The Wisdom of the Crowd: Ensemble Methodology

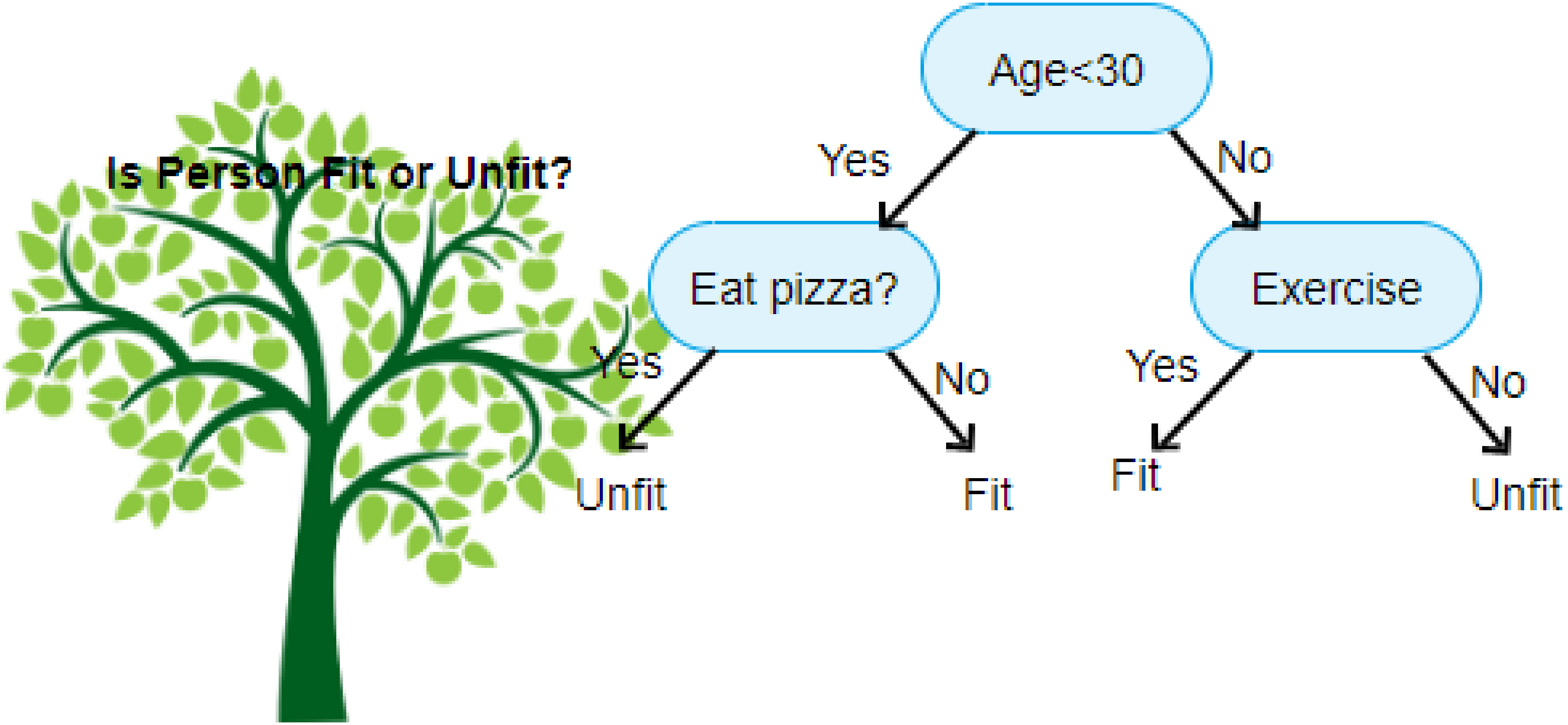
Alexander G. Ororbia II

Introduction to Machine Learning

CSCI-635

11/10/2023

Decision Trees – Interrogating Your Data!



On Continuous Variables & Gain Ratio

(Yay, More White Board!)

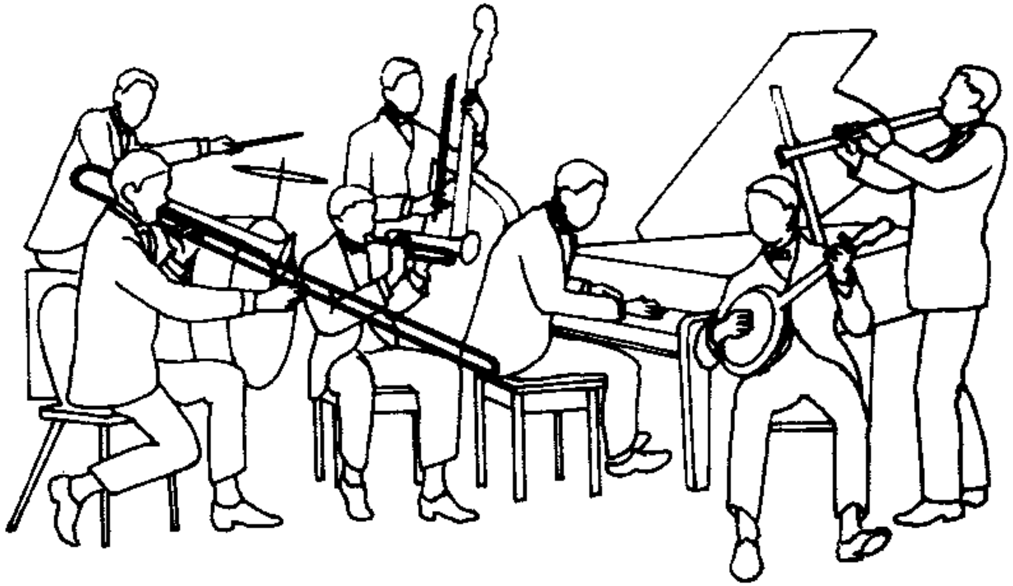
Class-Labeled Training Tuples from the *AllElectronics* Customer Database

RID	*age*	*income*	*student*	*credit_rating*	*Class: buys_computer*
1	youth	high	no	fair	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	fair	yes
4	senior	medium	no	fair	yes
5	senior	low	yes	fair	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	fair	no
9	youth	low	yes	fair	yes
10	senior	medium	yes	fair	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	fair	yes
14	senior	medium	no	excellent	no

Ok, but is one tree enough?

Ok, but is one tree enough?

No!



An ensemble is kinda like a garage band…

Some aspiring machine learning scientists can be musicians too, you know!

Good ol’ Led Zep!

The Wingmen

Ensembling: Bagging and Boosting

- Ensembles can be very powerful
  - Put together all classifiers (weak + strong learners)
  - ***Intuition*** = different learners will pick up on different (statistical) regularities in feature space
  - Combine different perspectives to build a more intelligent complex system (a "committee machine")
    - Ensemble = more powerful than component classifiers

- ***Bagging*** & ***boosting*** = create data subsets through sampling with replacement
  - Build a classifier on each subset
  - Bagging, introduced by Breiman (1996), stands for "**b**ootstrap **agg**regat**ing**"
    - *Bootstrap* => sampling with replacement

Enlarging the Hypothesis Space

Combinations of Simple Algorithms: Powerful

e.g. using linear separators, can represent non-linearly separable attribute regions

- *Enlarges the hypothesis space (i.e. space of representable functions)*

Decision stumps (single decision tree nodes partitioning attributes) used frequently to learn complex functions

If base algorithm supports simple and efficient learning (e.g. decision stump), ensemble permits complex functions to be learned without high algorithmic or computational complexity

Example of an Ensemble

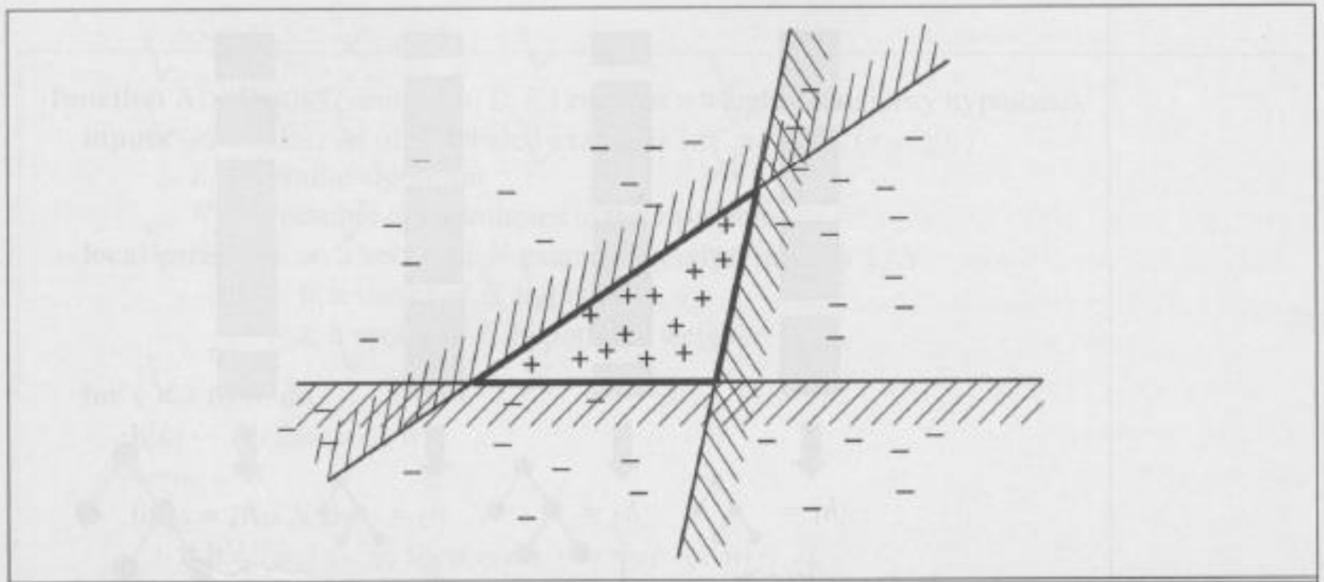
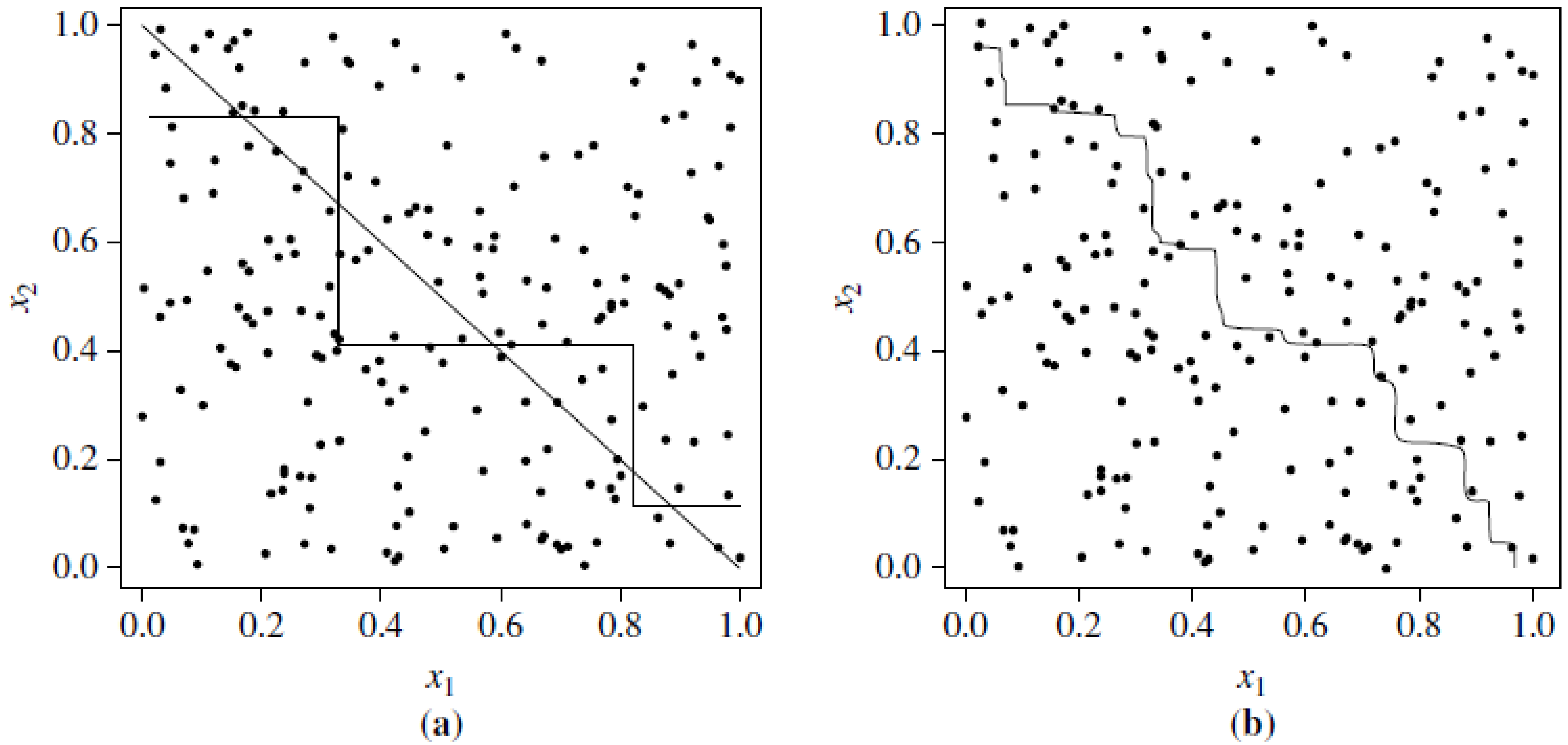


Figure 18.32 Illustration of the increased expressive power obtained by ensemble learning. We take three linear threshold hypotheses, each of which classifies positively on the unshaded side, and classify as positive any example classified positively by all three. The resulting triangular region is a hypothesis not expressible in the original hypothesis space.

An *ensemble* consisting of three linear separations of 2D data for a two-class classification problem

Ensemble Methodology (White Board)

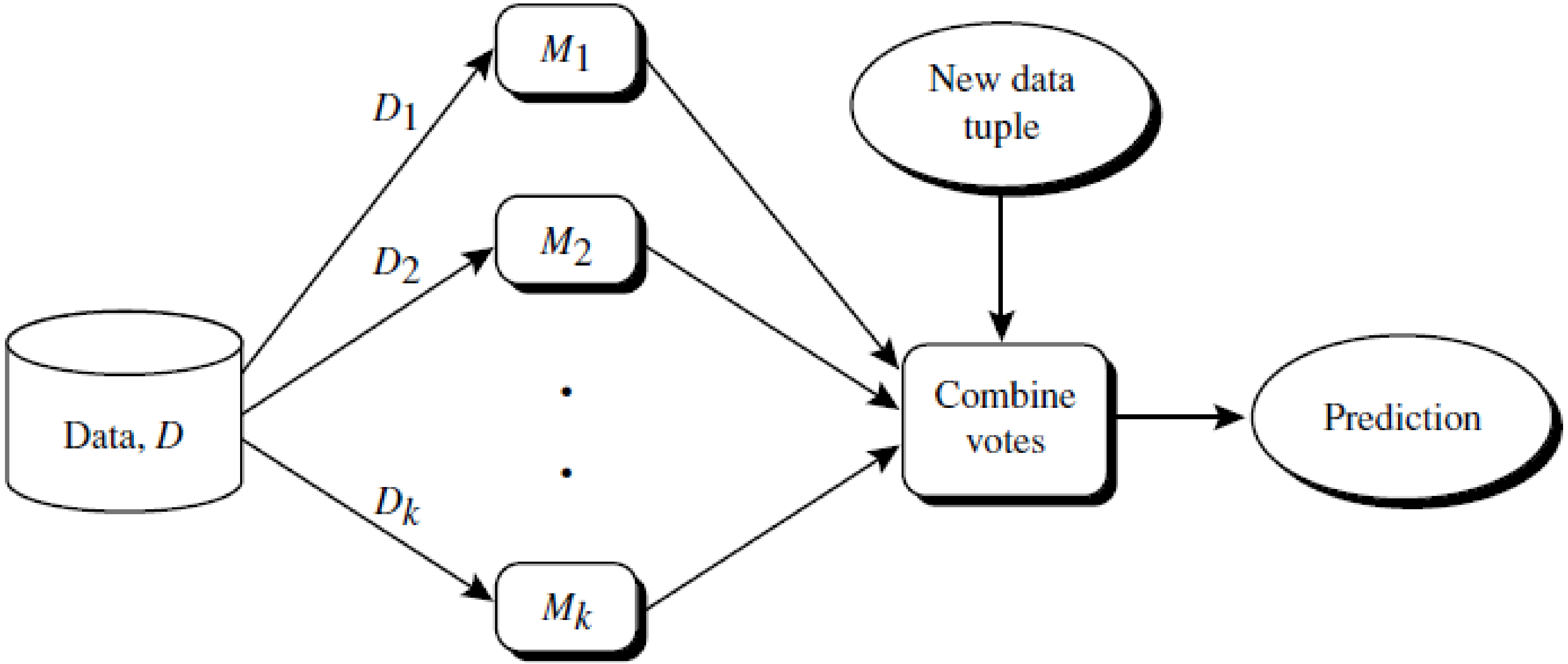
Decision Boundary of an Ensemble



Decision boundary by (a) a single decision tree and (b) an ensemble of decision trees for a linearly separable problem (i.e., where the actual decision boundary is a straight line). The decision tree struggles with approximating a linear boundary. The decision boundary of the ensemble is closer to the true boundary. *Source:* From Seni and Elder [SE10]. © 2010 Morgan & Claypool Publishers; used with permission.

Ensemble Methodology, Bagging (White Board)

A *Bagging* Ensemble System



Increasing classifier accuracy: Ensemble methods generate a set of classification models, $M_1, M_2, \ldots, M_k$. Given a new data tuple to classify, each classifier "votes" for the class label of that tuple. The ensemble combines the votes to return a class prediction.

Algorithm: Bagging. The bagging algorithm—create an ensemble of classification models for a learning scheme where each model gives an equally weighted prediction.

Input:

- D, a set of d training tuples;
- k, the number of models in the ensemble;
- a classification learning scheme (decision tree algorithm, naïve Bayesian, etc.).

Output: The ensemble—a composite model, $M*$.

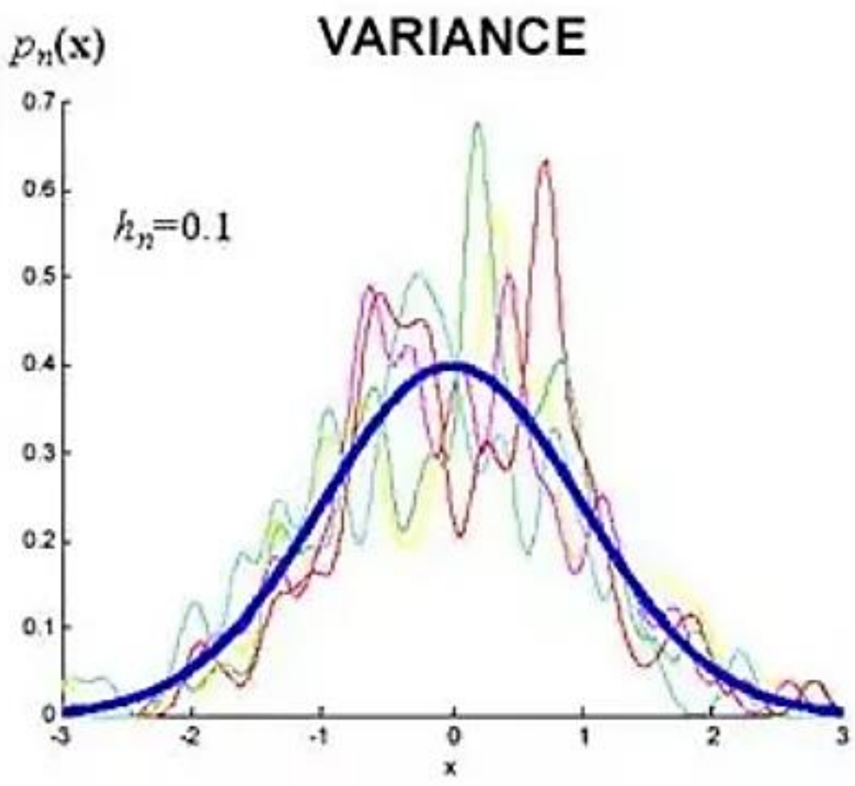
Method:

(1) **for** $i = 1$ to k **do** // create k models:
(2) create bootstrap sample, D_i, by sampling D with replacement;
(3) use D_i and the learning scheme to derive a model, M_i;
(4) **endfor**

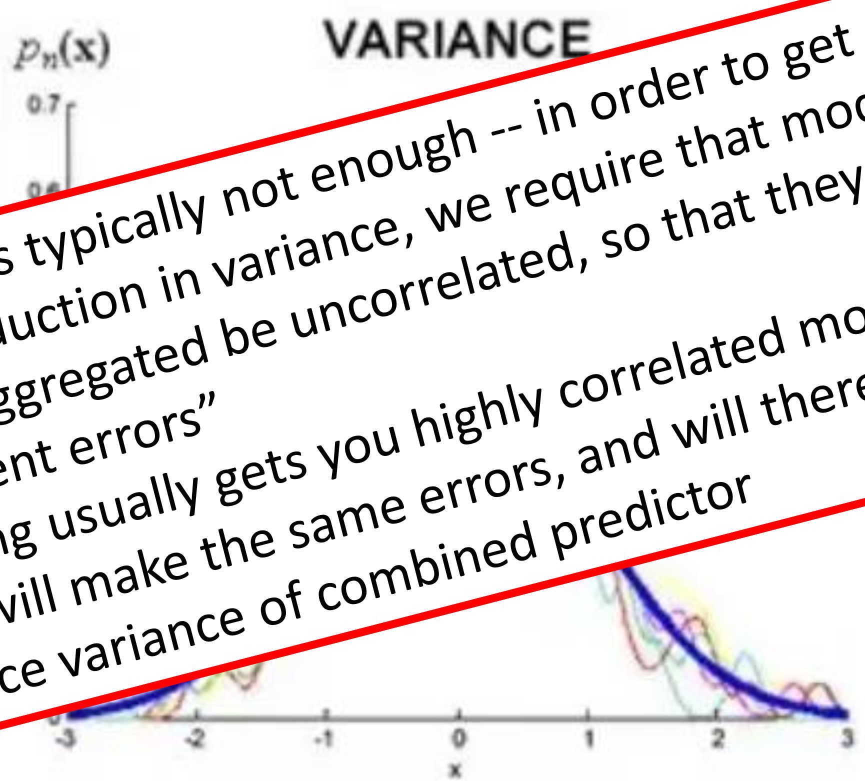
To use the ensemble to classify a tuple, X:

let each of the k models classify X and return the majority vote;

Insight: By ensembling predictors by averaging (or generally aggregating) many low bias, high variance predictors, we can reduce the variance while retaining low bias!



Insight: By ensembling predictors by averaging (or generally aggregating) many low bias, high variance predictors, we can reduce the variance while retaining low bias!



Bagging is typically not enough -- in order to get a good reduction in variance, we require that models being aggregated be uncorrelated, so that they make "different errors"

Bagging usually gets you highly correlated models that will make the same errors, and will therefore not reduce variance of combined predictor

QUESTIONS?