Fourth Industrial Summer School

Advanced Machine Learning

Tips for ML + KNIME

Session Objectives

- ✓ Tips on machine learning
- ✓ KNIME

Tips on Machine Learning

Important remarks

- Mainly for supervised learning but some aspects are general
- Not for academic research but for development

Some guidelines–1

- Understand your goal
  - What are you trying to achieve, translated into:
    - Classification
    - Regression
    - Unsupervised learning
    - …
- Understand your data
  - What data you have
  - How much data you have
  - Characteristics of your data

Some guidelines–2

- Prepare your data
  - Clean your data
  - Select your features
  - Continuous vs. categorical features
  - How to encode the target
  - Data scaling
  - Data imbalance
  - Train/validation(cross-validation)/test sets
- Data collection and annotation
  - Realistic conditions
  - Dev. and test sets from the same distributions

Some guidelines–3

- Select measures appropriate for the task (get to an agreement)
  - Also depends on the data
  - Accuracy vs. precision & recall (unify it)
  - Prioritize your measures (accuracy vs. runtime), optimizing vs. satisficing
  - Change them later if needed
- What should be the size of train/dev/test partitions?
  - 70/15/15?, 60/20/20?
  - What if you huge amount of data?
- Significance interval of differences in performance
  - 95% percent confidence intervals
  - The smaller the intended progress, the larger the dev set needed

Some guidelines–4

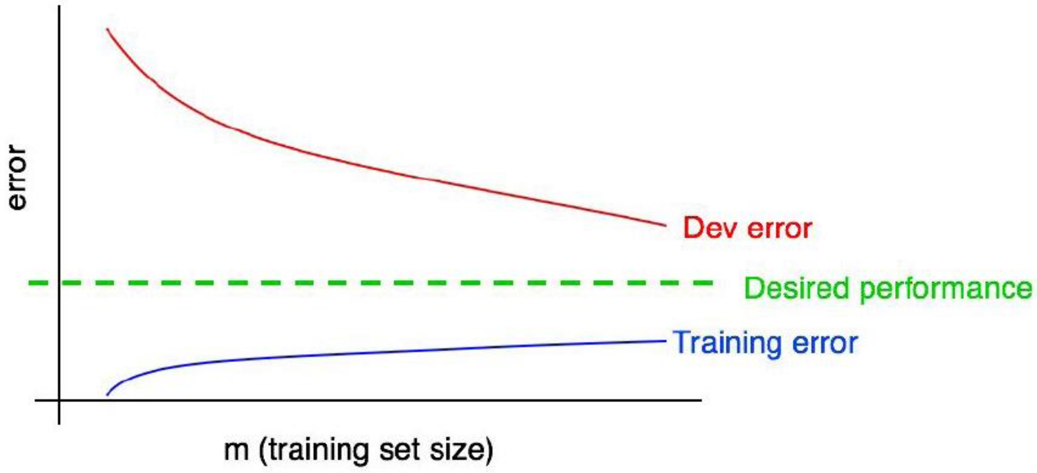
- Decide which algorithm to use
  - Start simple (linear/logistic regression) unless there is a clear case
  - Incrementally build complex ideas
  - Start small instead of making big goals from the very beginning
- Training
  - Trying to fit (overfit) and then,
  - Deal with variance
- Error analysis
  - Do error analysis to decide how to go forward by manually looking at the errors on the dev set
  - What if the dev set is large (create a small subset for manual analysis)

Some guidelines–5

- What about mislabeled data?
  - Correct them if they are a major cause of errors
  - Make sure to update the test set as well
  - What about errors in labels classified as correct category?
- Bias/variance and adding more data
  - What would be an optimal error rate?
  - All variance is avoidable (more data) but all bias is not
- High (avoidable) bias
  - Make your model more complex
  - More training
  - Error analysis on the training set

Some guidelines–6

- High variance
  - Add more training data
  - Regularization
  - Early stopping
  - Simpler model
  - Study the error curves



Andrew Ng, Machine Learning Yearning, deeplearning.ai

Some guidelines–7

- Data augmentation
- End-to-end recognition vs. a standard ML pipelines
- Use ensembles to give a final push to your results

KNIME

Data analysis and machine learning

KNIME (Konstanz Information Miner)

- It is a free and open-source data analytics, reporting and integration platform.
- Modular data pipelining concept.
- No programming needed.

- DEMO

References

- Andrew Ng, Machine Learning Yearning, deeplearning.ai
- Mehryar Mohri, Afshin Rostamizadeh, Ameet Talwalkar, Foundations of Machine Learning, second edition, The MIT Press
- Tom M. Mitchell, Machine Learning, McGraw-Hill, 1997