# CS-344 Guide 7 - Regression

- First steps with tensorflow:
  - Do you believe that tensorflow can be used to encode anything you can imagine?
    - Everything except for spiritual matters. (You can't encode God)
  - Compare and contrast tf.estimator vs. SciKit-Learn
    - Tf.estimater = high-level API specifying pre-defined architectures including linear regression and neural networks.
      - Actions: training, evaluation, prediction, and export for serving.
    - SciKit-Learn:
      - Classification: identifying to which category an object belongs to
      - Regression: predicting a continuous-valued attribute associated with an object.
      - Clustering: automatic grouping of similar objects into sets.
      - Dimensionality reduction: reducing the number of random variables to consider.
      - Model selection: comparing, validating and choosing parameters and models.
      - Preprocessing: feature extraction and normalization.
  - What is a tensor?
    - The primary data structure in TensorFlow programs.
    - N-dimensional data structures – scalars, vectors, matrices, etc.
    - Elements can hold integer, floating-point, or string values.
  - Note: we include the pandas tutorial below; save the tensorflow and synthetic features tutorials for the lab
- Generalization:
  - Occam's razor:
    - The less complex an ML model, the more likely that a good empirical result is not just due to the peculiarities of the sample.
  - IID:
    - Independently and identically.
    - Examples don't influence each other.
    - Refers to the randomness of variables.
  - Stationarity:
    - The distribution doesn't change within the data set.
- Training and test sets:
  - Should we randomize our examples before splitting the train/set sets? If so, why; if not, why not?
    - Yes, we should because the examples could be given in sorted order or otherwise organized in a way that isn't random.
    - This will affect our predictions.
- Validation set:
  - Compare and contrast train vs validation vs test datasets
    - Train set: a subset to train a model.
      - Used for learning to fit the parameters (weights) of classifiers, etc.
    - Validation set: a subset that is used to adjust hyper-parameters.
    - Test set: a subset to test the trained model.

    - Independent of training dataset but follows same probability distribution.
    - Minimal overfitting if model fits well to test dataset but overfitting if model fits better to training dataset.
- Pandas: Do Google’s Intro to Pandas
-