Lecture 5: Ensemble Learning

Course Teacher: Md. Shariful Islam Bhuyan

Assignment 1 Explanation

- Preprocessing
- Decision Tree
- Adaboost
- Evaluation

Ensemble Approaches

- Model diversification
  - Voting/Averaging of multiple model
  - Stacking: use model outputs as feature for next layer
  - Mixture of expert: gating network, different part of input space
- Dataset diversification
  - Boosting: weighted resampling with replacement
  - Bagging: random resampling with replacement
    - Bootstrap aggregating
    - Feature/attribute bagging or random subspace
    - Random forest

Aggregation

- Combine predictions with function

$$\hat{f}(y|\mathbf{x}) = \sum_{t=1}^{T} w_t f_t(y|\mathbf{x}) \text{ [Sum]}$$

$$\hat{f}(y|\mathbf{x}) = \prod_{t=1}^{T} f_t(y|\mathbf{x})^{w_t} \text{ [Product]}$$

$$\hat{f}(y|\mathbf{x}) = \text{sign}\left(\sum_{t=1}^{T} w_t f_t(y|\mathbf{x})\right) \text{ [Voting]}$$

- Less likely than a misclassification by a single hypothesis

$$P\left(X \geq \left\lceil \frac{n}{2} \right\rceil\right) = \sum_{k=\left\lceil \frac{n}{2} \right\rceil}^{n} \binom{n}{k} \epsilon^k (1-\epsilon)^{n-k}$$

Hypothesis space

- **Model** … defined by variables (measurable quantities) and their relationships (structure)
- Choose **representation** of the model (function, correlation, network, inequalities, equation …)
- **Hypothesis space**, set of all hypotheses, **restriction bias** … linear cannot represent quadratic
- **Realizable** hypothesis space, containing true function [Is there a true function?]
- Why not the set of all Turing-computable function?
  - Tradeoff between the **expressiveness** and the complexity of finding a good one
- **Training algorithm** and **objective function** to search target model from hypothesis space
  - optimization, approximation, greedy, dynamic, sampling etc.
- Incomplete/complete search/space (**preference bias**)

Bias-Variance Tradeoff

- Maximum overfit … each example, has separate rule
- Maximum underfit … don't look at example … declare class
- Complex hypotheses fit the training data well
- Simpler hypotheses may generalize test data well
- Stationarity assumption, independent and identically distributed
- https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff
- Ensemble reduce variance

Preprocessing

- http://www.cs.ccsu.edu/~markov/ccsu_courses/DataMining-3.html
- Real world data are generally
  - Incomplete
  - Noisy
  - Inconsistent
- Tasks in data preprocessing
  - Data cleaning: fill in missing values, smooth noisy data, remove outliers, resolve inconsistencies.
  - Data integration: using multiple databases, data cubes, or files.
  - Data reduction: reducing the volume but producing the same or similar analytical results.
  - Data discretization: part of data reduction, replacing numerical attributes with nominal ones.
  - Data transformation: normalization and aggregation, dimensionality reduction

Binarization

Cheat	No		No		No		Yes		Yes		Yes		No		No		No		No	
Taxable Income																				
Sorted Values →	60		70		75		85		90		95		100		120		125		220	
Split Positions →	55		65		72		80		87		92		97		110		122		172		230	
	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
Yes	0	3	0	3	0	3	0	3	1	2	2	1	3	0	3	0	3	0	3	0	3	0
No	0	7	1	6	2	5	3	4	3	4	3	4	3	4	4	3	5	2	6	1	7	0
Gini	0.420		0.400		0.375		0.343		0.417		0.400		*0.300*		0.343		0.375		0.400		0.420	