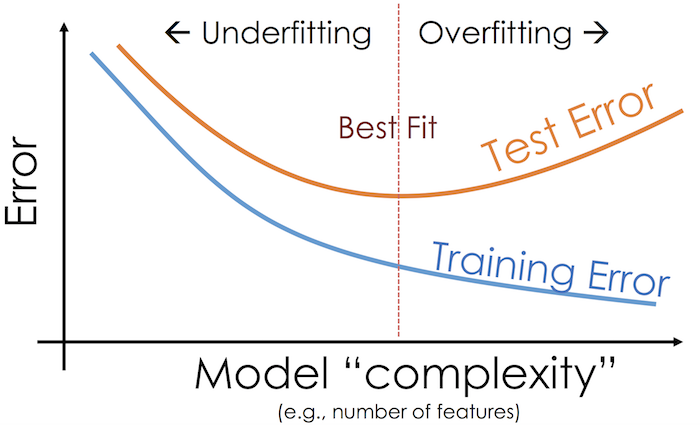
1. Why are evaluation metrics calculated from the training set misleading?

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.

**11.** What is an ROC curve?

supervised

unsupervised

train set, test set, bias variance,

ridge lasso elasticnet, changing parameters for linear model complexity change?

decision trees, k neearest neighbor,

what is the algorithm behind k nearest neighbors

whats the complexity for x

unsupervised

clustering, PCA, why do weuse them, how are they useful

go for the formulas, or main idea behind it.