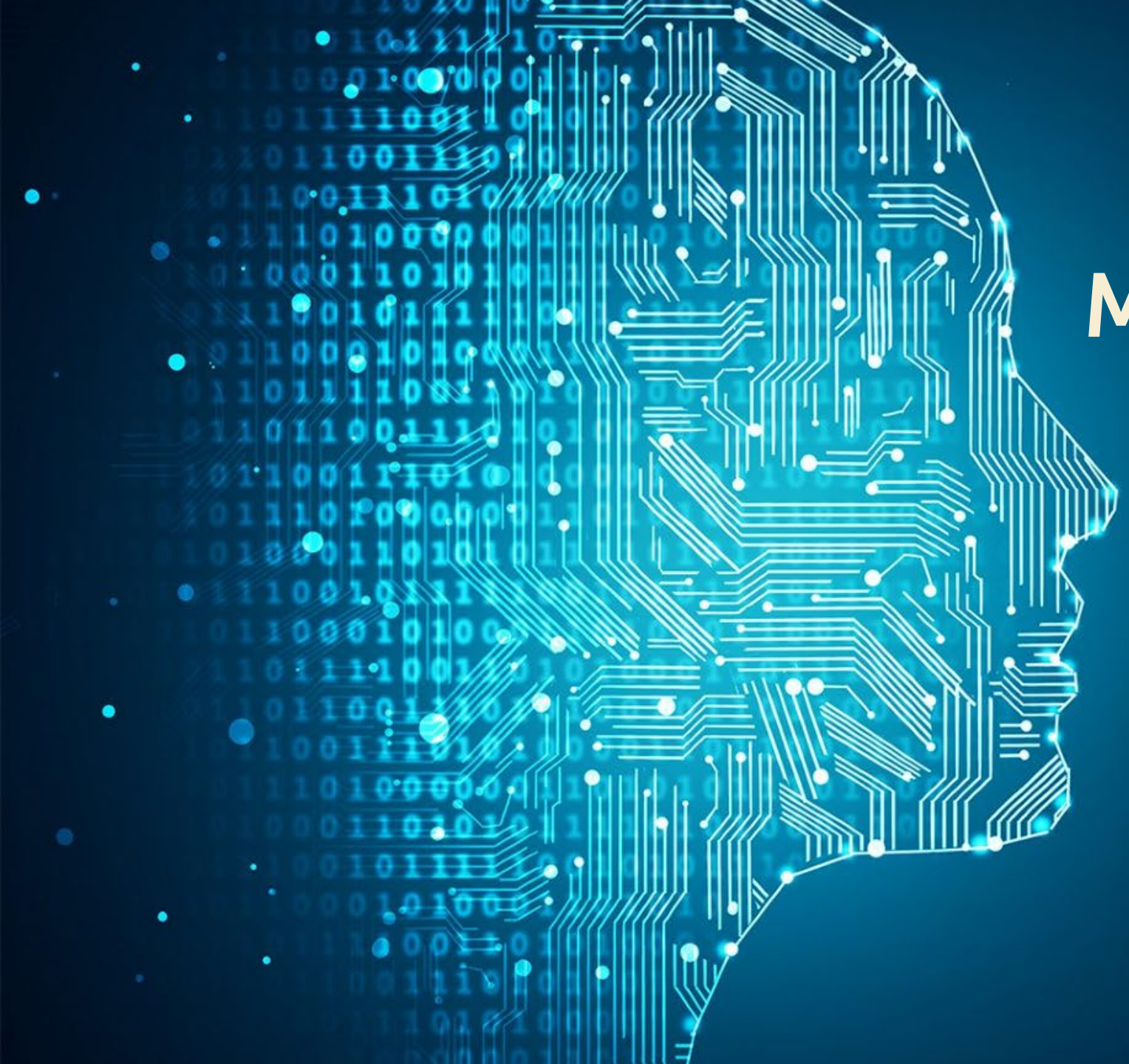


Applied Machine Learning

(BUAN 6341)

Lecture 6
Ensemble Models

Yingjie Zhang

University of Texas at Dallas

yingjie.zhang@utdallas.edu

Spring 2019

Group Project

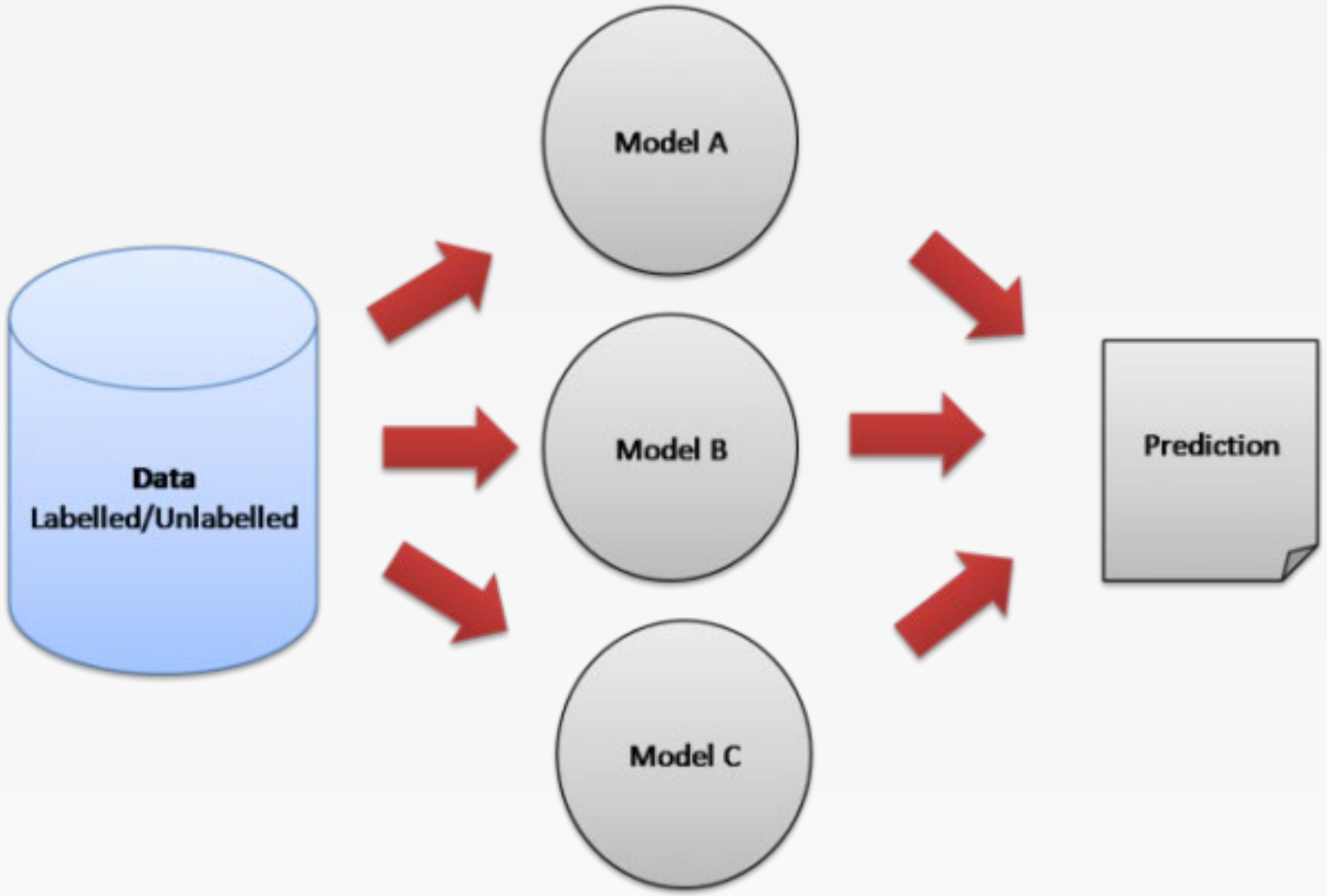
- 1-page proposal due on March 10, 2019 11:59pm

Agenda

- Ensemble Models
- Review for Exam I

Ensemble Models

Wisdom of the Crowd



Netflix Prize Competition

- Training data is a set of users and ratings (1,2,3,4,5 stars) those users have given to movies.
- Construct a classifier that given a user and an unrated movie, correctly classifies that movie as either 1, 2, 3, 4, or 5 stars
- $1 million prize for a 10% improvement over Netflix's current movie recommender

BellKor / KorBell

And, yes, the top team which is from AT&T…

“Our final solution (RMSE=0.8712) consists of blending 107 individual results. “

--	No Progress Prize candidates yet	--	--
Progress Prize - RMSE <= 0.8625			
1	BellKor	0.8705	8.50
Progress Prize 2007 - RMSE = 0.8712 - Winning Team: KorBell			
2	KorBell	0.8712	8.43
3	When Gravity and Dinosaurs Unite	0.8717	8.38
4	Gravity	0.8743	8.10
5	basho	0.8746	8.07
6	Dinosaur Planet	0.8753	8.00
7	ML@UToronto A	0.8787	7.64
8	Arek Paterek	0.8789	7.62
9	NIPS Reject	0.8808	7.42
10	Just a guy in a garage	0.8834	7.15
11	Ensemble Experts	0.8841	7.07
12	mathematical capital	0.8844	7.04
13	HowLowCanHeGo2	0.8847	7.01
14	The Thought Gang	0.8849	6.99
15	Reel Ingenuity	0.8855	6.93
16	strudeltamale	0.8859	6.88
17	NIPS Submission	0.8861	6.86
18	Three Blind Mice	0.8869	6.78
19	TrainOnTest	0.8869	6.78
20	Geoff Dean	0.8869	6.78
21	Rookies	0.8872	6.75
22	Paul Harrison	0.8872	6.75
23	ATTEAM	0.8873	6.74
24	wxyzconsulting.com	0.8874	6.73
25	ICMLsubmission	0.8875	6.72
26	Efratko	0.8877	6.70
27	Kitty	0.8881	6.65
28	SecondaryResults	0.8884	6.62
29	Birgit Kraft	0.8885	6.61

Ensemble Methods

- Why it works:
  - Diversity!
  - Image that we have 5 completely independent classifiers; each of them individually is correct 70% of the time
    - Prob(correctly classify a record by a majority vote)

$$= C_{(5,3)}(0.7)^3(0.3)^2 + C_{(5,4)}(0.7)^4(0.3)^1 + C_{(5,5)}(0.7)^5 = 0.837$$

- Downside:
  - Increased complexity, more difficult to interpret
  - Does not always guarantee performance improvements

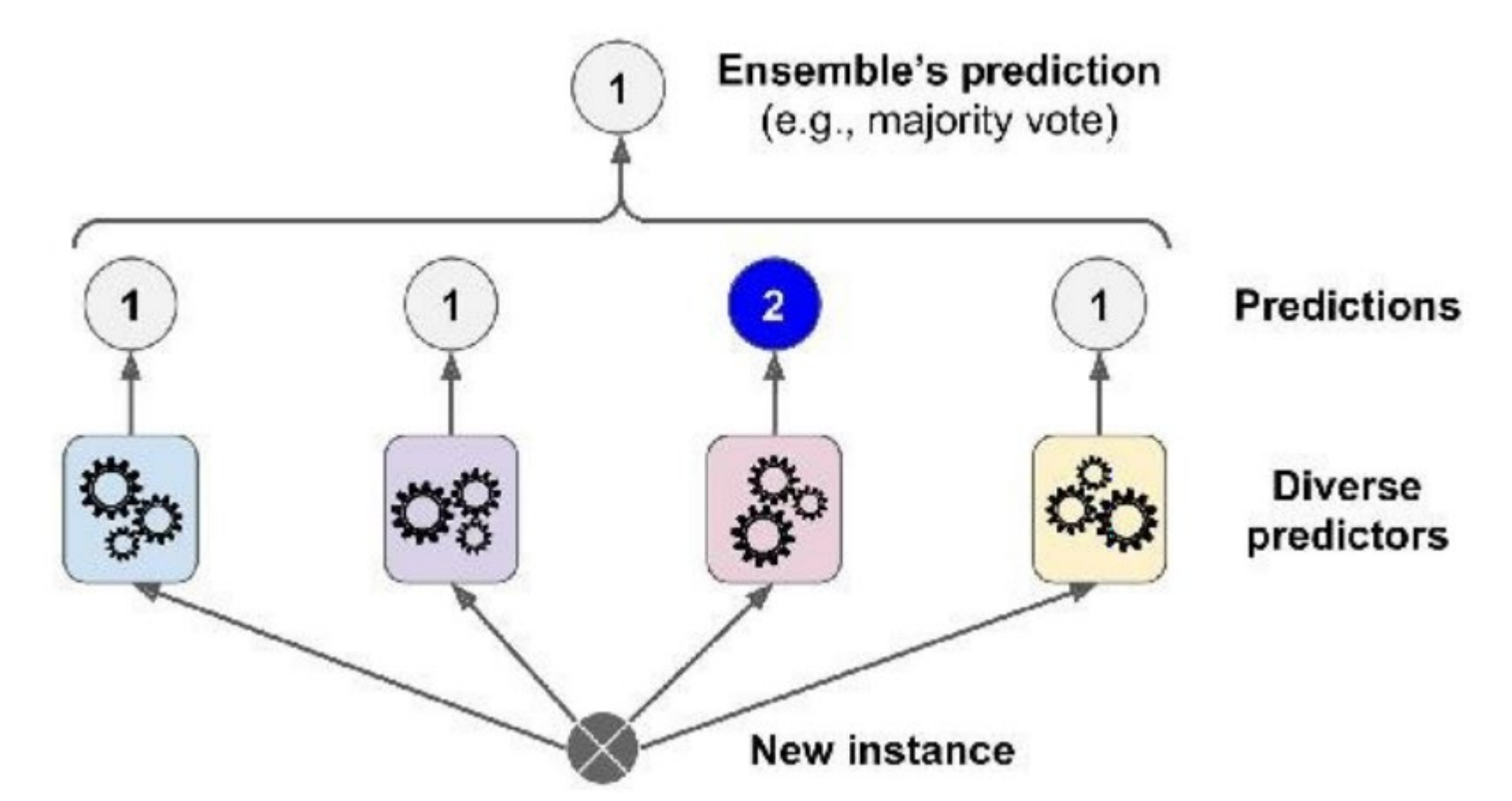
Ensemble Methods

- Voting Classifiers
- Bagging
- Boosting
- Stacking

Ensemble Methods

- Voting Classifiers
- Bagging
- Boosting
- Stacking

Voting Classifiers

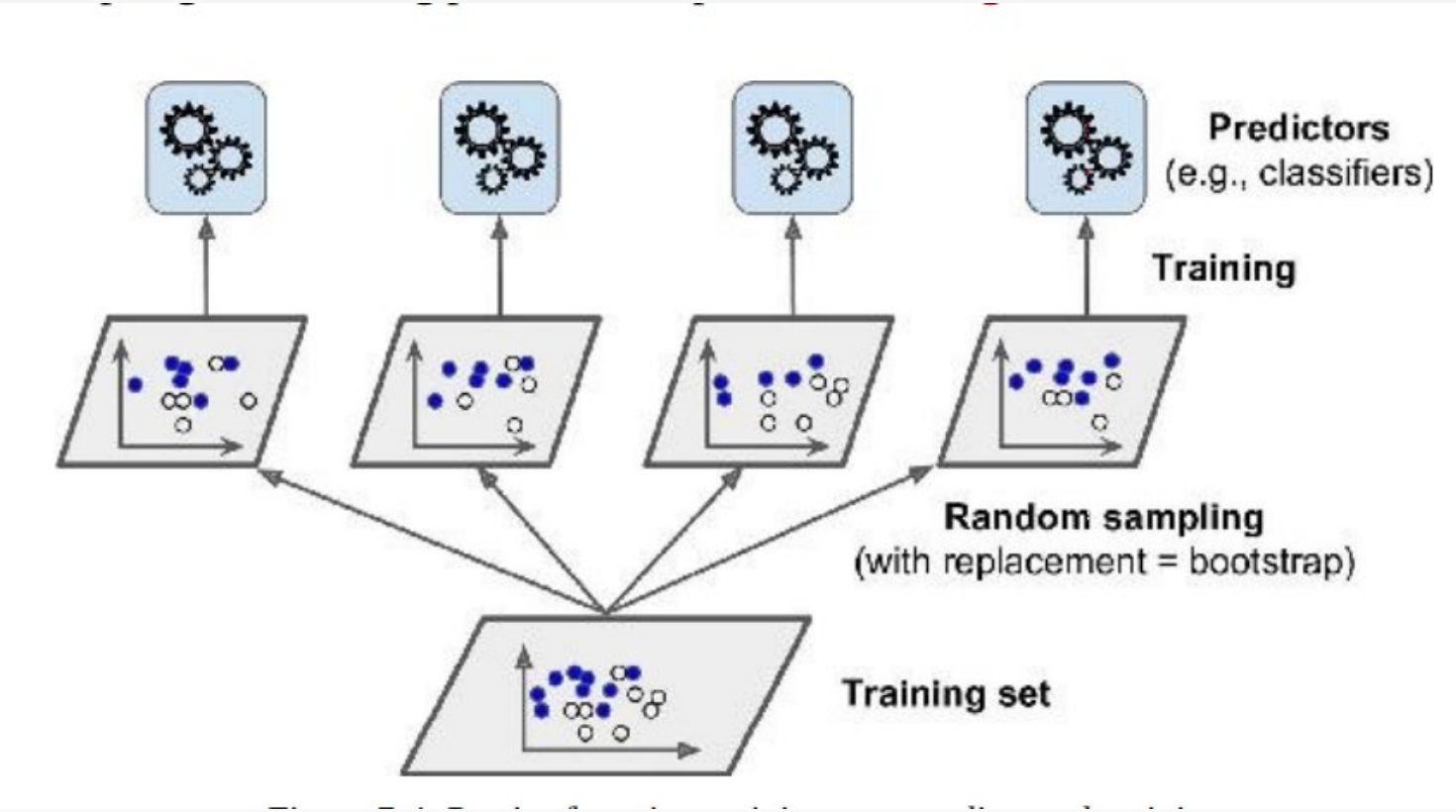


Ensemble Methods

- Voting Classifiers
- Bagging (Pasting)
- Boosting
- Stacking

Bagging: Bootstrap Aggregation

- Ideas:
  - Use the same training algorithm for every predictor, but to train them on different random subsets of the training set



Bagging

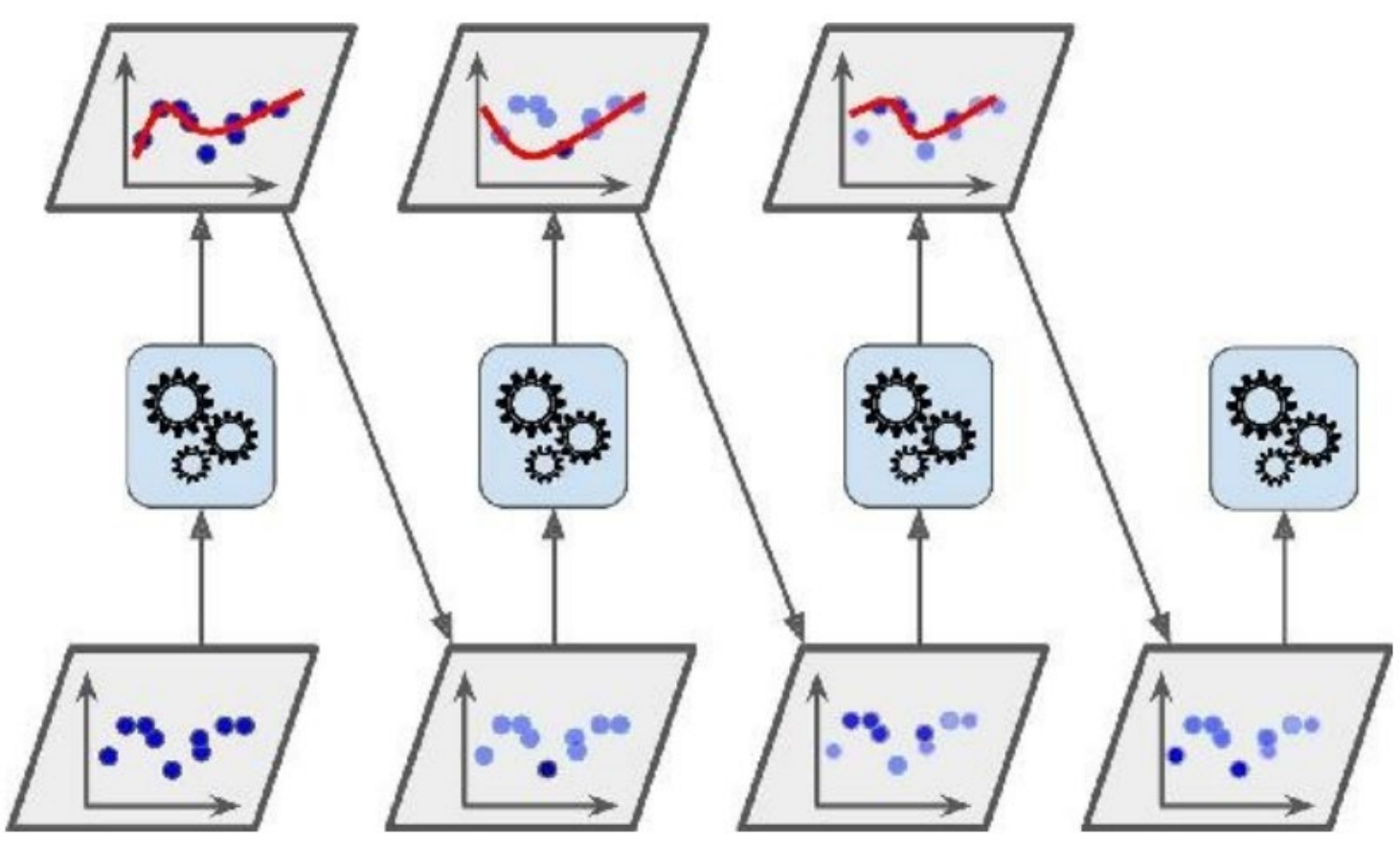
- Given
  - Labelled dataset
  - Specific predictive modeling techniques
- Train k models on different training data samples
  - Bootstrap samples: sampled with replacement, typically of the same size as the original training data
- Final prediction is done by combining (i.e., majority vote, averaging) the predictions of k individual models

Ensemble Methods

- Voting Classifiers
- Bagging
- Boosting
- Stacking

Boosting

- AdaBoost



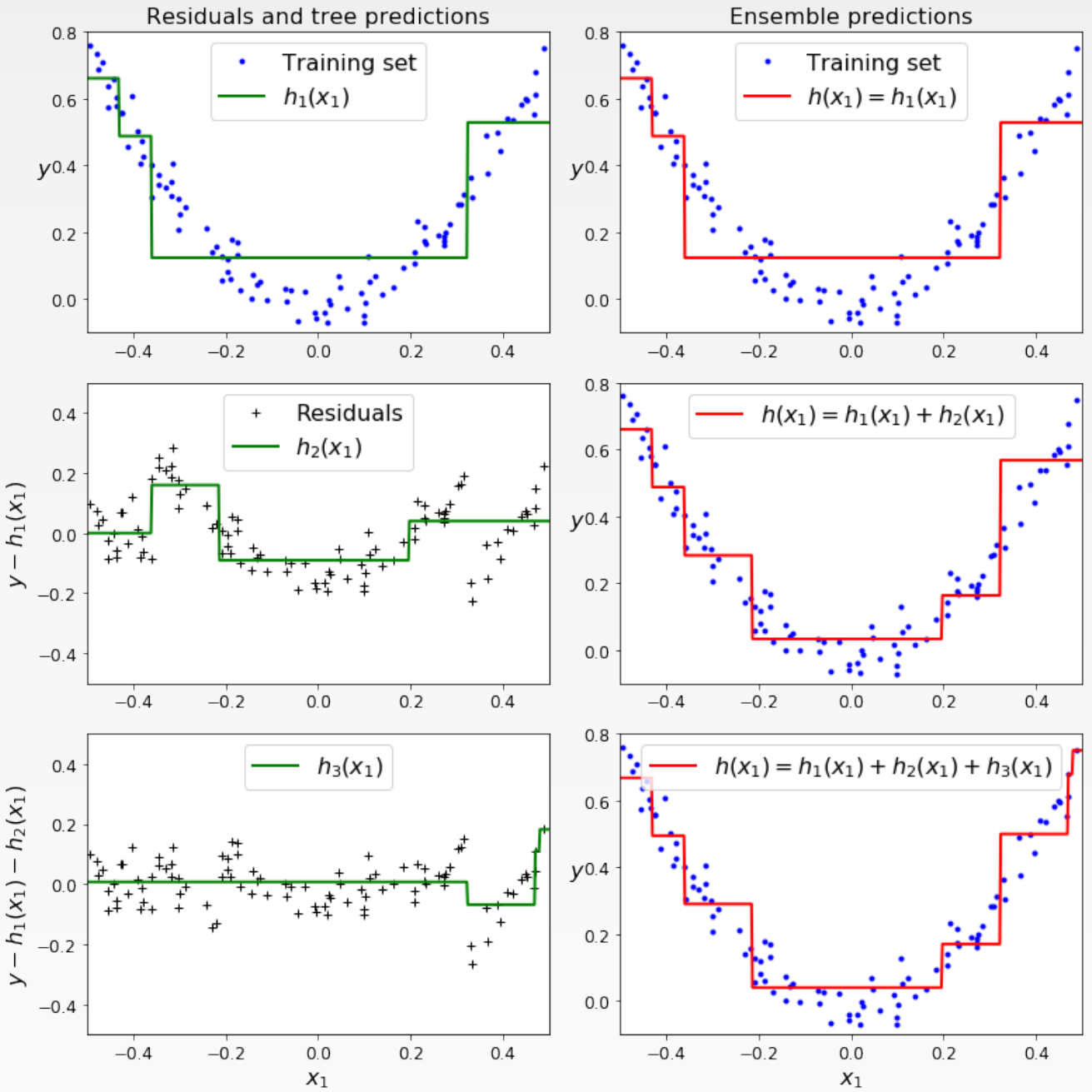
AdaBoost

- In each iteration
  - Build model on the sample (with replacement) from data
  - Evaluate the model on the original training data
  - Increase the weights of data points on which the current model makes misclassifications (so that subsequent iterations have higher chance to choose these data points for the sample)
  - Decrease the weights of other data points

- Final prediction is done by the weighted combination of the predictions of k individual models

Gradient Boosting

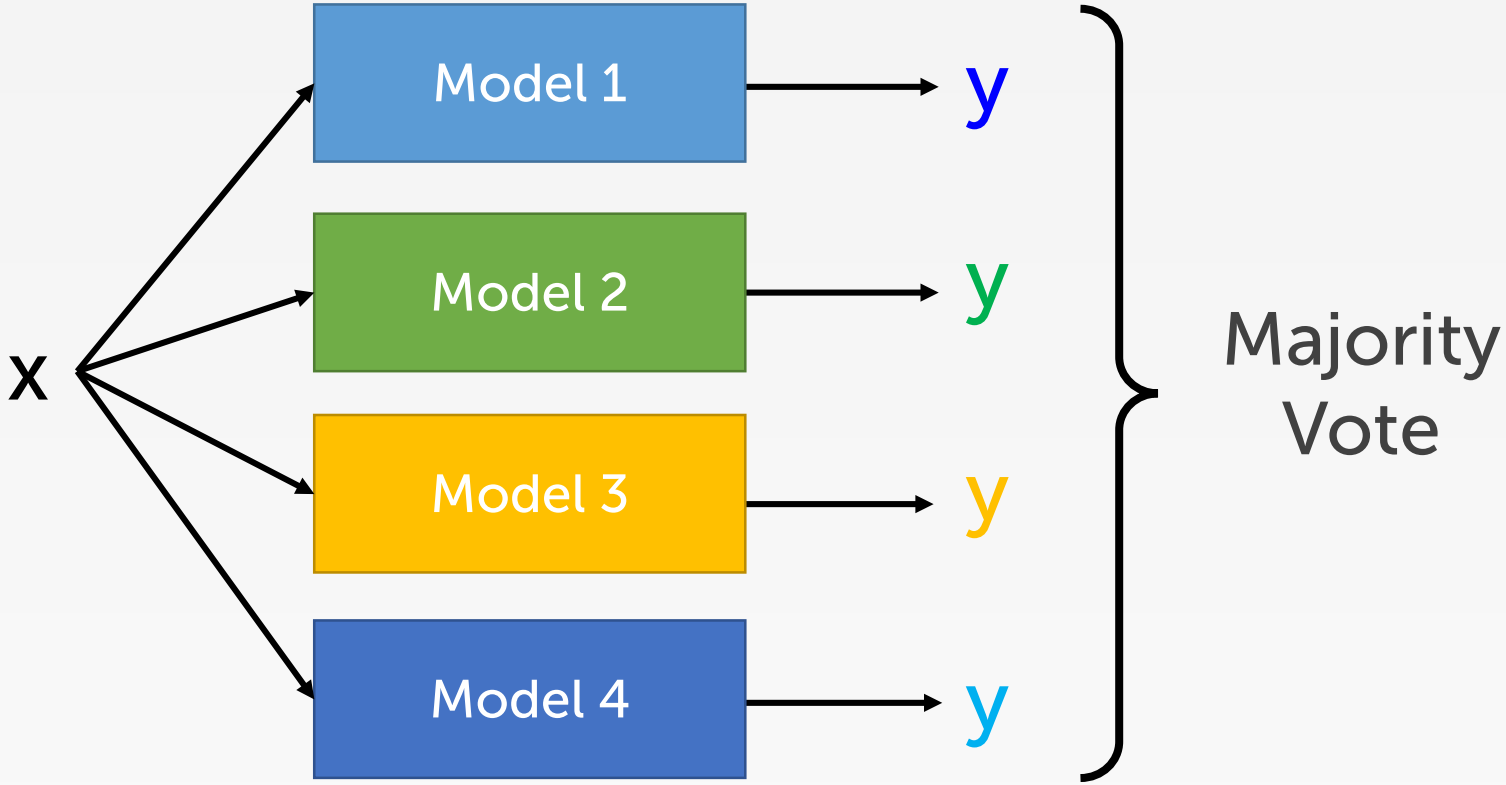
- AdaBoost: tweak the instance weights at each iteration
- Gradient Boosting: fit the new predictor to the residual errors made by the previous predictor


Residuals and tree predictions
Ensemble predictions
Training set
$h_1(x_1)$
$h(x_1) = h_1(x_1)$
Residuals
$h_2(x_1)$
$h(x_1) = h_1(x_1) + h_2(x_1)$
$h_3(x_1)$
$h(x_1) = h_1(x_1) + h_2(x_1) + h_3(x_1)$
y
$y - h_1(x_1)$
$y - h_1(x_1) - h_2(x_1)$
x_1

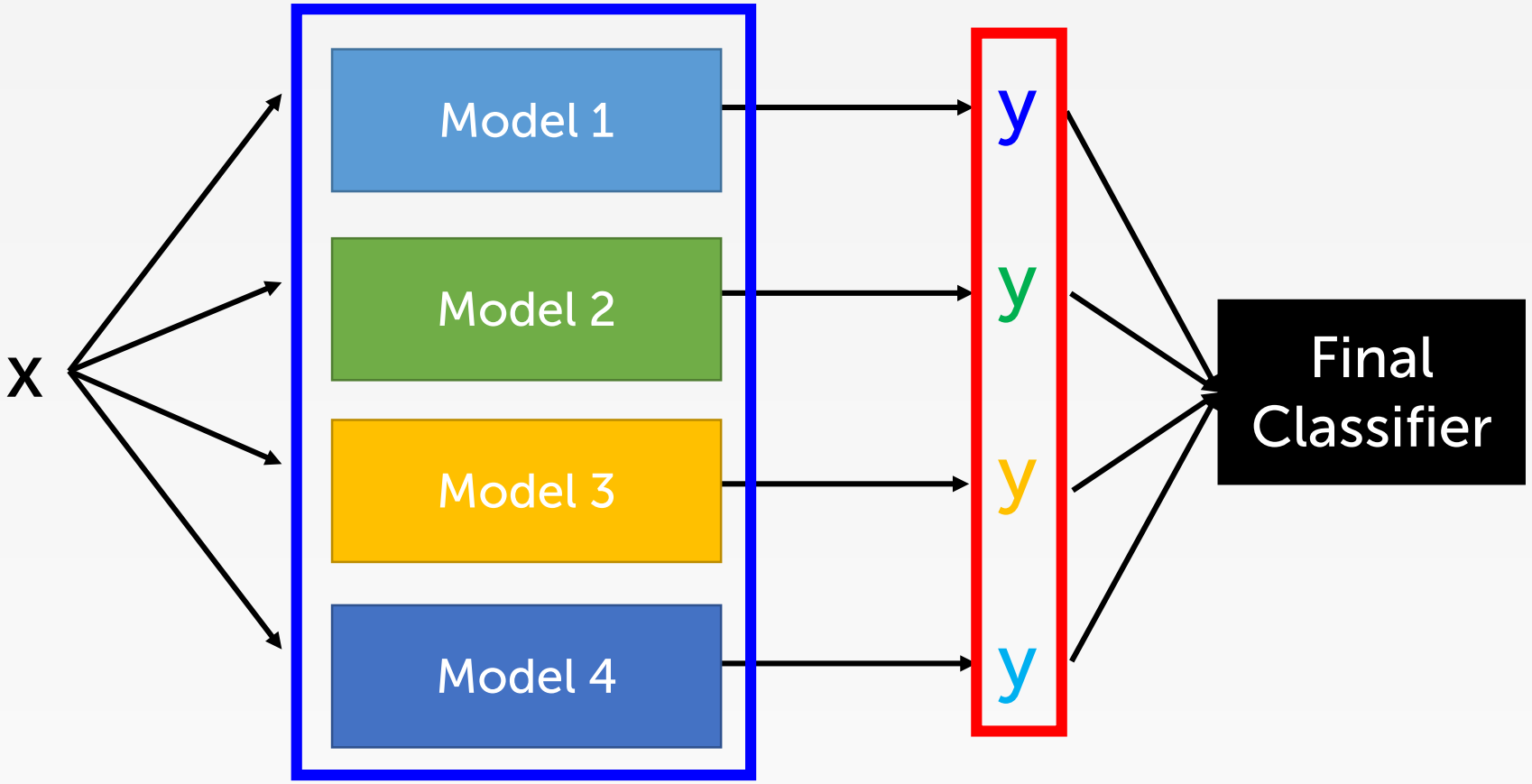
Ensemble Methods

- Voting Classifiers
- Bagging
- Boosting
- Stacking

Stacking Voting



Stacking



Random Forest

Overview

- Definition
  - Collection of unpruned trees
  - Rule to combine individual tree decisions
- Purpose
  - Improve prediction accuracy
  - Improve efficiency
- Principle
  - Encouraging diversity among the tree
- Solution: randomness
  - Bagging
  - Random decision trees

Details

- Build many “random” trees
- Randomness: using only a random sample of m attributes to calculate each split
- For each tree:
  - Choose a different training sample
  - For each node, choose m random attributes and find the best split
  - Trees are often fully grown (not pruned)
- Predication: majority vote among all the trees

Python Practice

Exam I

Logistics

- **February 28, 2019, 1-3pm (2 hours)**
- **Location: JSOM 2.714**

- Calculator is allowed.
- Closed-book
- Closed-notes

Review –General Ideas of ML

- How to define a learning problem?

- Differences between (among):
  - Supervised vs. Unsupervised learnings
    - Examples? Applications? Specific models?
  - Training data, Validation data, and Test data
  - Classification vs. Regression

Review -- Evaluations

- Overfitting?
  - What is overfitting issue?
  - Model feasibility
- Cross validations:
  - How does it work?
  - Why do we need it?
- Confusion matrix and classification performance measures
  - Accuracy, precision, recall, and specificity (why?)
  - Applications in the real world
  - Trade-offs between precision and recall
- ROC and precision-recall curve
  - X- and y-axis
  - Benchmarks in ROC curve

Review -- Regression

- Linear regressions:
  - Definitions? (what is a linear regression?)
  - What is the goal when training a linear regression?
  - How to estimate a linear regression? (2 techniques)
- Polynomial regressions:
  - Two steps: (polynomial features + linear regression)
  - Overfitting issues
- Ride and Lasso:
  - Goal?
  - What is λ? (too small? too large? $\lambda = 0$, $\lambda = \infty$)
  - Differences? (function forms and variable selections)
  - Penalty only works in the estimation process

Review -- Classification

- Logistic regression is a type of classification model
- KNN:
  - General process
  - Three keys (distance, k, and aggregations)
  - K values vs. overfitting/underfitting
  - Advantages vs. disadvantages
- Naïve Bayes:
  - General process
  - Advantages and disadvantages

Review – DT & SVM

- Decision tree:
  - General process using ID3 algorithm
    - Important concepts: Entropy (formula), information gain
  - How to avoid/eliminate overfitting issues?
  - Prune trees (why?)

- Support vector machine:
  - How does it work? (linear SVM)
  - Support vectors
  - Hard vs. soft margin (differences? Model parameters C: small or large?)
  - RBF SVM: how γ works?

Review – Optimizations

- Gradient descent
  - How does it work?
  - What is the learning rate? (if too small? Or to big?)
  - Advantages vs. disadvantages?

- Grid search:
  - Potential drawbacks?

Questions

- 8 True/false questions (with explanations) [20 points]
- 10 multiple choice questions [40 points]
- 4 short-answer questions [40 points]

Practice

True/False. If false, explain reasons or correct the statement

SQ1. When growing a decision tree, attributes can be used more than once.

SQ2. We cannot use Naïve Bayes classifier if we have a dataset with mixed attributes, containing both categorical and numerical attributes.

SQ3. The classification accuracy increases as k increases in k-NN classifiers

Practice

Multiple Choice Questions

SQ4. Which of the following statements are true?

A. SVM is a probability-based supervised learning model

B. Entropy is always non-negative

C. Naïve Bayes is not proper if we have a data set where all attributes are highly dependent on each other

D. Validation sets are required to achieve better performance of any machine learning methods.

Practice

Multiple Choice Questions

SQ5. The points on a model's Precision-Recall curve represent

A. the performance at different thresholds

B. the accuracy of different training sets

C. the cost of different regularization parameters

D. the generalization performance by increasing model complexity

Practice

Multiple Choice Questions

SQ6. Which technique(s) would be useful for the following business problem? "Predict whether a UTD alumni is likely to donate"

A. Linear Regression

B. Decision Tree

C. Unsupervised learning models

D. Logistic Regression

Practice

Multiple Choice Questions

SQ7. Given the following two linear regression models

$M1: y = ax + c$ $M2: y = ax + bx^2 + c$

Which of the two models is more likely to fit the training (test) data better?

A. M1

B. M2

C. both will fit equally well

D. impossible to tell

Practice

GPA	Studied	Passed
L	T	F
L	T	T
M	F	F
M	F	T
H	T	T

$\log_2 0.2 = -2.32$

$\log_2 0.3 = -1.73$

$\log_2(2p) = 1 + \log_2 p$

Short-answer questions

SQ8.1. What is the entropy before splitting?

SQ8.2. What is the entropy if we split on GPA?

SQ8.3. Draw the complete tree