Machine Learning Practice

Andrew H. Fagg

Know Your Problem and your Data

Spend time understanding:

- The problem that you are trying to solve
- The questions that you need to answer
- The meaning behind the different parts of your data
- The costs for collecting / labeling data
- The costs for making different prediction errors

Know Your Data

The meaning behind your feature vectors, especially individual features

- What are their distributions?
- How do they correlate with one-another and with the thing you are trying to predict?
- Continuous or enumerated (or both)?
- How much data do you have / can you get?

Always spend time visualizing your data

Know Your Problem

- What type of a prediction problem are you facing?
- Supervised, Semi-supervised, Unsupervised
- Continuous, probabilistic or categorical prediction?

Know Your Approaches

- What is the right learning approach for the job?
  - Depends on the details of your data and on the details of your predictions
- Don’t be afraid to try simple approaches
  - Quick to implement
  - Depending on the problem, this may be the solution that you need
  - Either way, you will learn useful things about your problem

The Full Machine Learning Process

- Try a few quick solutions & do a bit of hand tuning
  - Identify the right representations, right approaches, and (approximately) the right hyper-parameters
- Grid search + cross-validation
  - Systematic testing of hyper-parameter options
  - May need multiple grid search runs
  - Note that there are other non-grid search approaches
- Statistical comparisons
  - Across hyper-parameter choices (validation data sets)
  - Across modeling approaches (test data sets)

Over-fitting

It is easy to over-fit data sets!

- Primary issue: training set size is too small given the number of parameters that we trying to fit
- Always look at the over-fitting question by varying training set size
- Some methods have mechanisms that combat over-fitting directly
- Others need the data to be structured properly

Euclidean Distance

- Euclidean distance is at the center of many ML algorithms
- However, this metric is not always meaningful
- Especially an issue when we are working in high-dimensional feature spaces
- In these cases, manifold-sensitive approaches are appropriate
  - PCA, LLE, MDS, ISOmap, tSNE
- Or, approaches that don't focus on the full feature space:
  - Trees, Forests

Make Your Argument

Your customer/supervisor may or may not care about the details of the methods used

- Be ready to talk about the high level of your analysis
  - Specific methods / hyper-parameters are probably not important
- Show data, including intermediate results
- Be honest about what works and what doesn't
  - Specific examples + aggregate results
  - Make clear statistical arguments

Machine Learning Practice