10601
Machine Learning

Model and feature selection

Occam's Razor

- William of Ockham (1285-1349) *Principle of Parsimony*:
  - "One should not increase, beyond what is necessary, the number of entities required to explain anything."
- Regularization penalizes for *"complex explanations"*

- Alternatively (but pretty much the same), use *Minimum Description Length (MDL) Principle*:
  - minimize *length*(misclassifications) + *length*(hypothesis)

- *length*(misclassifications) – e.g., #wrong training examples
- *length*(hypothesis) – e.g., size of decision tree

Minimum Description Length Principle

- MDL prefers small hypothesis that fit data well:

$$h_{MDL} = \arg\min_h L_{C_1}(\mathcal{D} \mid h) + L_{C_2}(h)$$

  - $L_{C1}(D|h)$ – description length of data under code C_1 given *h*
    - Only need to describe points that *h* doesn't explain (classify correctly)
  - $L_{C2}(h)$ – description length of hypothesis *h*
- Decision tree example
  - $L_{C1}(D|h)$ – #bits required to describe data given *h*
    - If all points correctly classified, $L_{C1}(D|h) = 0$
  - $L_{C2}(h)$ – #bits necessary to encode tree
  - Trade off quality of classification with tree size

What you need to know about Model Selection, Regularization and Cross Validation

- Cross validation
  - (Mostly) Unbiased estimate of true error
  - LOOCV is great, but hard to compute
  - *k*-fold much more practical
  - Use for selecting parameter values!
- Model selection
  - Search for a model with low cross validation error
- Regularization
  - Penalizes for complex models
  - Select parameter with cross validation
  - Really a Bayesian approach
- Minimum description length
  - Information theoretic interpretation of regularization

Bayesian approach

- Start with a simple model
- As data comes, increase the complexity as necessary

- My research area: Nonparametric Bayes
- The complexity of the model is unbounded
- Select the correct complexity from data (posterior)
- For ex: the number of clusters

Feature selection

- Choose an optimal subset from the set of all N features
  - Only use a subset of a possible words in a dictionary
  - Only use a subset of genes
- Why?
- Can we do model selection to solve this? – 2^n models

Two approaches: 1. Filter

- Independent of classifier used
- Rank features using some criteria based on their relevance to the classification task
- For example, mutual information:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p_1(x)\, p_2(y)} \right),$$

- Choose a subset based on the sorted scores for the criteria used

2. Wrapper

- Classifier specific
- Greedy (large search space)
- Initialize F = null set
  - At each step, using cross validation or an information theoretic criteria, choose a feature to add to the subset [training should be done with only features in F + new feature]
  - Add the chosen feature to the subset
- Repeat until no improvement to CV accuracy

Problem Set 4

Q1.3:

- Take derivatives w.r.t to α first then w,b.

Q. 1.6

- Minimize the violations as much as possible.
- Assume C is large but not ∞.

Q. 2

- Either explain why some algorithm does not work well.
- Or draw the final result of the algorithms.

Q. 3

The contour of the distribution in 2-D:

- Spherical Gaussian: concentric circles
- Diagonal Gaussian: concentric eclipses with axes parallel to the coordinate axes.
- Unrestricted covariance Gaussian: concentric eclipses

Q. 4

- Cutting the tree by thresholding