Introduction to Machine Learning

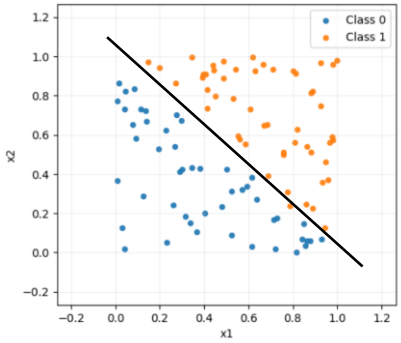
Lecture 15:

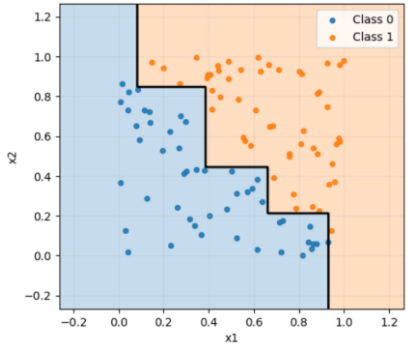
Ensemble Methods (Bagging, Random Forests)

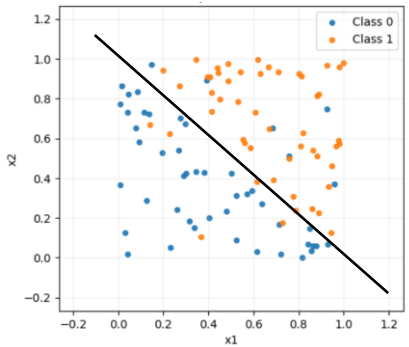
Prof. Subrahmanya Swamy Peruru
Department of Electrical Engineering
IIT Kanpur

Pruning to Prevent Overfitting

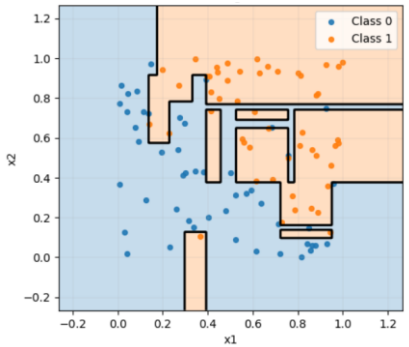
No noise in data



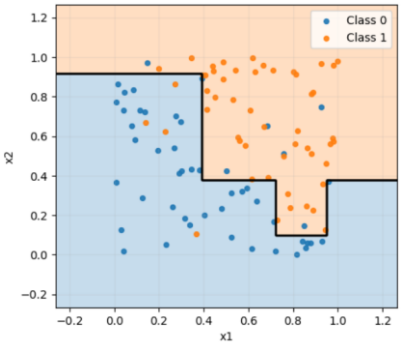




Noisy Data



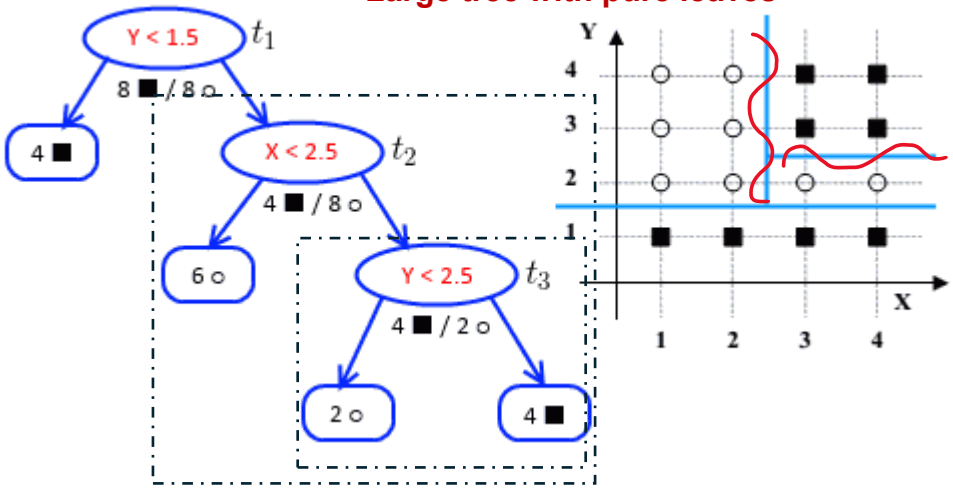
Overfitting



Pruned tree

Post Pruning

Large tree with pure leaves



Cost-Complexity Pruning

$$R_\alpha(T) = R(T) + \alpha|T|$$

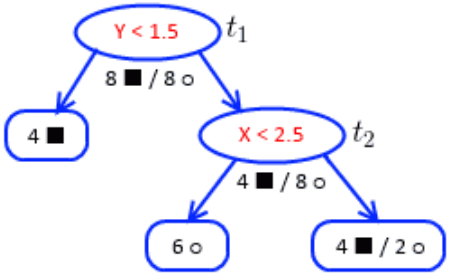
$|T|$: number of leaf nodes
$R(T)$: Total Misclassification error on training data

Interpretation:
First term: Encourages fewer mistakes
Second term: penalty for model complexity

Very Large $\alpha \Rightarrow$ simpler tree, high training error
Very Small $\alpha \Rightarrow$ complex tree, low training error

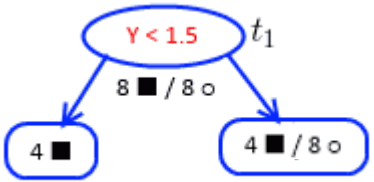
Use cross validation to tune α

Pruning at node t_3



misclass error: 2/6

Pruning at node t_2



misclass. error: 4/12

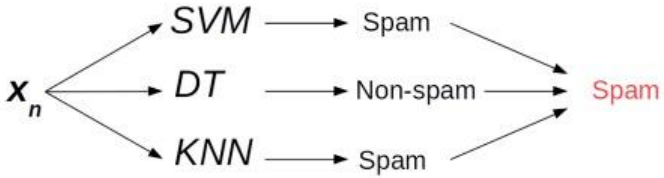
Fig Source: http://mlwiki.org/index.php/Cost-Complexity_Pruning

Ensemble Methods

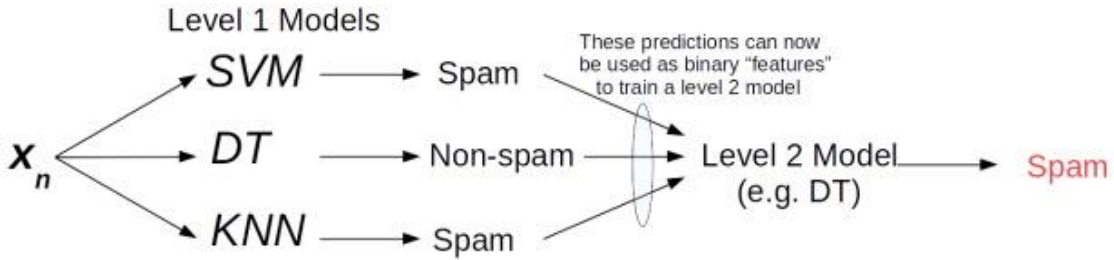
Slides adapted from Prof. Piyush Rai

Some Simple Ensembles

- Voting or Averaging of predictions of multiple models trained on the same data



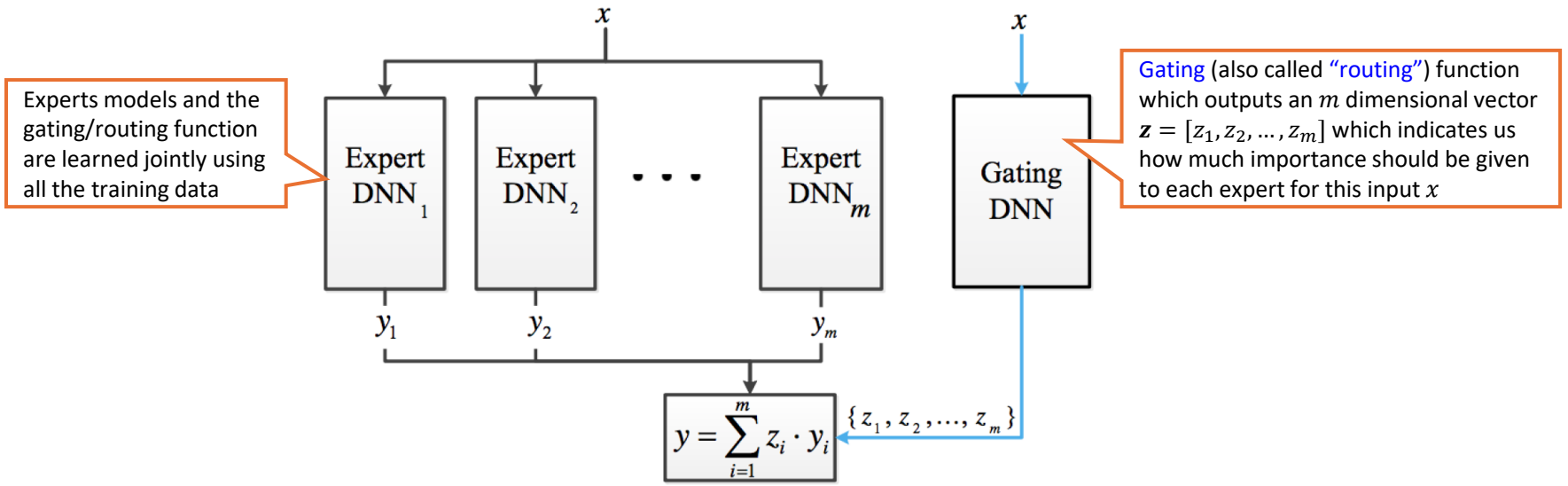
- "Stacking": Use predictions of multiple already trained models as "features" to train a new model and use the new model to make predictions on test data



In stacking, level 1 and level 2 models are trained independently (first level 1 and then level 2)

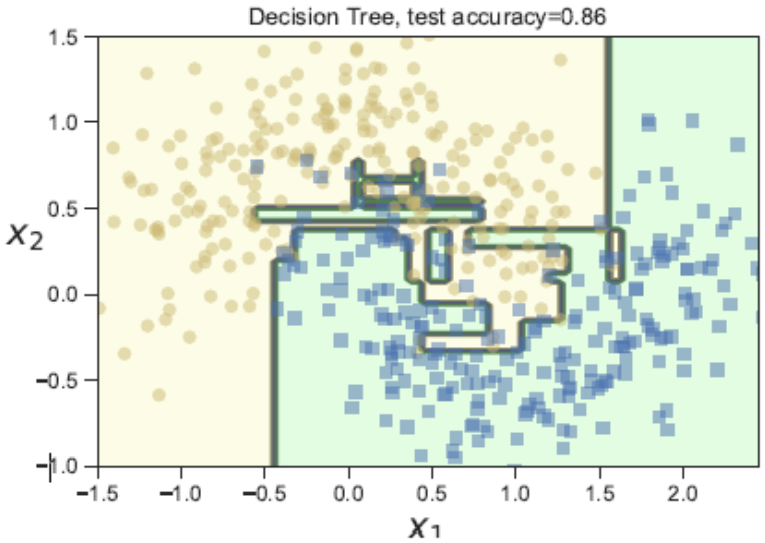
Mixture of Experts (MoE) based Ensemble

- We assume m "simple" models, usually of the same type, e.g., m linear SVMs or m logistic regression models, or m deep neural nets (usually all with same architecture)

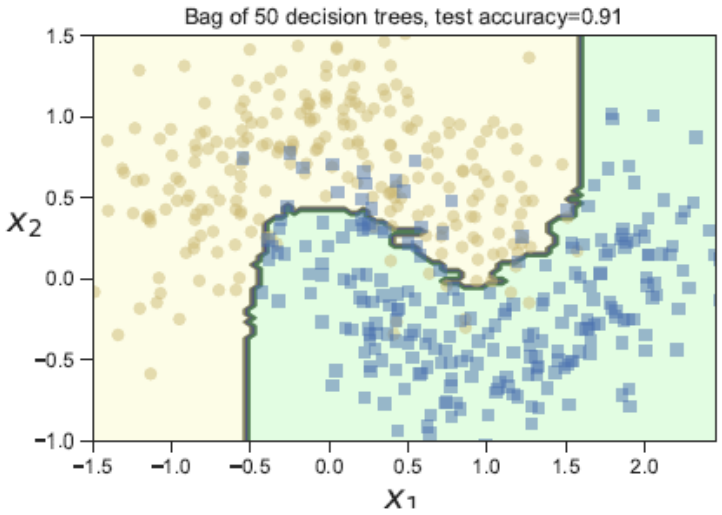


- MoE is very popular in classical ML as well as "modern" deep learning

Bagging (Bootstrap Aggregation)



A single decision tree



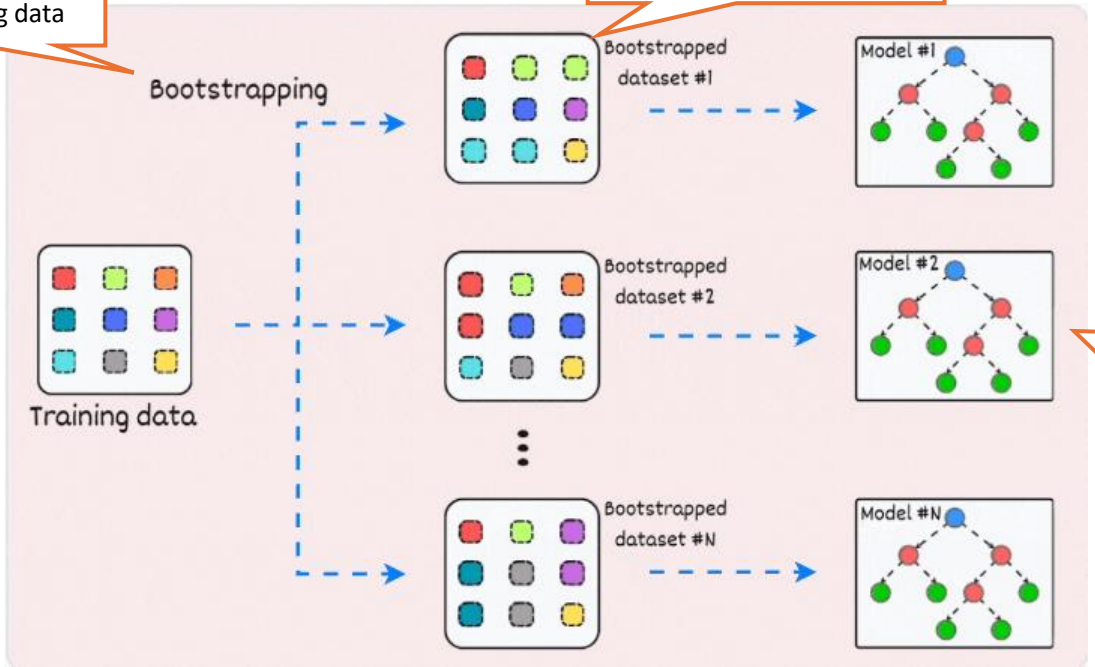
Bagging with 50 decision trees

Bagging is a technique to reduce variance by averaging using several decisions

Bagging (Bootstrap Aggregation)

sampling with replacement from original training data

Sampling with replacement will give about 63% unique training examples (if both datasets are of equal size)



To predict the output for a new test input, we apply all these models and predict using their **average or majority**

Parallel training for all the N models

Figure adapted from: https://www.blog.dailydoseofds.com/p/an-animated-guide-to-bagging-and

Random Forests

Limitations of Bagging:

Trees are correlated because

- Strong predictors chosen at top splits
- All trees look similar

Solution: Random Forest (Decorrelate trees)

At each split:

- Randomly select m features out of p
- Choose best split among those m

Typical:

$$m = \sqrt{p} \quad \text{(classification)}$$

Rest same as bagging.

Thank You