COMP30120

Ensembles

Derek Greene

School of Computer Science and Informatics

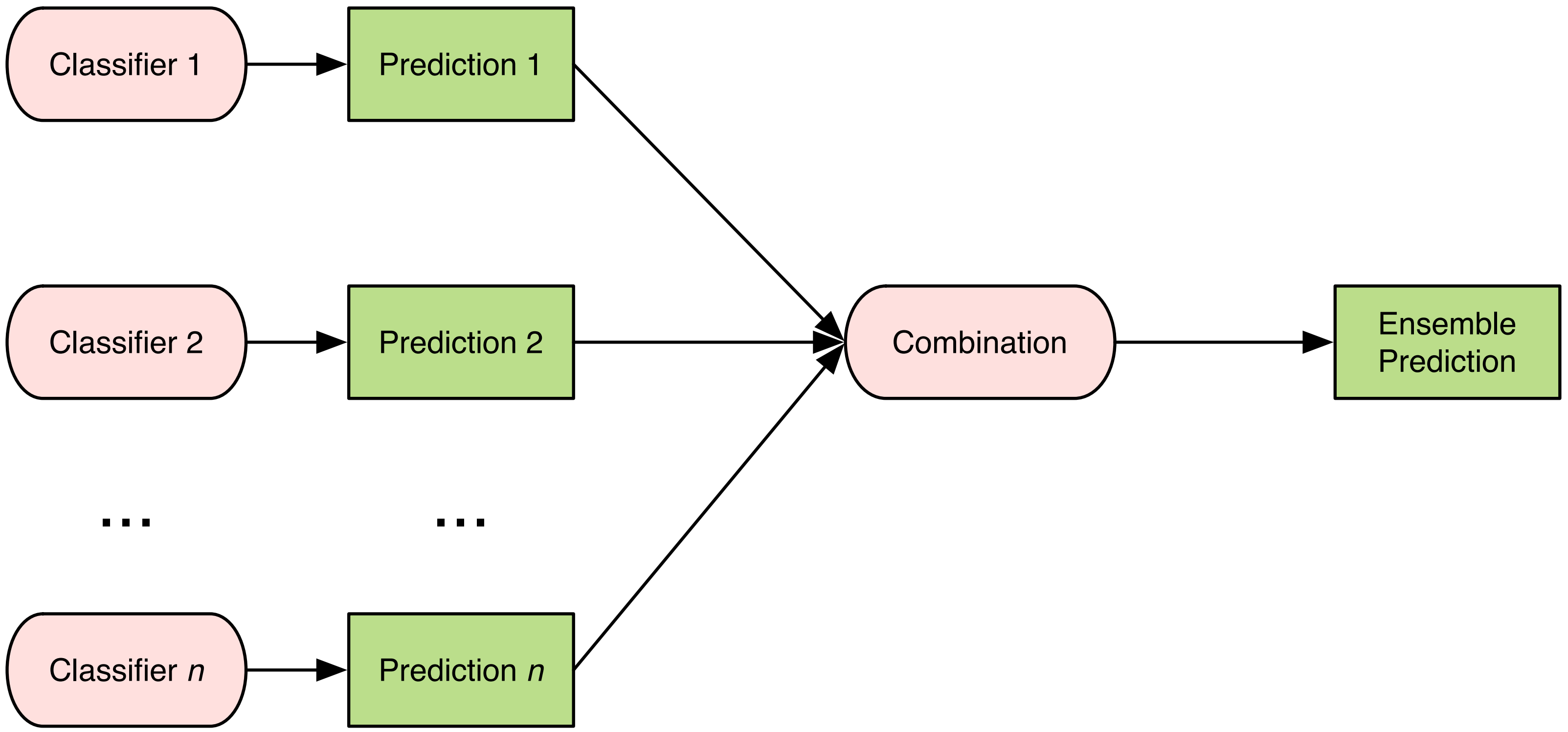
Autumn 2015

Overview

- Ensemble Classification
- Why do ensembles work?
  - Condorcet Jury Theorem
- Ensemble Generation
  - Bagging v Boosting
- Ensemble Combination
  - Voting v Weighted Voting
- Bias/Variance decomposition of error

Ensemble Idea

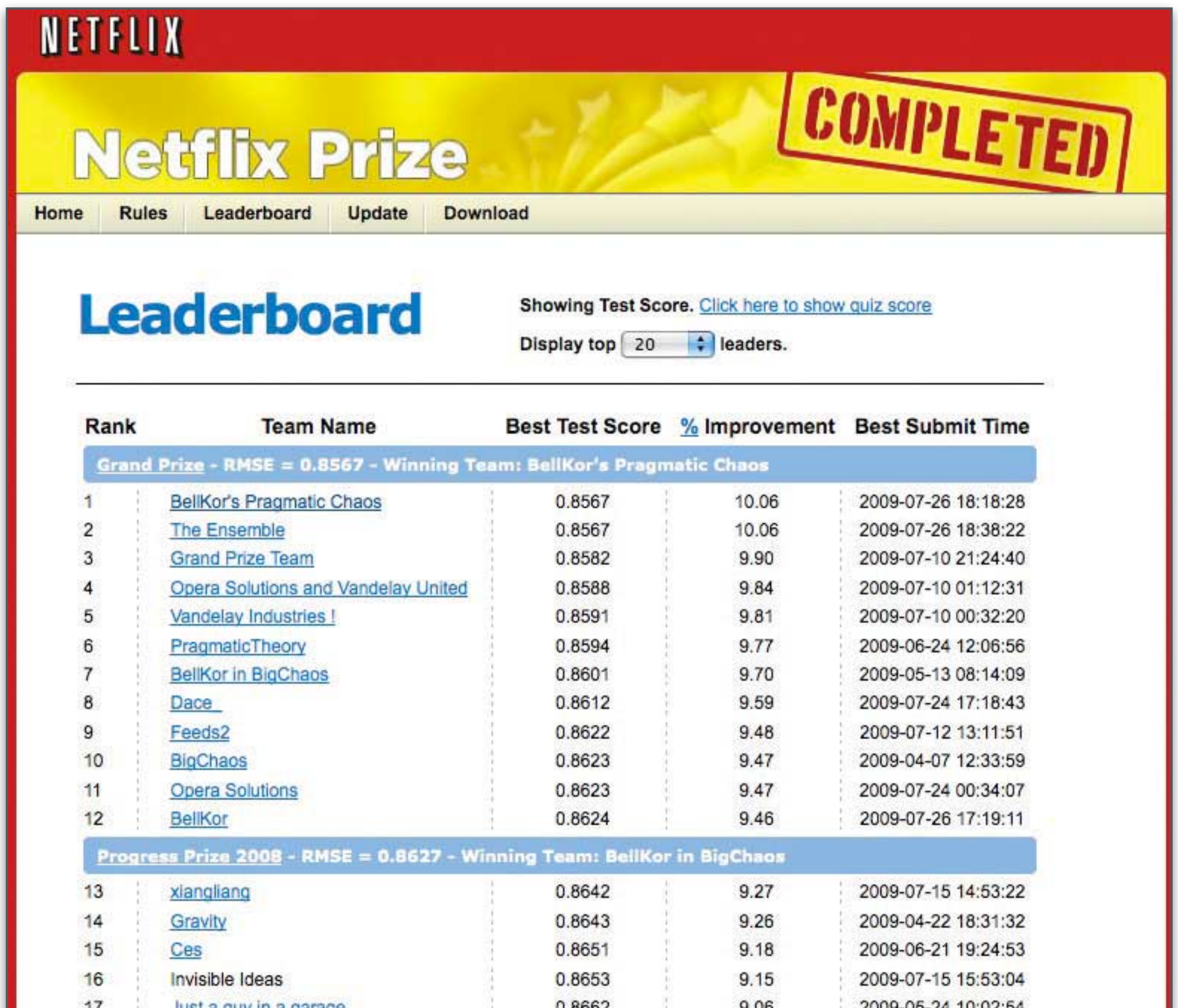
- Ensemble Classification: Aggregation of predictions made by multiple classifiers with the goal of improving accuracy.



- An ensemble of "weak learners" can provide a strong committee.
- Applied using many different types of classifiers - decision trees, *k*-NNs, neural networks, support vector machines…

Application: Netflix Prize

In 2006, Netflix announced a machine learning competition for movie rating prediction. Prize of $1 million to whoever improved the accuracy of existing system by 10%.



Rank	Team Name	Best Test Score	% Improvement	Best Submit Time
Grand Prize - RMSE = 0.8567 - Winning Team: BellKor's Pragmatic Chaos				
1	BellKor's Pragmatic Chaos	0.8567	10.06	2009-07-26 18:18:28
2	The Ensemble	0.8567	10.06	2009-07-26 18:38:22
3	Grand Prize Team	0.8582	9.90	2009-07-10 21:24:40
4	Opera Solutions and Vandelay United	0.8588	9.84	2009-07-10 01:12:31
5	Vandelay Industries !	0.8591	9.81	2009-07-10 00:32:20
6	PragmaticTheory	0.8594	9.77	2009-06-24 12:06:56
7	BellKor in BigChaos	0.8601	9.70	2009-05-13 08:14:09
8	Dace_	0.8612	9.59	2009-07-24 17:18:43
9	Feeds2	0.8622	9.48	2009-07-12 13:11:51
10	BigChaos	0.8623	9.47	2009-04-07 12:33:59
11	Opera Solutions	0.8623	9.47	2009-07-24 00:34:07
12	BellKor	0.8624	9.46	2009-07-26 17:19:11
Progress Prize 2008 - RMSE = 0.8627 - Winning Team: BellKor in BigChaos				
13	xiangliang	0.8642	9.27	2009-07-15 14:53:22
14	Gravity	0.8643	9.26	2009-04-22 18:31:32
15	Ces	0.8651	9.18	2009-06-21 19:24:53
16	Invisible Ideas	0.8653	9.15	2009-07-15 15:53:04
17	Just a guy in a garage	0.8662	9.06	2009-05-24 10:02:54

Top submissions all combine several teams and algorithms as an ensemble.

BellKor Team:
"Our final solution consists of blending 107 individual results"

http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html

Ensembles: Motivation

The Condorcet Jury Theorem

- Proposed by the Marquis of Condorcet in 1784, and relates to the relative probability of an ensemble of individuals arriving at a correct decision.
- If each voter has a probability p of being correct and the probability of a majority of voters being correct is M…
  - Then $p > 0.5$ implies $M > p$
  - Also if p always > 0.5, then M approaches 1.0 as the number of voters approaches infinity.

➡ "When the average probability of an individual being correct is > 50%, the chance of the ensemble of them reaching the correct decision increases as more members are added".

Ensembles: Motivation

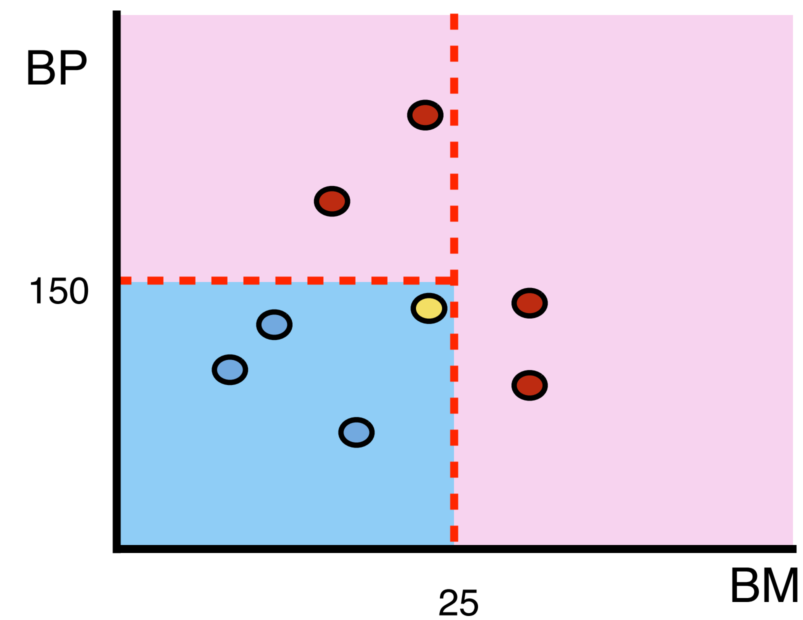
- Condorcet Jury Theorem revisited…
  - We now know that M will be greater than p only if there is diversity in the pool of voters - i.e. there is some disagreement between their decisions.
  - The probability of a majority of voters being correct will increase as the ensemble grows only if the diversity in the ensemble continues to grow as well.
- Eventually, new ensemble members will have voting patterns collinear with existing members.
- Typically the diversity of the ensemble will plateau as will the accuracy of the ensemble at some size between 10-50 members.

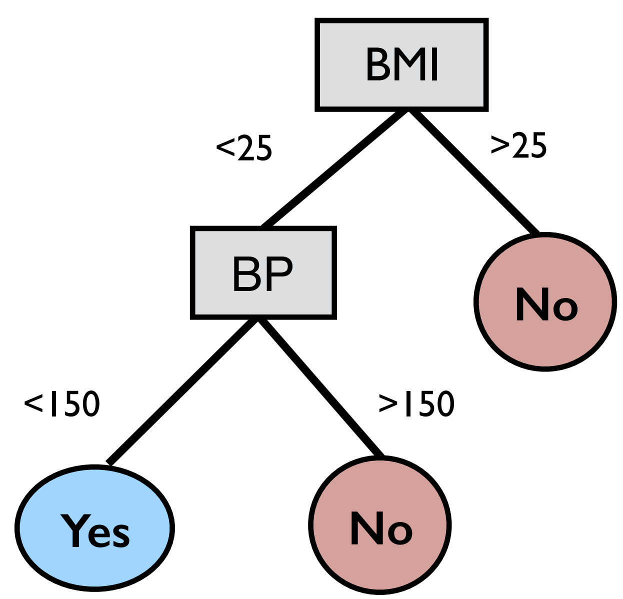
Example: Classification

- **Data:** Heart attack patient admitted. 19 variables measured during first 24 hours. Blood pressure, age, BMI + 16 other variables, considered important indicators of patient's condition.
- **Task:** Identify high risk patients (i.e. will not survive 30 days), based on evidence of initial 24-hour data.

No	Age	BMI	BP	Ok?
1	60	20	140	Yes
2	60	21	145	Yes
3	85	23	130	Yes
4	81	22	160	No
5	70	24	170	No
6	72	26	135	No
7	81	26	145	No
8	66	23	155	No

Q	66	24	148	?

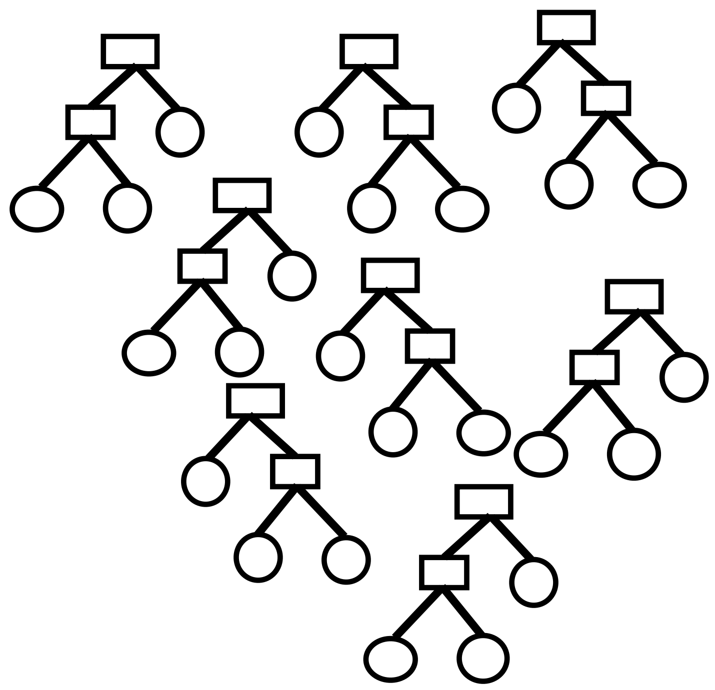




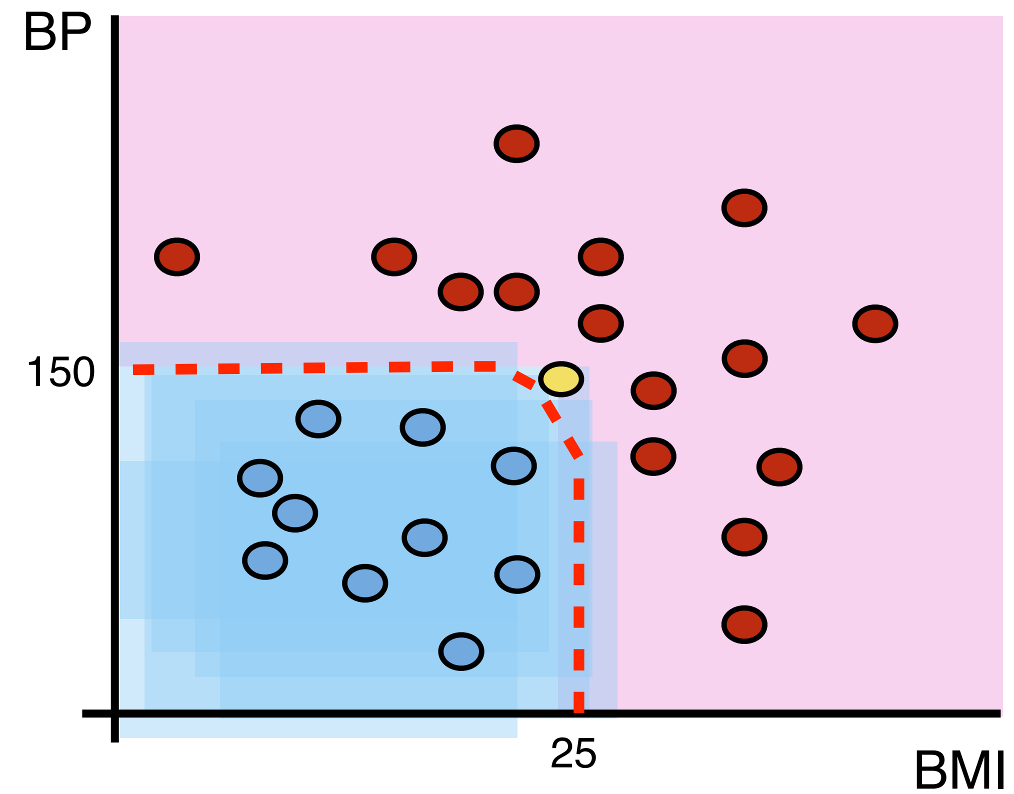
Ensemble Idea

- Build many "base" decision trees, using different subsets of the data. These trees can vote on the class of a new input example.

➡ Accuracy of ensemble should be better than the individual trees.



Ensemble of Decision Trees



Q. How do we <u>generate</u> base classifiers that complement each other?

Q. How do we <u>combine</u> the outputs of base classifiers to maximise accuracy?

Ensemble Generation: Bagging

- **Key Idea:** Train *n* classifiers on different subsets of the training data.
- Bootstrap aggregation / Bagging (Breiman, 1996):
  - Randomly sample from training data with replacement.
  - 100% bootstrap sample will contain ~63% of training examples. Remaining data is "out-of-bag" (OOB).

Complete dataset has 8 examples

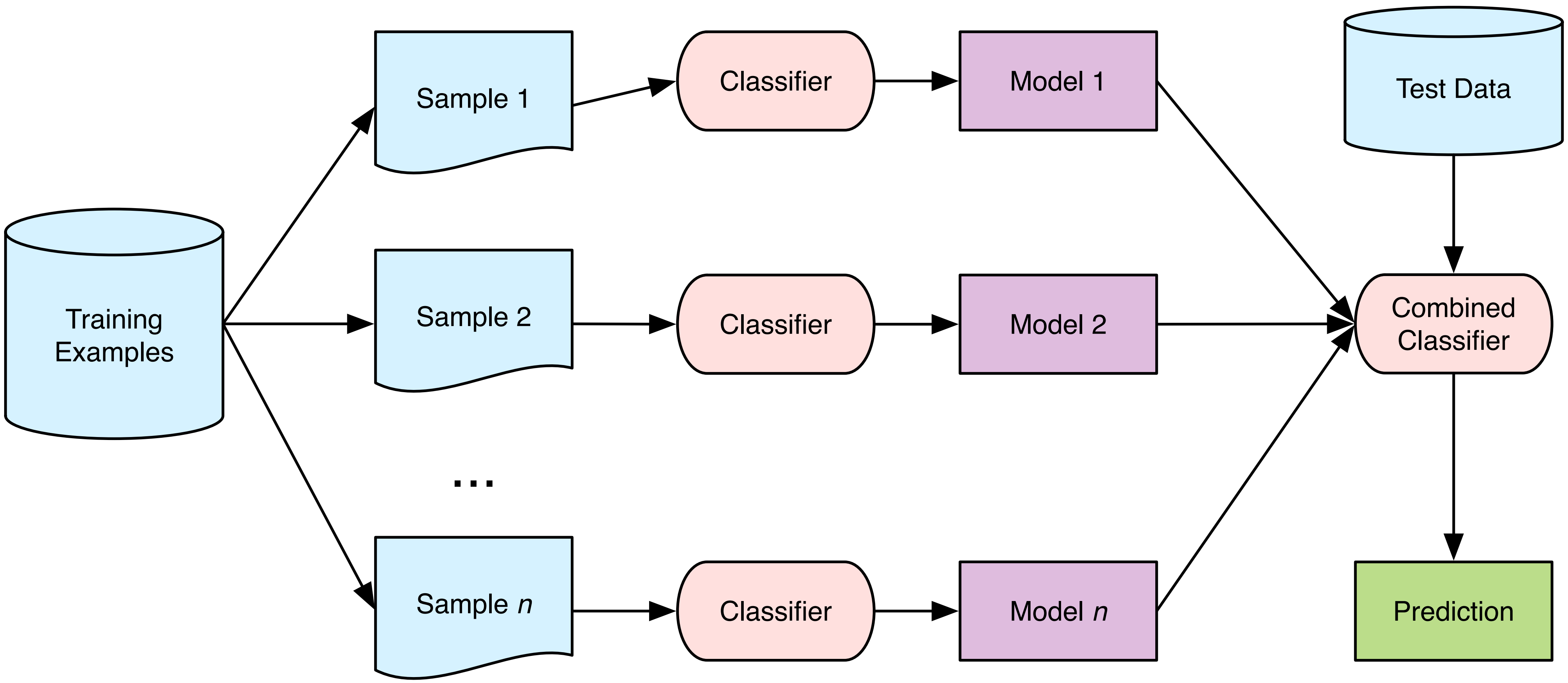
Original	A	B	C	D	E	F	G	H
Set 1	B	G	H	C	G	F	C	A
Set 2	G	H	E	F	D	B	G	A
Set 3	C	F	B	G	E	F	B	B
Set 4	D	E	A	D	E	D	C	H
Set 5	E	F	A	C	E	F	A	H
Set 6	C	H	B	F	D	B	H	F

Each bootstrap subset has 8 examples.

Some examples may be duplicated, others left out.

Ensemble Generation: Bagging

- Bootstrap aggregation: Randomly sample from training data with replacement, apply a classifier to each sample.



➡ Encourages diversity in the ensemble, works better for "unstable" classifiers - e.g. decision trees, neural networks.

Ensemble Generation: Random Features

- **Key Idea:** Train *n* base classifiers, each on a different subset of features.
- Random Subspace Method:
  - A subset of features is randomly selected without replacement.
  - Train a classifier using only selected features to represent the training data.
  - ➡ Encourages diversity in the ensemble, works well for *k*-NNs.

Original dataset
(5 features)

X	Col	H	W	Wgt	HW
1	210	60	62	186	0.97
2	220	70	53	180	1.32
3	215	55	50	152	1.10
4	180	76	40	152	1.90
...	...	...	...	...	...
10	190	60	58	175	1.03

Feature Subset 1
(3 features)

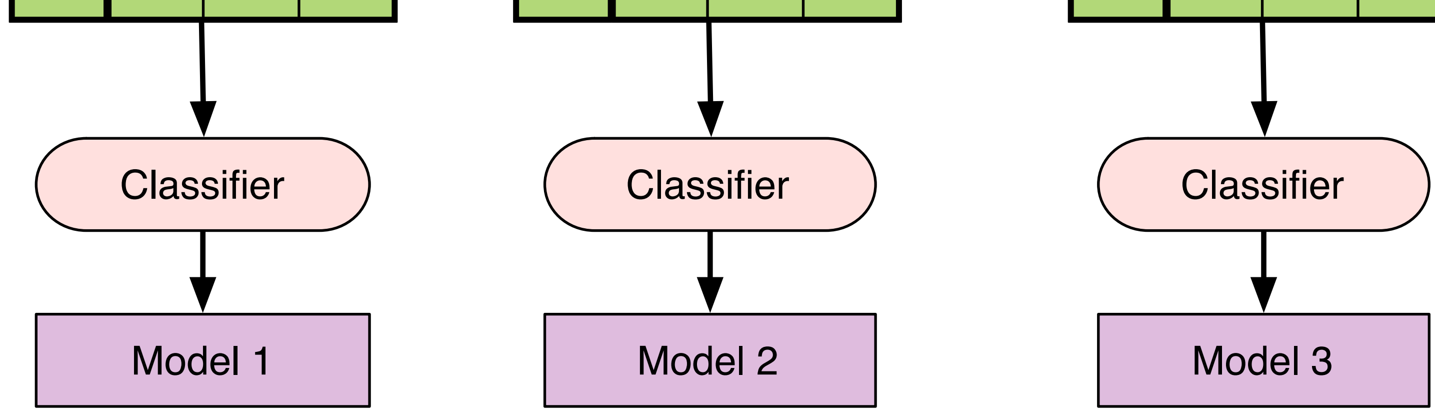
X	H	W	HW
1	60	62	0.97
2	70	53	1.32
3	55	50	1.10
4	76	40	1.90
...	...	...	...
10	60	58	1.03

Feature Subset 2
(3 features)

X	Col	H	HW
1	210	60	3.50
2	220	70	3.14
3	215	55	3.91
4	180	76	2.37
...	...	...	...
10	190	60	3.17

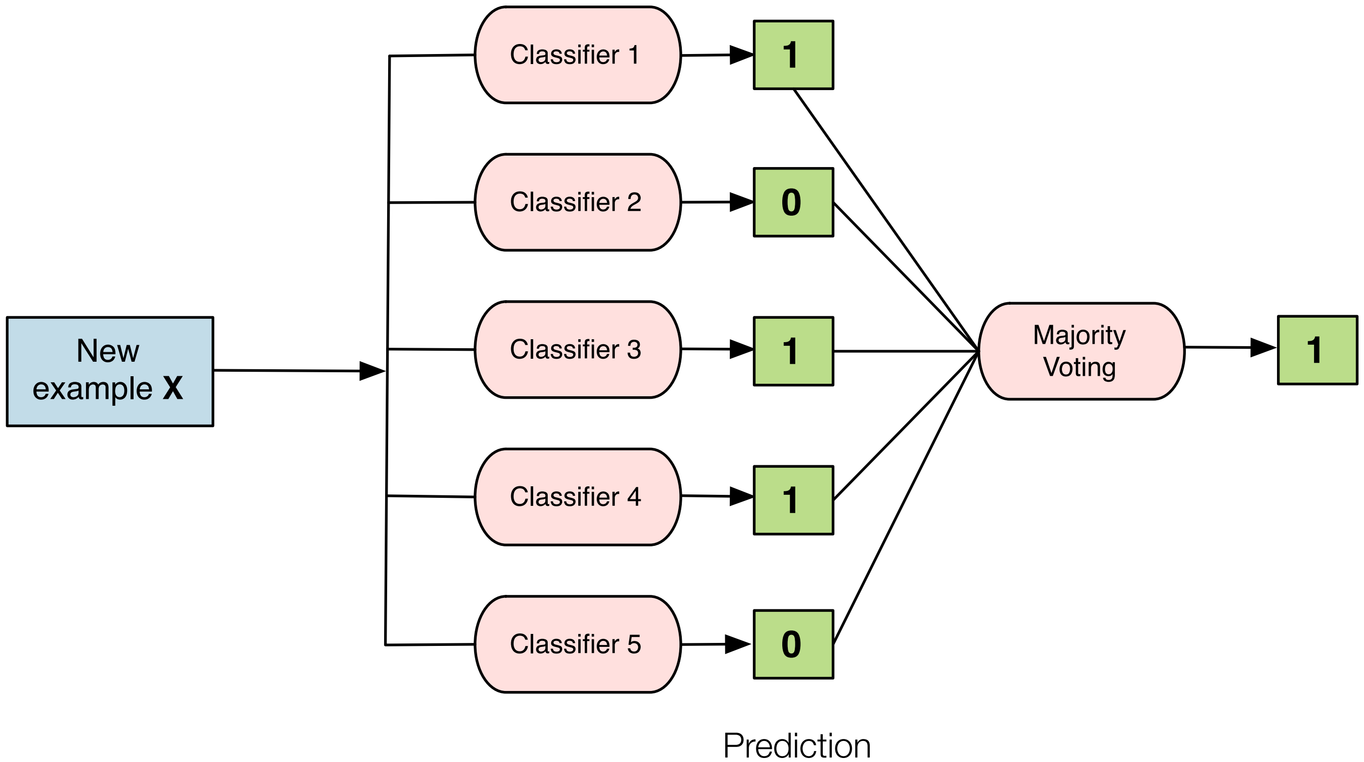
Feature Subset 3
(3 features)

X	Col	H	W
1	210	60	62
2	220	70	53
3	215	55	50
4	180	76	40
...	...	...	...
10	190	60	58



Ensemble Combination: Voting

- Simplest way to combine the output of multiple classifiers is to use majority voting.
- All classifiers are run independently in parallel. Results are combined when all runs have completed.
- Each classifier "votes" for a particular class, where all classifiers carry equal weight. The class with the majority vote in the ensemble wins.



Ensemble votes for label "1" for **X** by 3:2

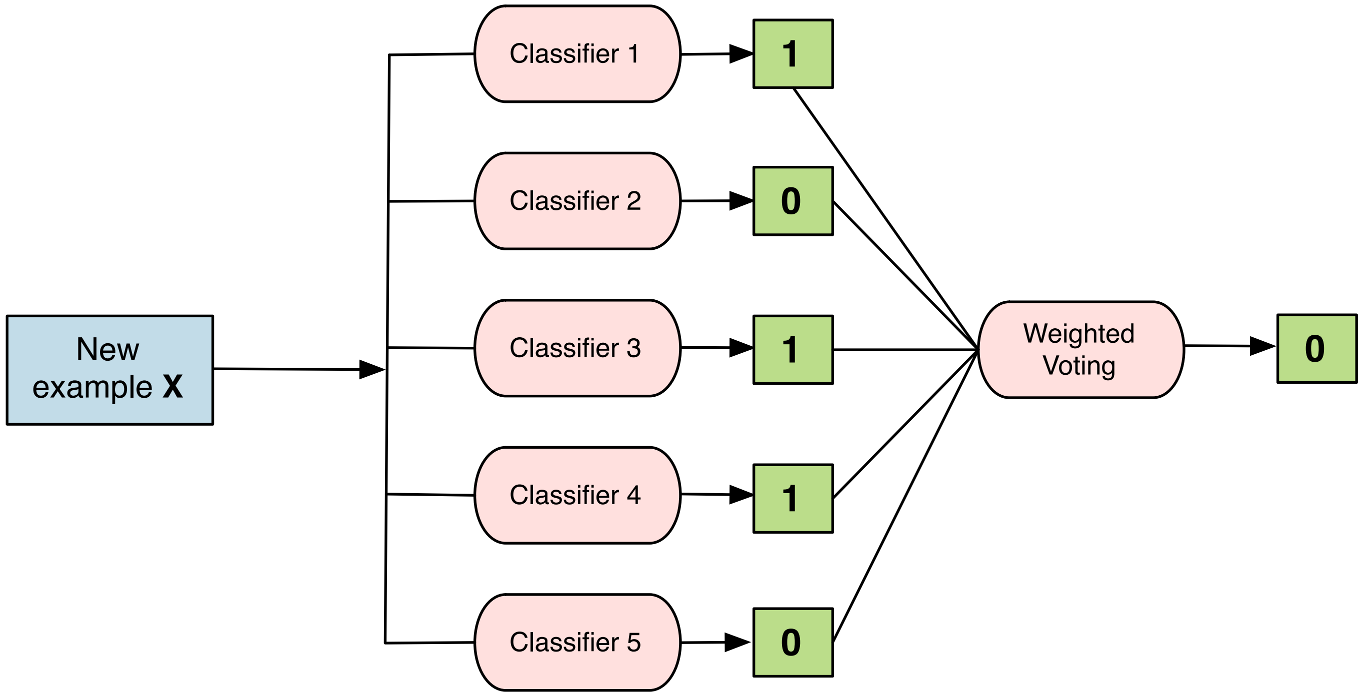
Ensemble Combination: Weighted Voting

- **Intuition:** If individual classifiers do not give equal performance, we should give more power to better classifiers.
- Weighted Voting Combination:
  - Rather than treating every classifier's vote equally, we weight each classifier's vote based on its accuracy/error.

  ➡ More accurate classifiers contribute more to the ensemble.

Classifier	Accuracy	Weight
1	0.52	0.14
2	1.00	0.28
3	0.57	0.16
4	0.55	0.15
5	0.95	0.26
TOTAL	**3.59**	**1.00**

e.g. C1: 0.52/3.59 = 0.14



Vote "0": 0.28 + 0.26 ≅ 0.54

Vote "1": 0.14 + 0.16 + 0.15 ≅ 0.46

Committees of Experts

- **Consider:** *" … a medical school that has the objective that all students, given a problem, come up with an identical solution"*.
- No value in a committee of experts from such a group - the committee will not improve on the judgement of an individual.
- There needs to be disagreement for the committee to have the potential to be better than an individual.

- Fundamental work by Krogh & Vedelsby (1995) for regression:
  - Increasing "ambiguity" (disagreement) decreases overall combined error, provided it does not result in an increase of average error.

$$\overline{E} - \overline{A} = E$$

Average error of classifiers

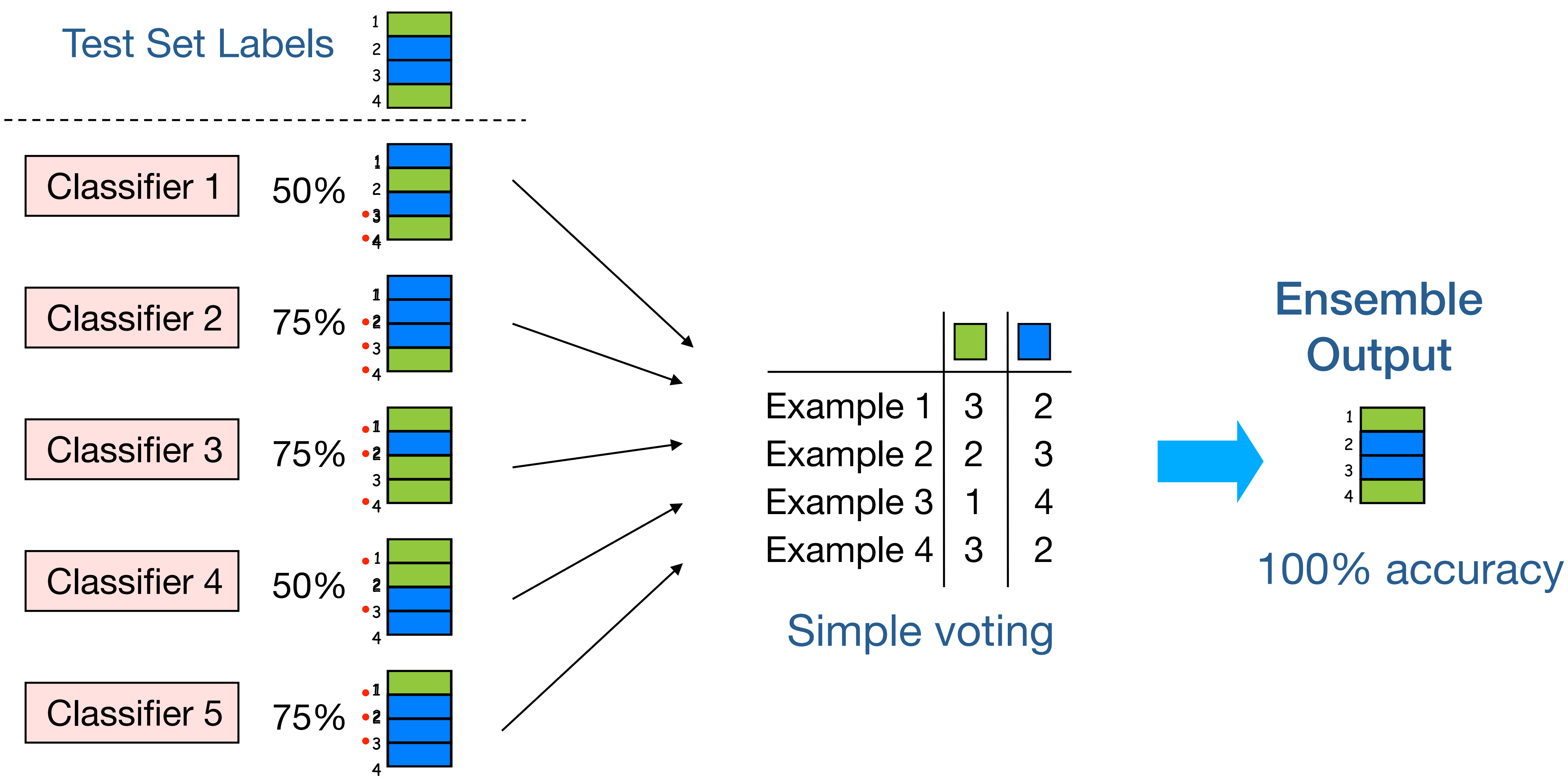
Ensemble ambiguity

Ensemble error

We need accuracy + diversity in classifier ensembles!

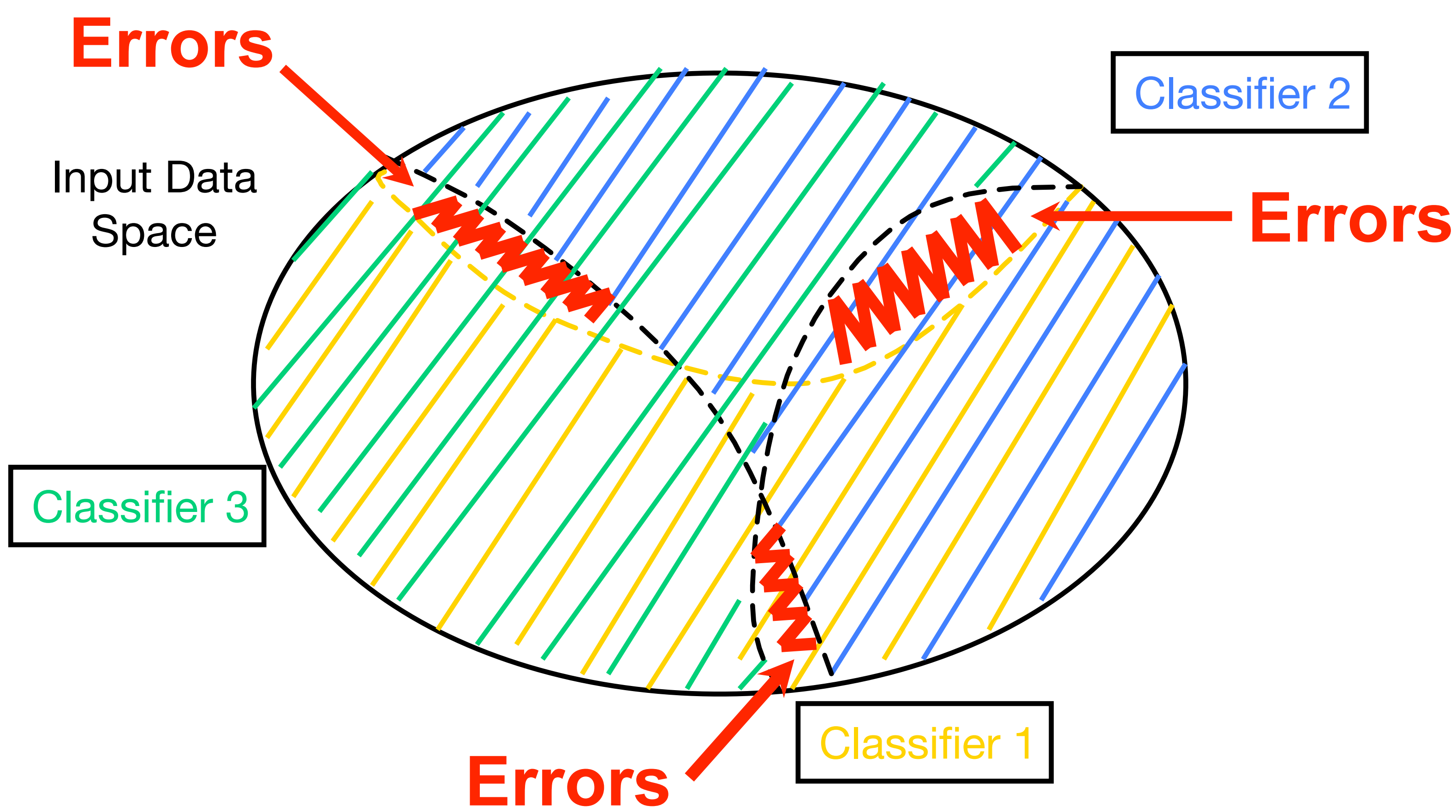
Ensemble Diversity

- **Local Learning in Ensembles**
  - Every single classifier performs well on a subset of the test set.
  - The mistakes that one classifier makes are “corrected” by the other classifiers.



Ensemble Specialisation

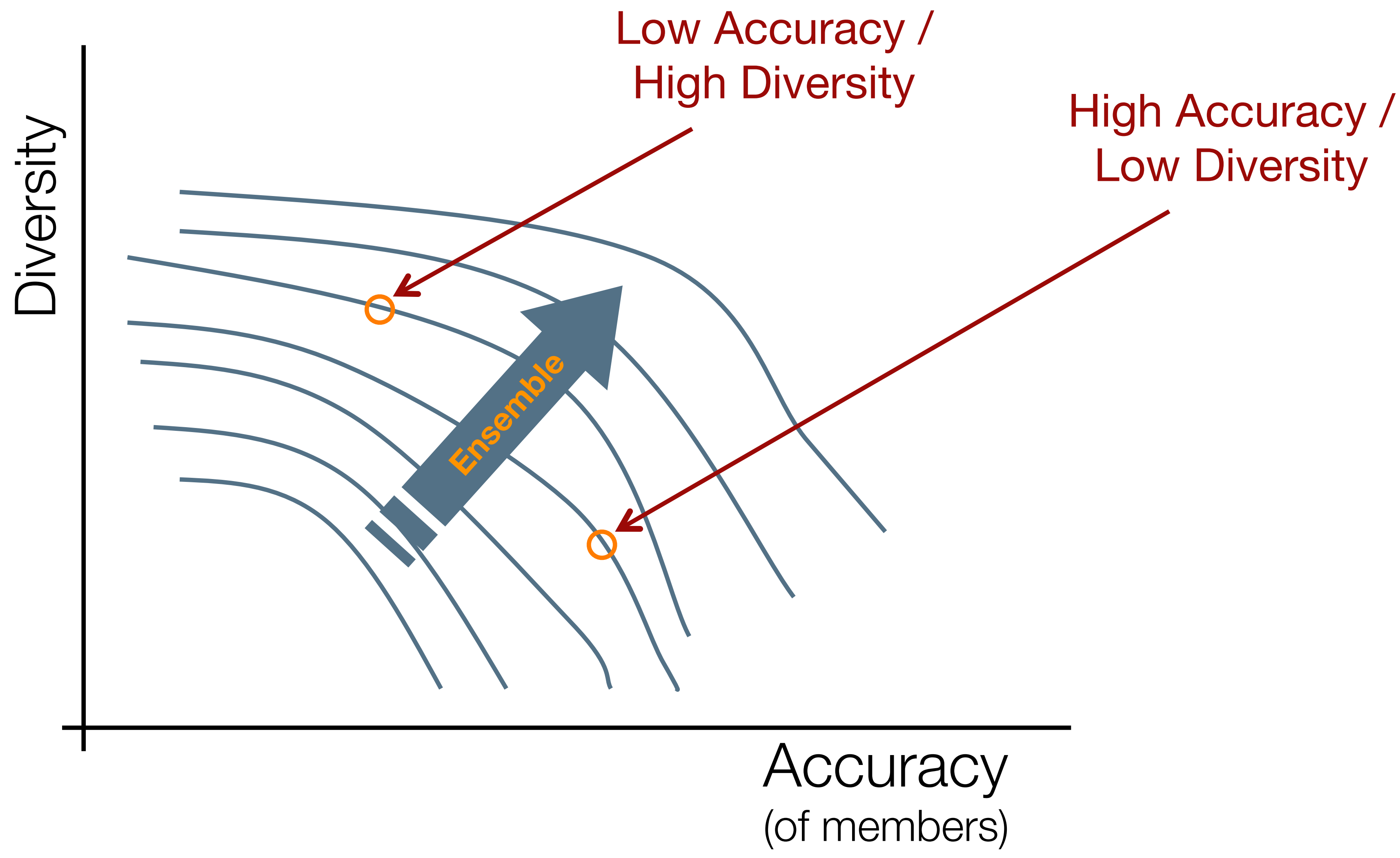
- Classifiers can specialise on accurately classifying only related examples from certain regions of the input data space.
- **Example:** Visualisation of specialisation in classifier ensembles…



➡ Want ensemble members to make mistakes in different areas.

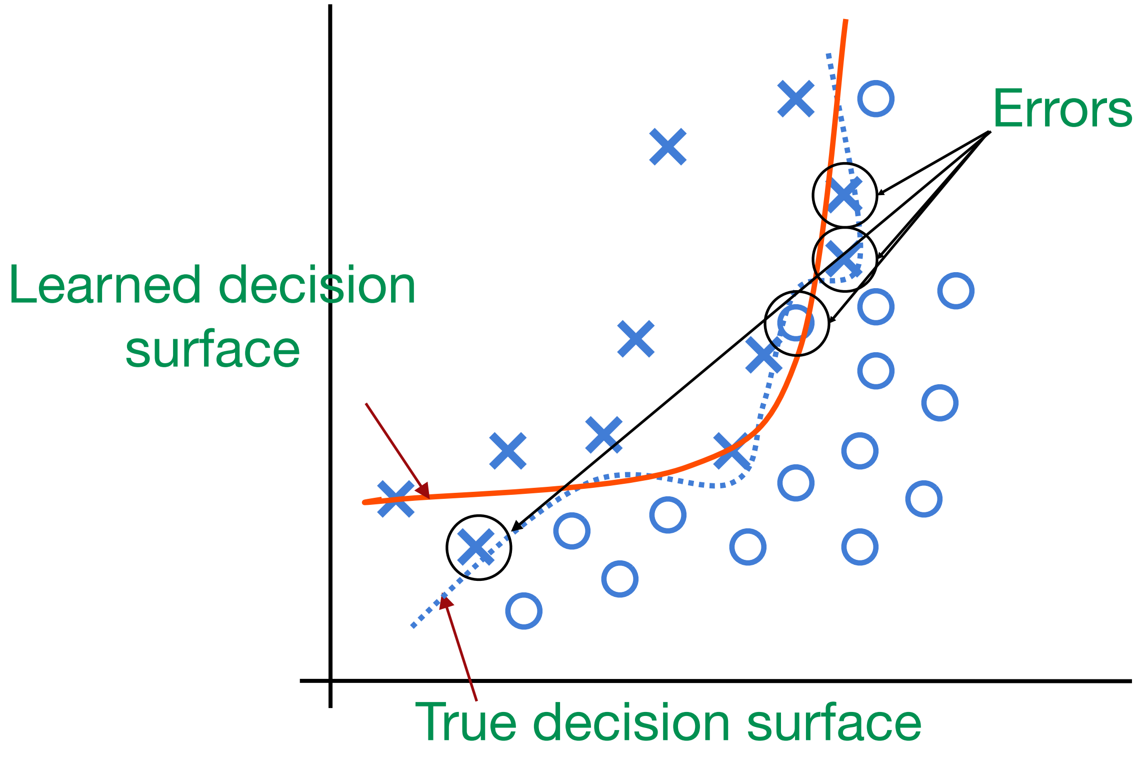
Ensemble Diversity

- **Recall:** Krogh & Vedelsby said an ideal ensemble is one that consists of highly accurate members which at the same time disagree.
- Often face a trade-off between diversity and accuracy when constructing an ensemble of classifiers.



Ensemble Generation: Boosting

- **Key idea:** Train a series of classifiers such that later classifiers are trained to better predict on examples that earlier ones perform poorly on.
- Focus on previous errors when building next ensemble member.



Boosting Approach:

1. Weight all training examples equally.
2. FOR *i = 1 to T*
   (a) Train classifier using current weights.
   (b) Compute errors.
   (c) Increase weights for misclassified examples, decrease weights for those classified correctly.
3. Output final model.

Example: Boosting

- **Problem:** Apply a classifier to a training set with 8 examples *{A, B, C, D, E, F, G, H}*, where example *A* is an outlier and difficult to classify.

- Selected training sets for 4 runs of bagging - i.e. simple random sampling with replacement.

➡ All examples equally weighted.

Original	A	B	C	D	E	F	G	H
Set 1	B	G	H	C	G	F	C	A
Set 2	G	H	E	F	D	B	G	A
Set 3	C	F	B	G	E	F	B	B
Set 4	D	E	A	D	E	D	C	H

- Selected training sets for 4 runs of boosting - i.e. increase weights for misclassified examples.

➡ The "hard" example *A* appears more frequently in later sets.

Original	A	B	C	D	E	F	G	H
Set 1	B	G	H	C	G	F	C	A
Set 2	A	D	E	D	A	E	F	D
Set 3	G	A	E	H	A	H	A	D
Set 4	A	A	F	A	A	C	A	E

D. Opitz & R. Maclin. "Popular Ensemble Methods: An Empirical Study" (1999)

Bias and Variance

- We can view the error of a classifier predicting a given target function on a dataset as consisting of three parts:
  1. Bias: Measures how close the average classifier's predictions are from the correct target function.
  2. Variance: Measures the error from sensitivity to small fluctuations in the training set.
  3. Minimum classification error (i.e. the noise in the data).

- Theories relating to ensemble generation methods:
  - Bagging can often reduce variance part of error.
  - Boosting can often reduce variance AND bias, since it focuses on misclassified examples.
  - Boosting may sometimes increase error, as it is susceptible to noise and may lead to overfitting.

Summary

- Ensemble Classification
- Why do ensembles work?
  - Condorcet Jury Theorem
- Ensemble Generation
  - Bagging v Boosting
- Ensemble Combination
  - Voting v Weighted Voting
- Bias/Variance decomposition of error

References

- D. Opitz and R. Maclin. "Popular Ensemble Methods: An Empirical Study" (1999). Journal of Artificial Intelligence Research.
- Breiman, L., (1996) “Bagging predictors”. Machine Learning, 24:123-140.
- Krogh, A., Vedelsby, J., (1995) “Neural Network Ensembles, Cross Validation and Active Learning”, in Advances in Neural Information Processing Systems 7
- Baur, E., and Kohavi, R. (1999) “An Empirical Comparison of Voting Classification Algorithms: Bagging, Boosting, and Variants”, Machine Learning.