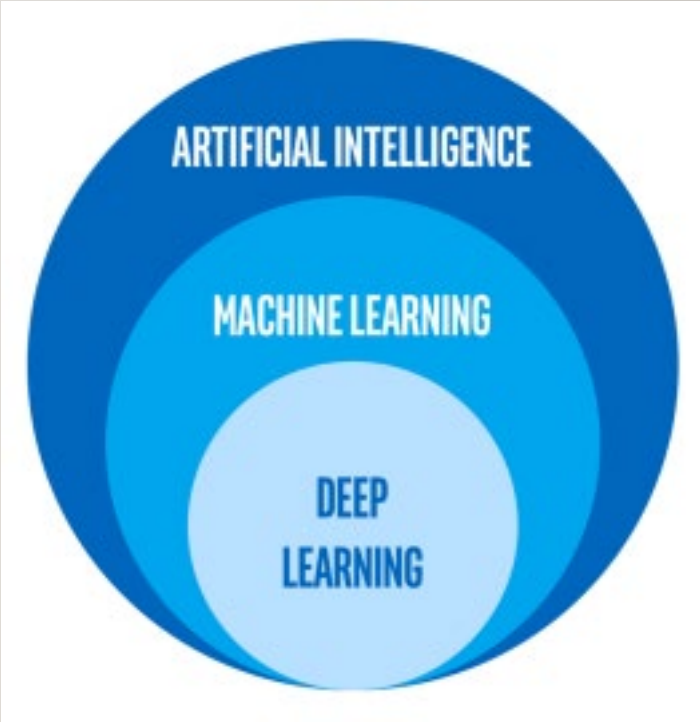
MACHINE LEARNING BASICS

WHAT IS MACHINE LEARNING?

- Machine learning is one part of artificial intelligence.
- ML is a way to build models* that can help make predictions about the future.
- We build these models by letting a computer learn from old data, and learning how to predict new data.

- * A model, in this context, is something that translates inputs into some output. The spreadsheet calculations we did before this are models – they translate the things we know, into what we want to know.



MACHINE LEARNING PROCESS

- The concepts behind ML are not all that complex, there are two key portions:
- Training a model:
  - This is where the computer learns to make predictions from old data.
  - We provide training data which contains features (the variables that we can track) and a target, the thing you want to predict. We know the target for all this data because it is old.
  - The algorithm processes the data, does some math, and learns some way to make predictions of what the target will be.
  - The end result of this is a model – we can enter inputs, and get a prediction of a target.
- Using a model:
  - The training part is all behind the scenes, using old data. That part is strictly creating a model.
  - To use a model, we plug in the features and get a prediction.
- An example…

MACHINE LEARNING EXAMPLE

- Suppose we are real estate investors who want to make predictions about house prices.
  - If we can predict which houses are actually worth more we can buy them and make money!
- Step 1 – get old data on house prices.
  - We want the target, the house prices, and the features, the things that we can know ahead of time.
  - We can always get data like sqft or year built, we want to accurately predict price.

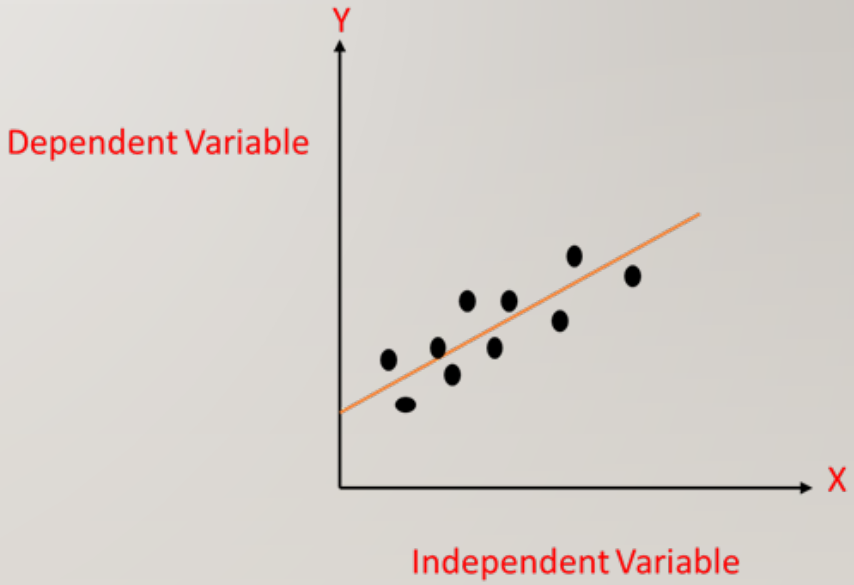
Target

Features

price	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built
221900.0	3	1.00	1180	5650	1.0	0	0	3	7	1180	0	1955
538000.0	3	2.25	2570	7242	2.0	0	0	3	7	2170	400	1951
180000.0	2	1.00	770	10000	1.0	0	0	3	6	770	0	1933
604000.0	4	3.00	1960	5000	1.0	0	0	5	7	1050	910	1965
510000.0	3	2.00	1680	8080	1.0	0	0	3	8	1680	0	1987

MACHINE LEARNING EXAMPLE

- Step 2 – feed data to the algorithm so it can learn to make predictions.
  - There are many algorithms, they all work slightly differently, but all do the same thing.
  - Ideally we have a lot of data here so the algorithm has a lot to learn from.
  - The algorithm will find a way to make predictions for the target, based on the features.
  - This will produce a model that can translate features to a prediction for the output.
- This process can be though of like a line of best fit, or a linear regression.
  - The reality is it is a larger version, with more than one input.
  - The line of best fit is a model, if we enter an input (x), we get a prediction (y).



MACHINE LEARNING EXAMPLE

- Once a model is created and trained, we can use it to make predictions.
- In real usage, we can know the features, we can't know the target, so we predict it.
  - For houses, we can know all the features (size, bedrooms, etc…) and we can use that info to make a prediction for the correct price.
- We can gather data that we want to predict (one per row).
  - This data has all of the features, but not the target – that's what we are predicting!

bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade	sqft_above	sqft_basement	yr_built
3	2.50	1530	1131	3.0	0	0	3	8	1530	0	2009
4	2.50	2310	5813	2.0	0	0	3	8	2310	0	2014
2	0.75	1020	1350	2.0	0	0	3	7	1020	0	2009
3	2.50	1600	2388	2.0	0	0	3	8	1600	0	2004
2	0.75	1020	1076	2.0	0	0	3	7	1020	0	2008

price
360000.0
400000.0
402101.0
400000.0
325000.0

MACHINE LEARNING IS EASY!

- ML is really just that simple!
- We can make things more large and more complex, but the core is just that process.

TYPES OF PREDICTIONS

- When we are doing ML there are two main types of predictions that we will look at:
  - Regression: regression is when we are predicting a raw value, like the house example.
  - Classification: classification is when we are predicting a discrete value, like true/false.
    - E.g. will someone be a customer or not, is this image a cat or not.
- Which one we are doing depends only on the output.
- The steps are almost entirely the same between the two, except…

HOW MUCH CAN THE PREDICTIONS BE TRUSTED?

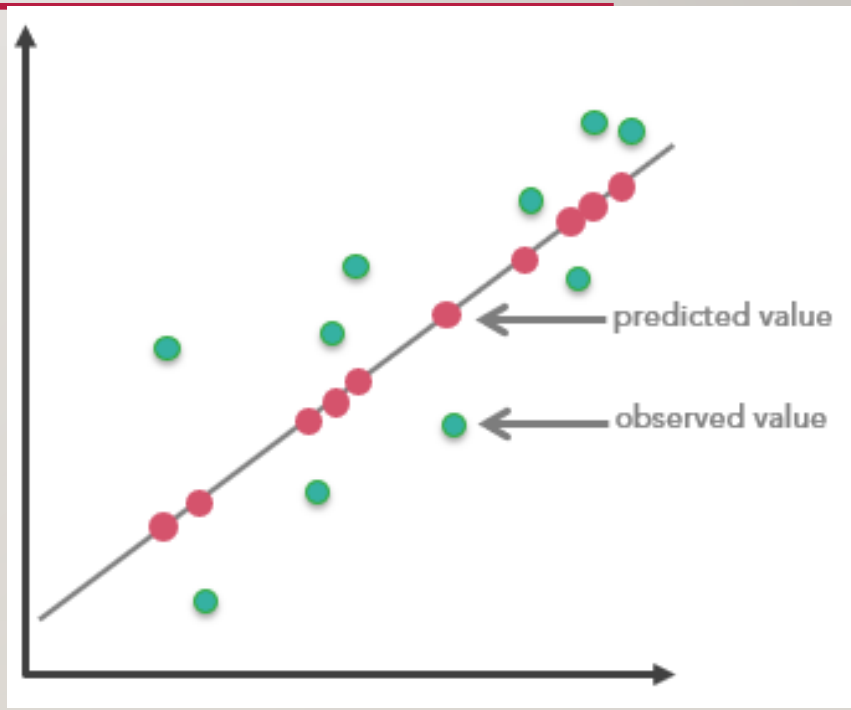
- … the measurements of error.
- When we build models we need some way to evaluate their accuracy.
  - The more accurate the model, the more we can trust it.
  - Think about the time series stuff, we can compare models and select the best.
- The error metrics we use are different between regression and classification:
  - Regression: RMSE and R2
  - Classification: Accuracy.

DETERMINING ERRORS

- As part of the training process the algorithm also produces an estimate of error.
- During the training the system does something called cross validation:
  - Take all the training data and split it into two sets, roughly 70/30.
  - Use the 70% to train the model, hold the 30% aside.
  - Take the features from the saved 30% (remember, we know the target) and make dummy predictions with the model that was just trained.
  - Compare those predictions to the real values to calculate error.
  - Repeat this process several times and average the error rates.

REGRESSION ERRORS

- Regressions have two error metrics to monitor.
- RMSE is root mean squared error:
  - The same error we explored doing time series in Excel.
  - The amount of error we expect in an average prediction.
  - Lower RMSE is less error in prediction, which means predictions are close to real values, which means better predictions.
- R2 is R squared:
  - R2 is a measure of completeness.
  - It estimates how much of the variation in the data is captured by the model. If we capture most, that's good. If we aren't capturing it, bad.
  - It is between 0 and 1, close to 1 is better.



CLASSIFICATION ERRORS

- Classification error is more simple, at least for now, it is the accuracy.
- The accuracy is just what percentage of the predictions we got right.