COMP9417

# Regression

# Ensemble Learning

## Simple ensembles: combining several learning algorithm

* Majority vote or unweighted average will be used for prediction
* Using weighted average or weighted votes to predict the output
* Treat the output of each algorithm as a feature and train another learning algorithm on them

## Mixture of experts

* Each learning algorithm defines which indicated expertise of that algorithm for that particular location of in the input space
* It may use a weighted average or just pick the model with the largest expertise

## Bagging method: (“Bootstrap Aggregation”)

* Training many classifiers, but each on a Bootstrapped dataset
  + Bootstrap: Create a random subset of data by sampling with replacement
  + Bagging: Repeat times to generate subsets
* Then aggregate through model averaging / majority voting
* Bagging is applied on a collection of low-bias high variance models
  + By averaging them the bias would not get affected
  + By averaging them the variance will be reduced

## Add randomization to the models to introduce more diversity in the models for example

* For every model use a subset of features, selected randomly, e.g. in Random Forest(it can also help with training time)
* For algorithms that are dependent on initial weights, use different random initial weights

## Boosting: A sequence of weak learners, each trying to correct its predecessor

* Learners are trained sequentially
* New learners focus on errors of earlier learners
* New learners try to get misclassified samples right by operating on a weighted training set in favor of misclassified instances
* Combine all learners in the end using weighted majority / weighted average of k learners
* AdaBoost is a boosting algorithm using stump trees
  + Misclassified instances gain higher weights
  + Correctly classified instances lose weight
* Main advantages:
  + Use very simple(weak) learners
  + It boost the performance
  + Decrease bias
  + Decrease variance
* Slow during training and lack of interpretability
* Gradient Boosting is a boosting algorithm using stump tree for regression:
  + At every step models the residuals