

430.457

Introduction to Intelligent Systems

Prof. Songhwai Oh
ECE, SNU

GENERALIZATION ERROR

Overfitting and Other Considerations

- **Overfitting:** When a complex model is used for a small number of data, it may fit the training set well but have a large generalization error on future data.
- Decision tree pruning: A method to prevent overfitting in a decision tree. For each node having leaf nodes as descendants, perform a significance test. If a null hypothesis is accepted, prune away the node.
- Other considerations:
 - Missing data. Solution: Weighting all possible paths.
 - Multivalued attributes: When an attribute has many possible values, the information gain measure gives an inappropriate indication of the attributes usefulness. Solution: $\text{Gain ratio} = (\text{information gain}) / (\text{intrinsic value})$.
 - Continuous and integer-valued input attributes: Continuous or integer-valued attributes have an infinite set of possible values. Solution: Finding a split point with the highest information gain.
 - Continuous-valued output attributes. Solution: Regression tree.

Loss Function

- **Loss function:** $L(x, y, \hat{y})$ is defined as the amount of utility lost by predicting $h(x) = \hat{y}$ when the correct answer is $f(x) = y$.

Absolute value loss: $L_1(y, \hat{y}) = |y - \hat{y}|$

Squared error loss: $L_2(y, \hat{y}) = (y - \hat{y})^2$

0/1 loss: $L_{0/1}(y, \hat{y}) = 0$ if $y = \hat{y}$, else 1

- Suppose each example (x, y) is independent and identically distributed (i.i.d.) from $P(X, Y)$.
- **Generalization loss** for a hypothesis h with respect to loss function L is

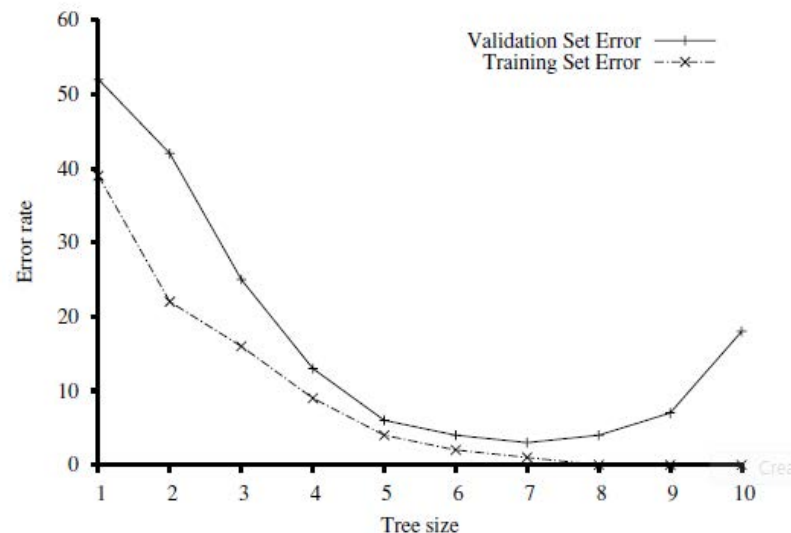
$$GenLoss_L(h) = \sum_{(x,y) \in \mathcal{E}} L(y, h(x))P(x, y),$$

where $\mathcal{E}$ is the set of all input-output pairs.

- Our goal is to find $h^* = \arg \min_{h \in \mathcal{H}} GenLoss_L(h)$.
- **Empirical loss:** $EmpLoss_{L,E}(h) = \frac{1}{N} \sum_{(x,y) \in E} L(y, h(x))$, where E is the training set, estimates generalization loss.
- The estimated best hypothesis may differ from the true function due to: unrealizability, variance, noise, and computational complexity.

Cross Validation, Model Selection

- **k -fold cross validation:** to measure how well a hypothesis generalizes for the future data.
 1. Split the data into k equal subsets.
 2. Perform k rounds of learning; on each round $1/k$ of the data is held out as a test set and the remaining examples are used as training data.
 3. Compute the average test set score of the k rounds.
- **Leave-one-out cross-validation** when $k = n$.
- **Model selection:** the problem of finding an optimal model class (hypothesis space) in terms of the expressiveness-complexity tradeoff.



Simplifying Models

- **Regularization:** penalizes complex hypotheses.

$$Cost(h) = EmpLoss(h) + \lambda Complexity(h)$$

$$\hat{h}^* = \operatorname{argmin}_{h \in \mathcal{H}} Cost(h) .$$

← Regularization
function

- **Feature selection:** discard attributes that appear to be irrelevant, e.g., pruning.
- Model selection methods:
 - Akaike information criterion (AIC)
 - Bayesian information criterion (BIC)
 - Minimum description length (MDL)
 - Bayes factor
 - And more.