MACHINE LEARNING CDS503

Topic 11: Ensemble Learning

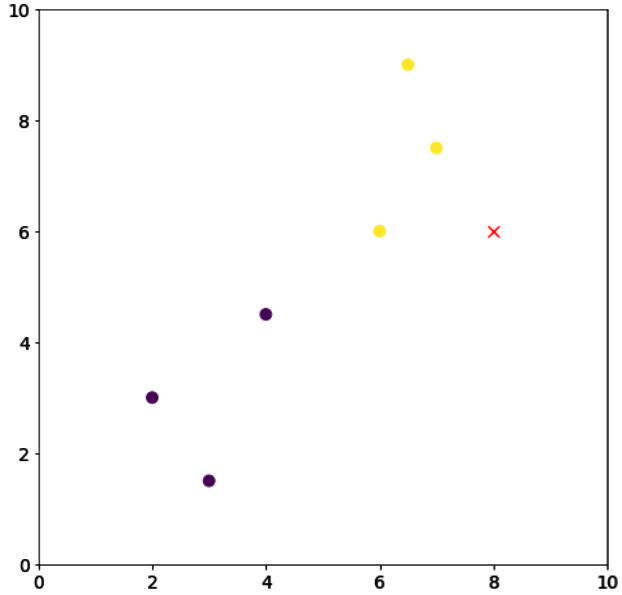
Mohd Halim Mohd Noor, PhD

Outline

- Introduction
- Generating Diverse Learners
- Model Combination Schemes
- Voting
- Bagging
- Random Forest
- Boosting
- Stacking

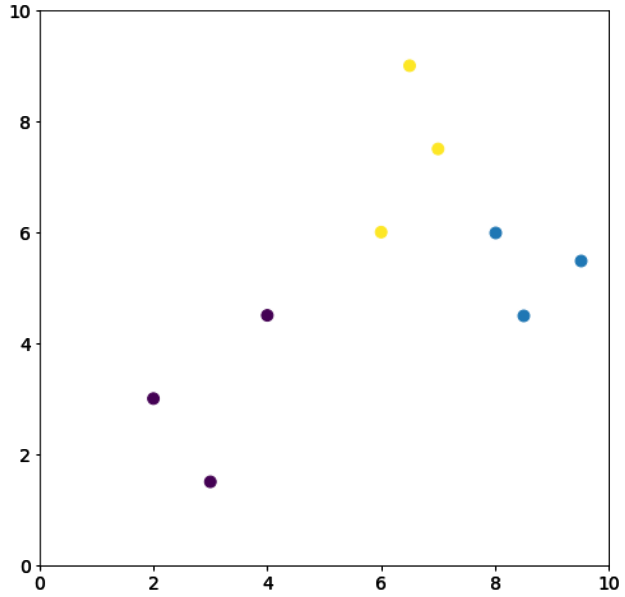
Introduction

- A prediction model is a simplified representation of the classification or regression problems
- The simplifications are based on assumptions where it may hold in some situations but may not hold in other situations
- Assume we see a sequence of random numbers, $x = 1,3,9, \ldots$ What would be the next number?
- What would be class of the sample x?



Introduction

- What would be the next number? The answer most probably 27
- But the sequence of numbers could be the output of a random number generator
- What would be class of the sample x? The answer most probably class 'Yellow’
- But it could belong to another class



Introduction

- The only reason machine learning works because of the assumptions we make about the problem
- We call these assumptions the model
- The assumptions may hold in some situation, but it may not hold in other situations especially in learning real-world problems

Introduction

- There is no single model that will always do better than any other model in every problem – No Free Lunch Theorem[a]
  - e.g. SVM is not always better than DT or DT is not always better than NB etc.
- It is common to try multiple models and find one that works best for a particular problem
- Fine-tune the model's parameters to get the highest possible accuracy on a test set

[a] To learn more, read this

Introduction

- The best model might fail or not accurate enough
- There might be another best model or learning algorithms that is more suitable

- How do we improve the accuracy of the best model?

Introduction

- By combining the strengths of multiple *base-learners*, accuracy can be improved
- Crowd is *smarter* than the individuals in the crowd
- How do we generate base-learners that complement each other?
- How do we combine the outputs of base-learners for maximum accuracy?

Generating Diverse Learners

- Diverse refers to learners that differ in their decisions
- Aim: to find a set of diverse learners that will complement each other
- How do we generate a diverse learners?

Generating Diverse Learners

- Train different base-learners using different learning algorithms
- Different algorithms make different assumptions about the data and lead to different classifiers
- Combination of parametric and nonparametric classifiers
- Train the base-learners using the same learning algorithm but with different hyperparameters
- k in k-nearest neighbour, number of hidden layers in a artificial neural network, kernel function in support vector machine etc.

Generating Diverse Learners

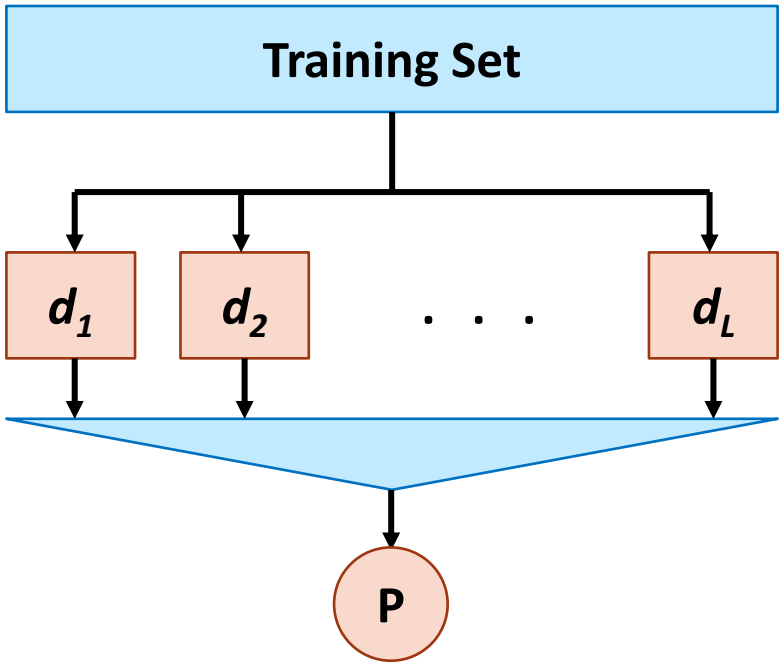
- Train the base-learners using different input representations of the same object or event
- There are multiple ways of extracting information from the object or event
- Make separate predictions based on different sources using separate base-learners
- In speech recognition to recognize the uttered words
- Acoustic input, video image of the speaker's lips – sensor fusion
- Train different base-learners using different subsets of the training set
- Subsets of training set are generated by drawing randomly from the training set; this is called bagging
- Train the base-learners serially so that instances on which the preceding classifiers are not accurate are given more emphasis in training later classifiers; this is called boosting

Model Combination Schemes

- Voting
- Bagging
- Boosting
- Stacking
- Cascading

Voting

- The base-learners work in parallel
- Combining the predictions from base-learners by voting



Voting

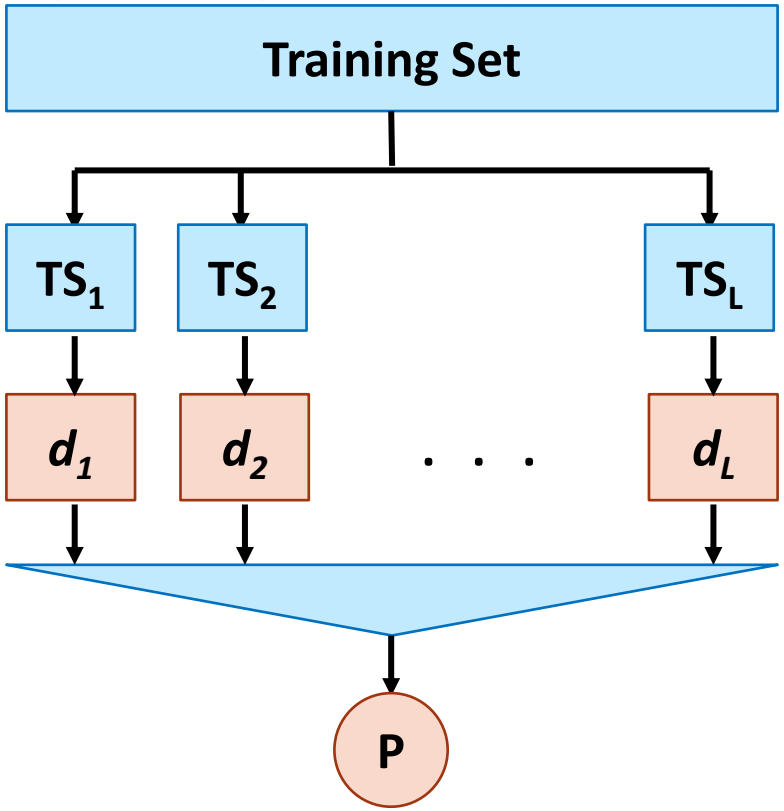
- All learners are given weights, w_j where $w_j \geq 0, \sum_j w_j = 1$
- Calculate the weighted sum, $y_k = \sum_j w_j d_{jk}$
- Simple voting – the weights have equal values, $w_j = \frac{1}{L}$
- Other combination rules
- Sum: $y_k = \frac{1}{L} \sum_{j=1}^{L} d_{jk}$
- Median: $y_k = \underset{j}{\text{med}}\, d_{jk}$
- Minimum: $y_k = \min_j d_{jk}$
- Maximum: $y_k = \max_j d_{jk}$

Voting

- In the case of regression, (weighted) averaging or median can be used to fuse the outputs of base-regressors

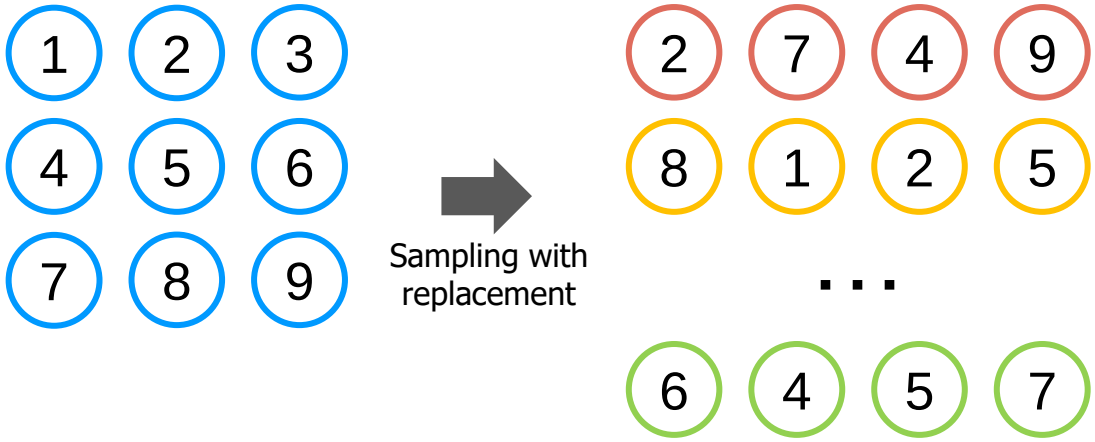
Bootstrap AGGregatING (Bagging)

- The base-learners are trained using different training sets
- The learners are aggregated using their average or a voting system
- An ensemble model with lower variance
- The different training sets are generated by bootstrapping



Bootstrapping

- A resampling technique – selecting samples from original training set and using these samples for estimating various statistics or model accuracy
- Given a training set, X of size N
- Draw M samples randomly from X with replacement (same samples can be drawn more than once)
- Estimate the statistic of interest e.g. mean, median, variance etc.
- We can generate smaller training sets of size $M < N$ if the original training set is large

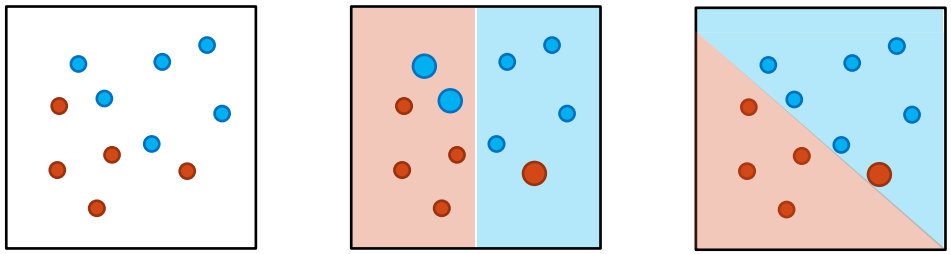


Random Forest

- A supervised learning algorithm that builds a large collection of decision tree using bagging method
- A subset of features (columns) is used (randomly selected) for splitting a node
- For b = 1 to B:
  - Draw a bootstrap sample X^* of size N from the original training set
  - Train a random-forest tree (decision tree) d_b using X^* by recursively repeating the following steps for each terminal node of the tree, until the minimum node size is reached
    - Select m variables at random from the p variables
    - Pick the best variable/split-point among the m
    - Split the node into two daughter nodes
- To make prediction
  - Regression: $d_{\mathrm{rf}}(x) = 1/B \sum_{b=1}^{B} d_b(x)$
  - Classification: $d_{rf}(x) = \text{majority vote}\{d_b(x)\}_1^B$

Boosting

- Training of models are done sequentially
- Each learner in the sequence is fitted giving more importance to samples that were misclassified in the previous learners
  - Focus on efforts on the most difficult samples
  - At the end obtain a strong ensemble learner with low bias
- Train weak learners that has an error rate less than 0.5 (otherwise it would be random guessing) but not always accurate
  - e.g one-level decision tree



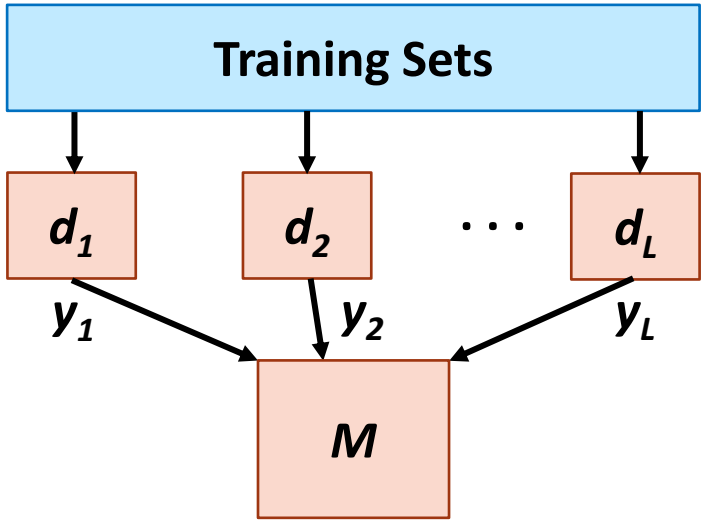
d_1 d_2 d_L

Stacking

- Combine the prediction of (diverse) weak learners into a single learner known as meta-model
- The meta-model is trained to output prediction based on the predictions returned by the weak learners

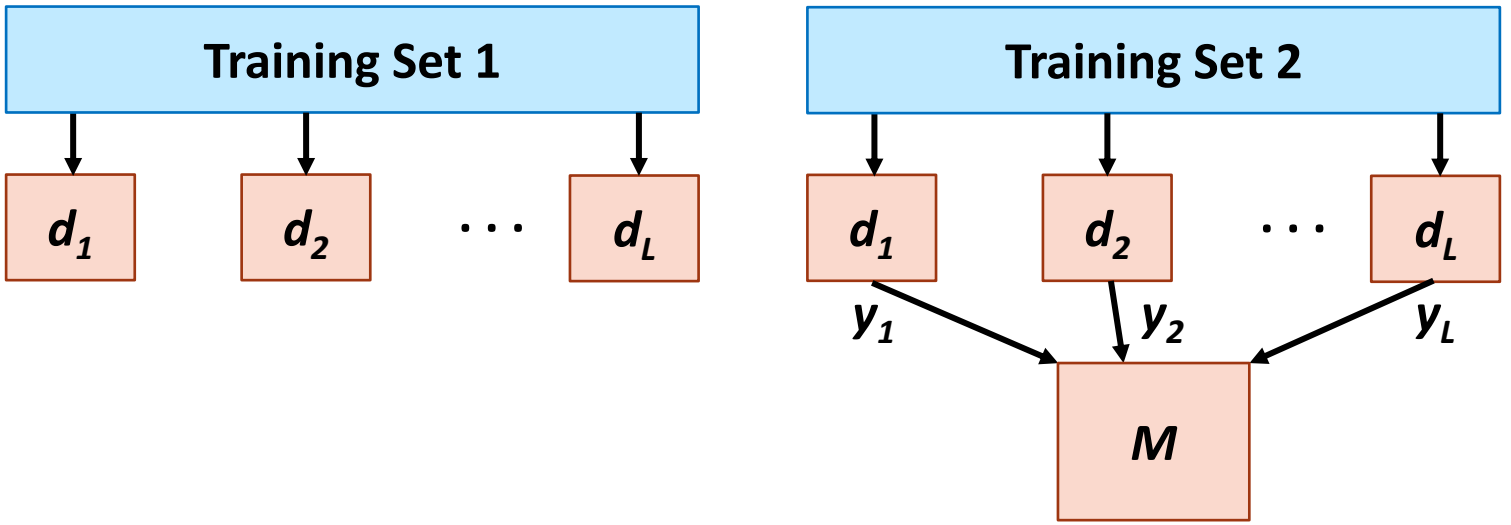
Stacking

- Weak leaners takes the data as inputs
  - Decision tree
  - K-nearest neighbours
  - Naïve bayes
- Meta-model: neural network takes the predictions of weak learners as inputs to output final prediction



Stacking

- Split the training data into two groups, D_1 and D_2
- Train weak learners, d_l using D_1 where $l = 1, 2, \ldots, L$
- For each weak learners, make predictions, y_l for each sample in D_2
- Train the meta-model on D_2 using y_l (predictions made by weak learners) as inputs



Summary

- To divide a complex task into simpler tasks that are handled by separately trained base-learners
- base-learners should be diverse and accurate – they should provide useful information
- A base-learner should be discarded if a base-learner does not add to accuracy
- One of two base-learners is not needed if they are correlated
- An ensemble classification system is not interpretable
- Ensemble methods are usually computationally expensive
  - Long learning time
  - More parameters

UNIVERSITI SAINS MALAYSIA

End