CS 584 – Introduction to machine learning

General characteristics:

- Supervised vs unsupervised learning
- Generative vs discriminative learning
- Regression vs classification
- Parametric vs non-parametric
- Online vs offline
- Graphical models

Training vs Testing

- Train/test on different collections
- Training error vs Testing error (want small Testing error)
- Generalization vs Overfitting (good testing vs good training)
- Normally training error $<$ testing error

Model selection

- How to select model
- Model capacity
  - Too large: testing error
  - Too small: testing and training error

Cross validation

- Split data into training and testing sets.
- To maximize utilization, repeat the process with different splits.
- K-fold cross-validation:
  - Split data to K parts
  - Leave one part out, and train on remaining K-1 parts
  - Repeat K times
  - Compute average error
  - For final model train on all examples (report cross-validation result)
- Leave one out K=m (# examples). Speed up by random sampling.

Typical ML process

- Data collection (collect, label, clean, extract features)
- Apply algorithm (train, test, error analysis, make changes)
- Test on validation set

Feature types

- Qualitative (categorical)
  - discrete ordinal
  - discrete nominal
- Quantitative (numerical)
  - continuous (can be discretized)
  - discrete numerical

Pitfalls

- Correlation vs causation
- Confounding factor
- Selection bias
- Reporting bias
- Survival bias
- Backward looking studies