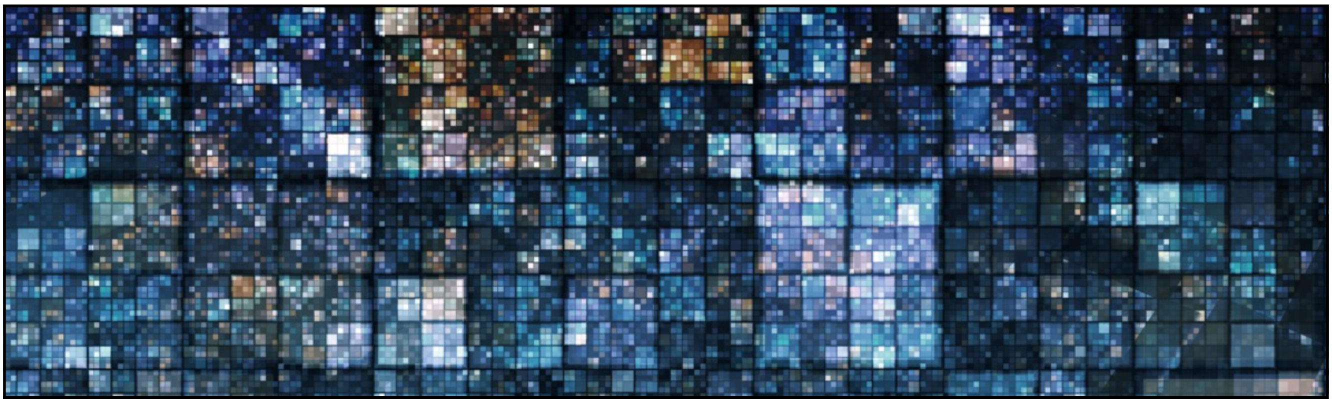


COMP47590

Advanced Machine Learning

Supervised Learning - Ensembles 1

Dr. Brian Mac Namee

1

Contents

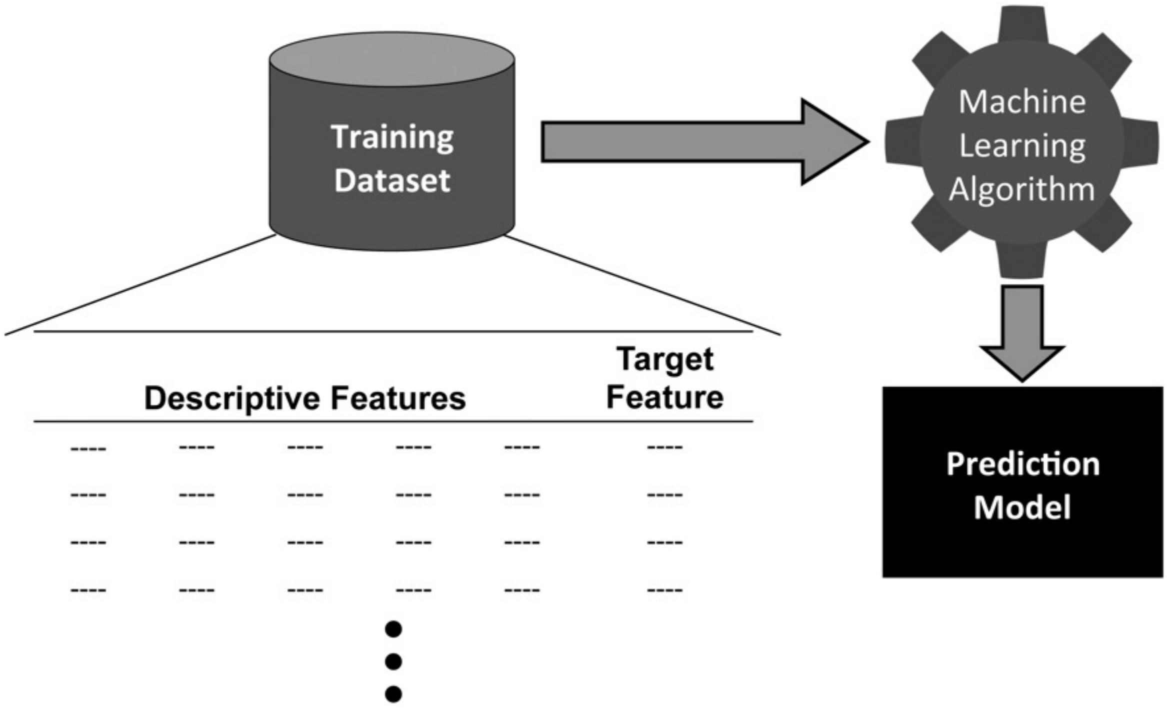
Today we will cover

- Supervised learning
- Wisdom of the crowds
- Ensembles
- Random forests
- Gradient boosting

2

SUPERVISED LEARNING

3



Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

4

$$\mathcal{D} = [(\mathbf{d}_1, t_1), (\mathbf{d}_2, t_2), \ldots , (\mathbf{d}_n, t_n)]$$

where $\mathbf{d}_i$ is a set of descriptive features

$\mathbf{d}_i[0], \mathbf{d}_i[1], \ldots , \mathbf{d}_i[m]$

t_i is the corresponding target feature value

5

6

$$\mathrm{t} = \mathbb{M}(\mathbf{q})$$

where $\mathbf{q}$ is a set of descriptive features $\mathbf{q}[0], \mathbf{q}[1], \dots, \mathbf{q}[\mathrm{m}]$ describing a query instance

t is a predicted target feature value

Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

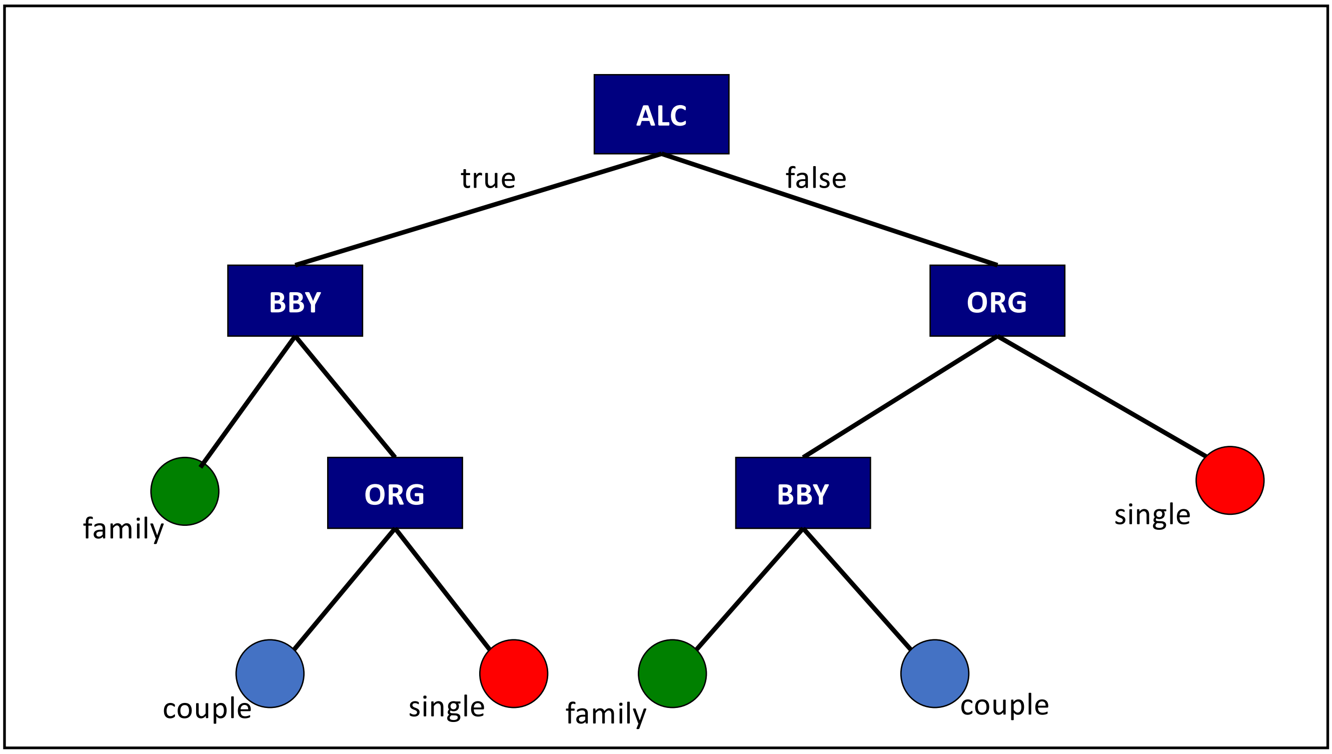
7

A simple retail dataset

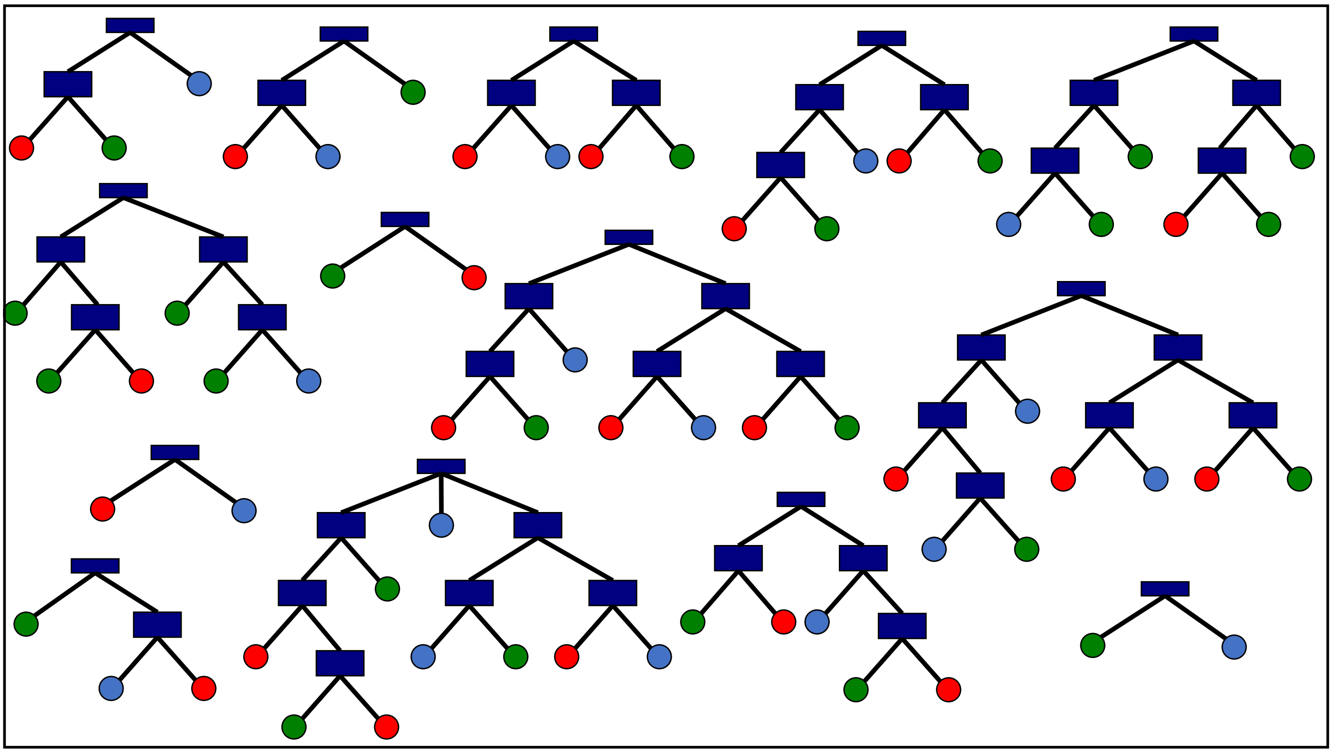
ID	BBY	ALC	ORG	GRP
1	no	no	no	couple
2	yes	no	yes	family
3	yes	yes	no	family
4	no	no	yes	couple
5	no	yes	yes	single

Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

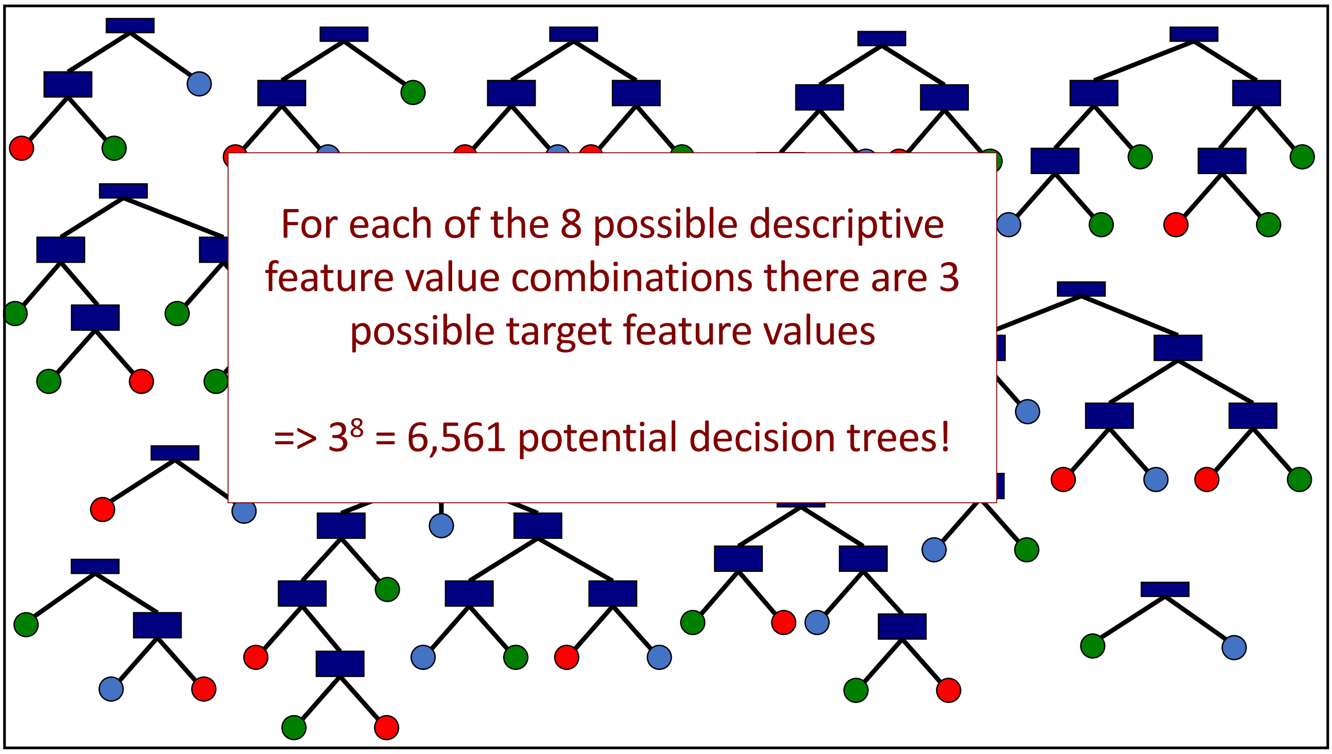
8


ALC
true
false
BBY
ORG
family
ORG
BBY
single
couple
single
family
couple

9



10



11

Consistency?

Consistency ≈ memorizing the dataset

Consistency with noise in the data isn't desirable

Coverage through memorization is never possible in real problems

12

Consistency?

Consistency ≈ memorizing the dataset

Consistency with noise in the data isn’t desirable

Coverage through memorization is never possible in real problems

GOAL: a model that **generalises** beyond the dataset and that **invariant** to the noise in the dataset

13

Inductive Bias

The solution is **inductive bias**, a set of assumptions that define the model selection criteria of an ML algorithm

There are two types of bias that we can use:

- restriction bias
- preference bias

Inductive bias is necessary for generalisation

14

WISDOM OF THE CROWD

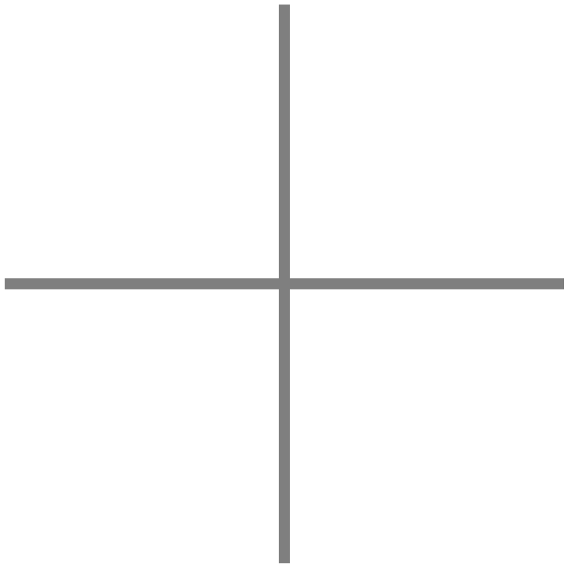
15

Experiment

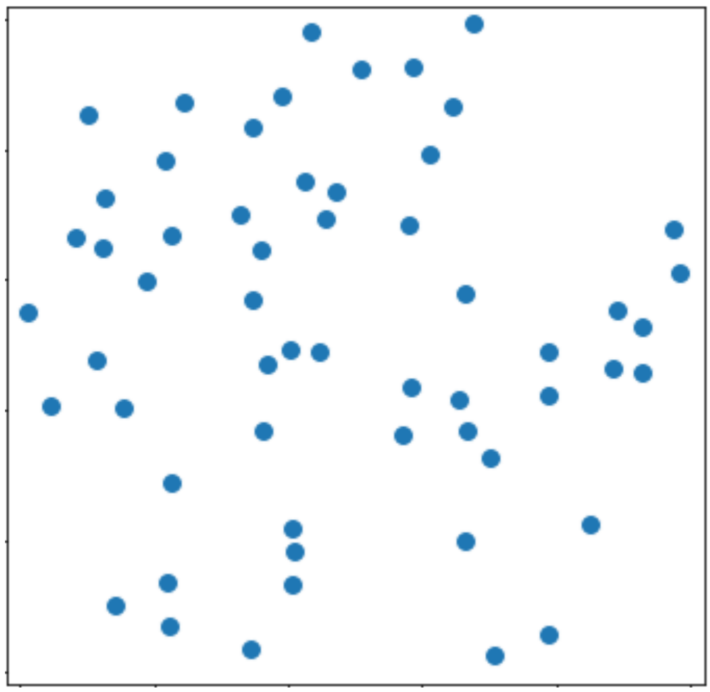
Estimate the number of dots on the graph on that appears and enter your estimate online.

NOTE you will only see the graph for a short amount of time.

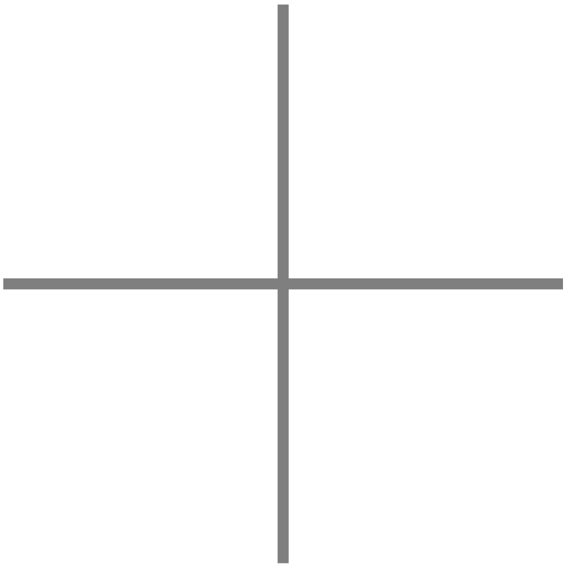
16



17



18



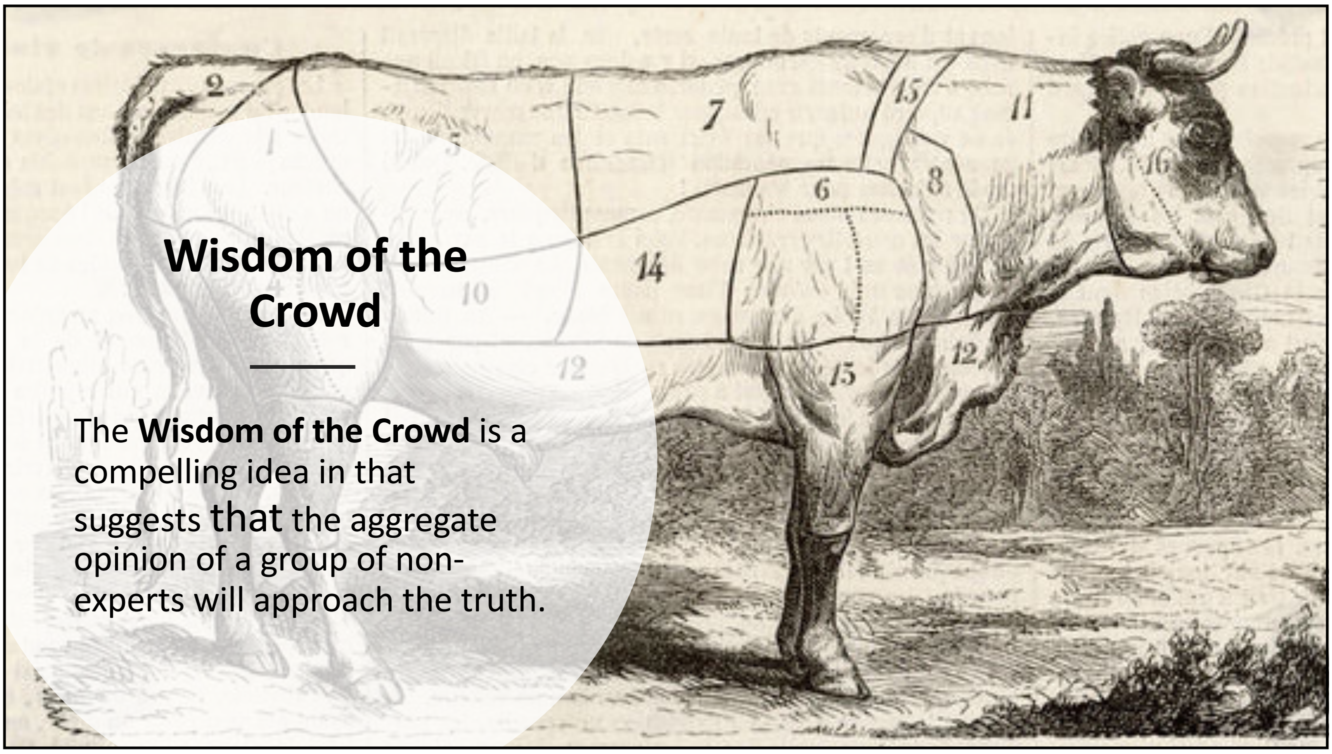
19

Experiment

Enter your estimate here:

http://bit.ly/40Ba9rs

20



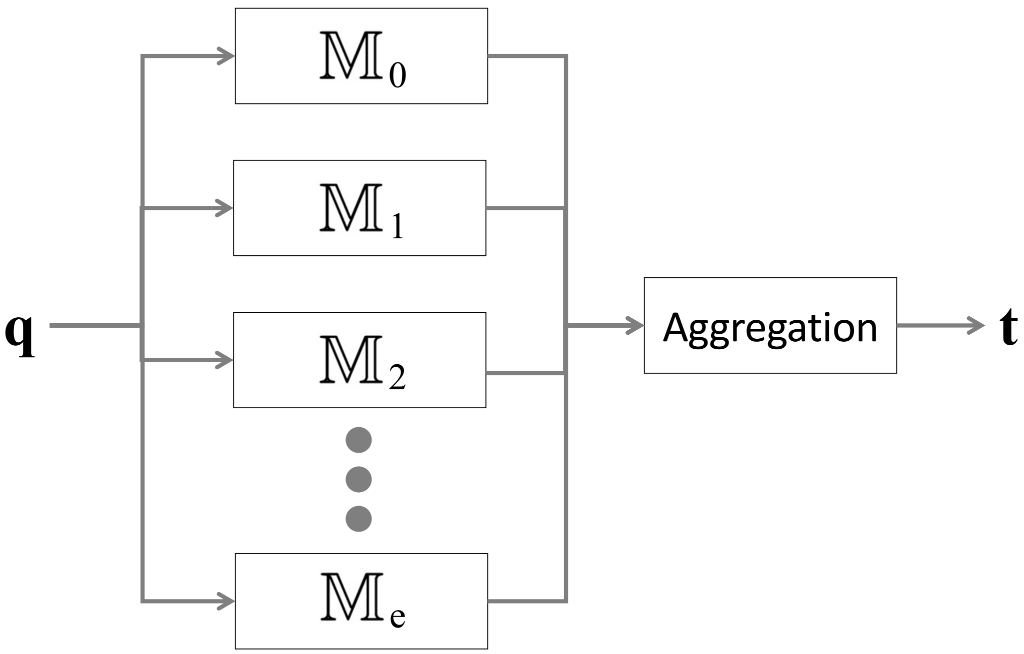
Wisdom of the Crowd

The **Wisdom of the Crowd** is a compelling idea in that suggests that the aggregate opinion of a group of non-experts will approach the truth.

21

ENSEMBLES

22



23

Ensembles

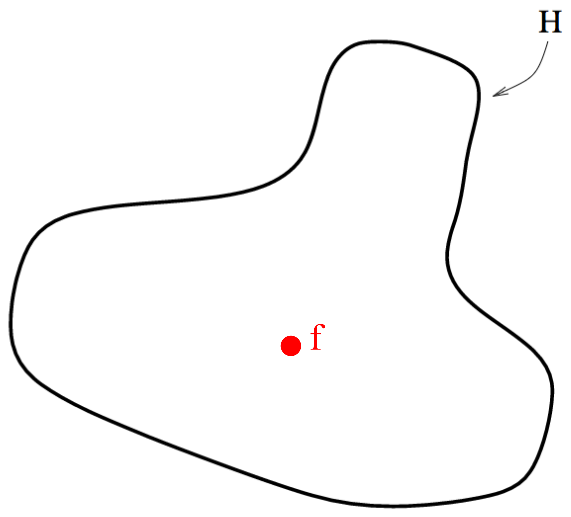
The aggregate of multiple combined models is more effective than any individual model

Thomas Dietterich describes 3 motivations for using ensembles:

- Statistical
- Computational
- Representational

Ensemble Methods in Machine Learning
Thomas G Dietterich
http://web.engr.oregonstate.edu/~tgd/publications/mcs-ensembles.pdf

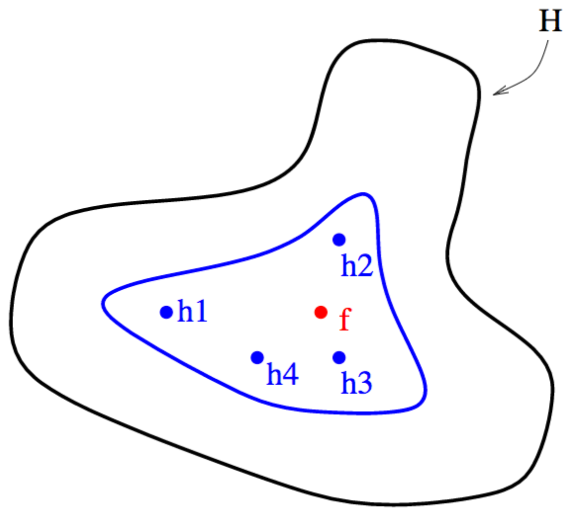
24



Ensemble Methods in Machine Learning
Thomas GDietterich
http://web.engr.oregonstate.edu/~tgd/publications/mcs-ensembles.pdf

25

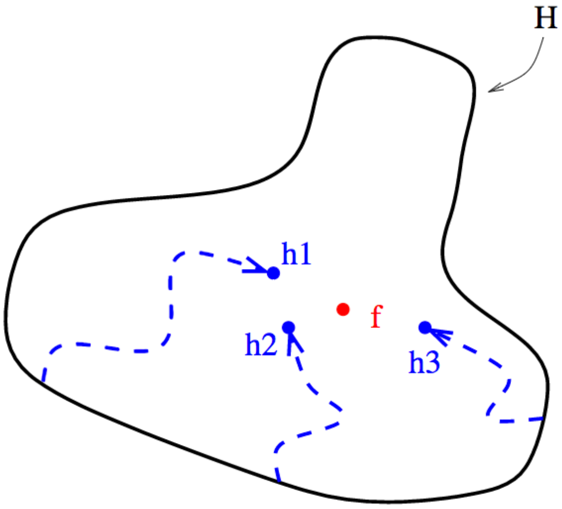
Statistical



Ensemble Methods in Machine Learning
Thomas GDietterich
http://web.engr.oregonstate.edu/~tgd/publications/mcs-ensembles.pdf

26

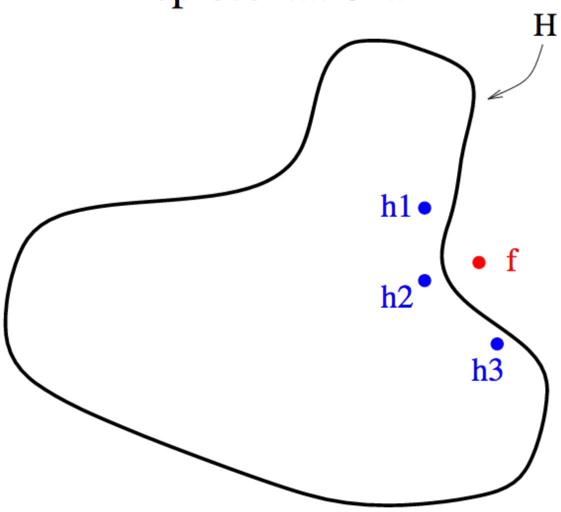
Computational



Ensemble Methods in Machine Learning
Thomas GDietterich
http://web.engr.oregonstate.edu/~tgd/publications/mcs-ensembles.pdf

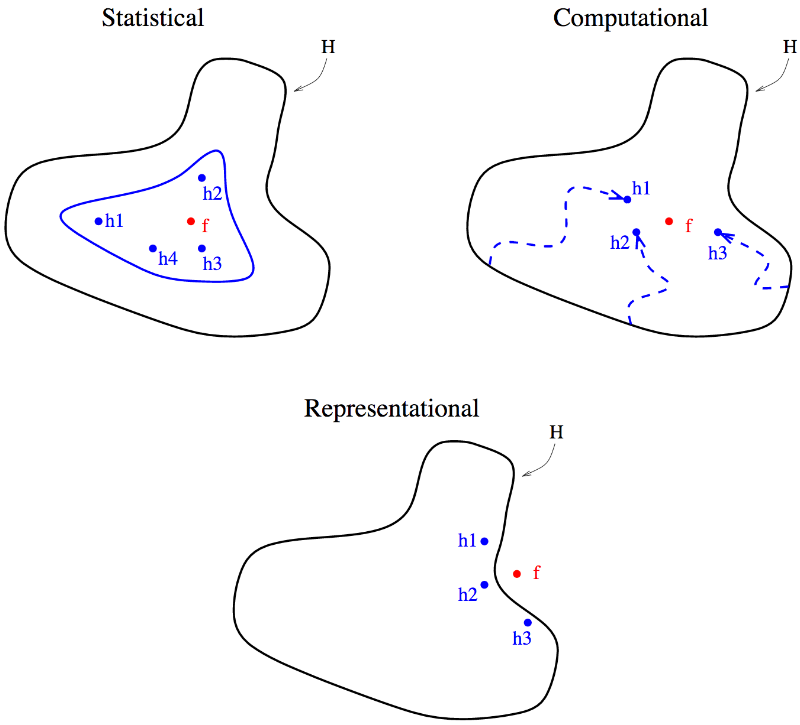
27

Representational



Ensemble Methods in Machine Learning
Thomas GDietterich
http://web.engr.oregonstate.edu/~tgd/publications/mcs-ensembles.pdf

28



Ensemble Methods in Machine Learning
Thomas GDietterich
http://web.engr.oregonstate.edu/~tgd/publications/mcs-ensembles.pdf

29

Ensembles

Imagine we have an ensemble for a binary prediction problem with 21 models, each with a classification error of 0.3

The big idea behind ensembles is that if we have multiple learners that are diverse, when one is wrong there is a very good chance that others are correct

30

Aggregation

$\mathbb{M}_0$	$\mathbb{M}_1$	$\mathbb{M}_2$	$\mathbb{M}_3$	$\mathbb{M}_4$	$\mathbb{M}_5$	$\mathbb{M}_6$
$\mathbb{M}_7$	$\mathbb{M}_8$	$\mathbb{M}_9$	$\mathbb{M}_{10}$	$\mathbb{M}_{11}$	$\mathbb{M}_{12}$	$\mathbb{M}_{13}$
$\mathbb{M}_{14}$	$\mathbb{M}_{15}$	$\mathbb{M}_{16}$	$\mathbb{M}_{17}$	$\mathbb{M}_{18}$	$\mathbb{M}_{19}$	$\mathbb{M}_{20}$

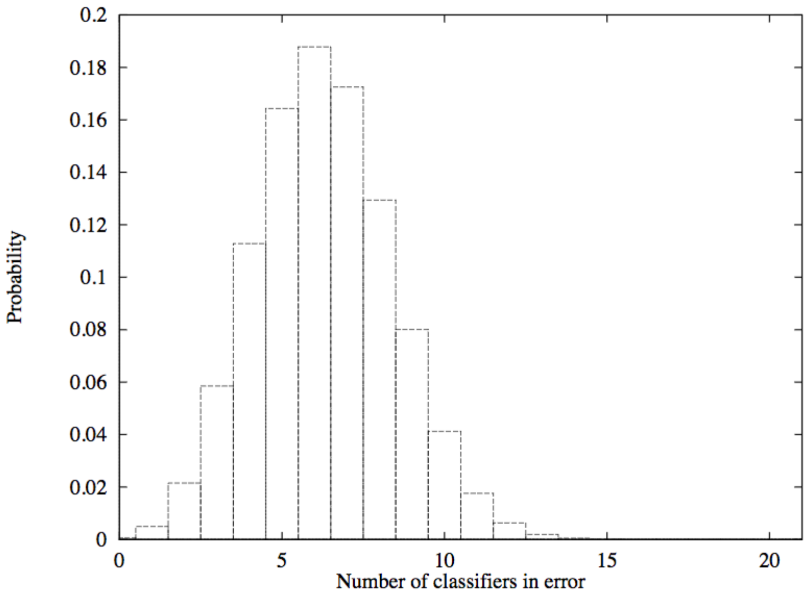
Query Instance ($\mathbf{q}$)

31

Ensembles

More formally if the error rate of each of the L models in an ensemble is less than ½ and if the errors are independent, then the probability that the majority vote of the ensemble will be wrong will be the area under the binomial distribution where more than $L/2$ models are wrong

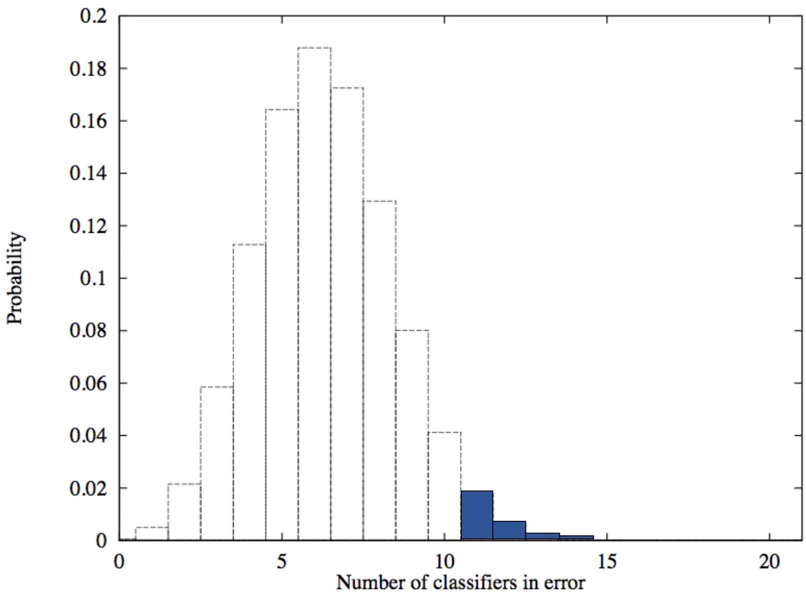
32



An ensemble of **21 models** each with an error rate of **0.3**

Ensemble Methods in Machine Learning
Thomas GDietterich
http://web.engr.oregonstate.edu/~tgd/publications/mcs-ensembles.pdf

33



An ensemble of **21 models** each with an error rate of **0.3**

The area under the curve for 11 or more models being wrong is **0.026**

This is much less than the error of any individual model

Ensemble Methods in Machine Learning
Thomas GDietterich
http://web.engr.oregonstate.edu/~tgd/publications/mcs-ensembles.pdf

34

Ensembles

But models in a real ensemble are never independent so we don't quite do that well

In general we build our ensembles to have two competing characteristics

- Individual models in the ensemble should be strong
- The correlation between the models in the ensemble should be weak (diversity)

35

Practical Ensembles

There are however a series of pracical ensemble approaches

- Bagging
- Random forests
- Boosting
- Gradient boosting
- Stacking

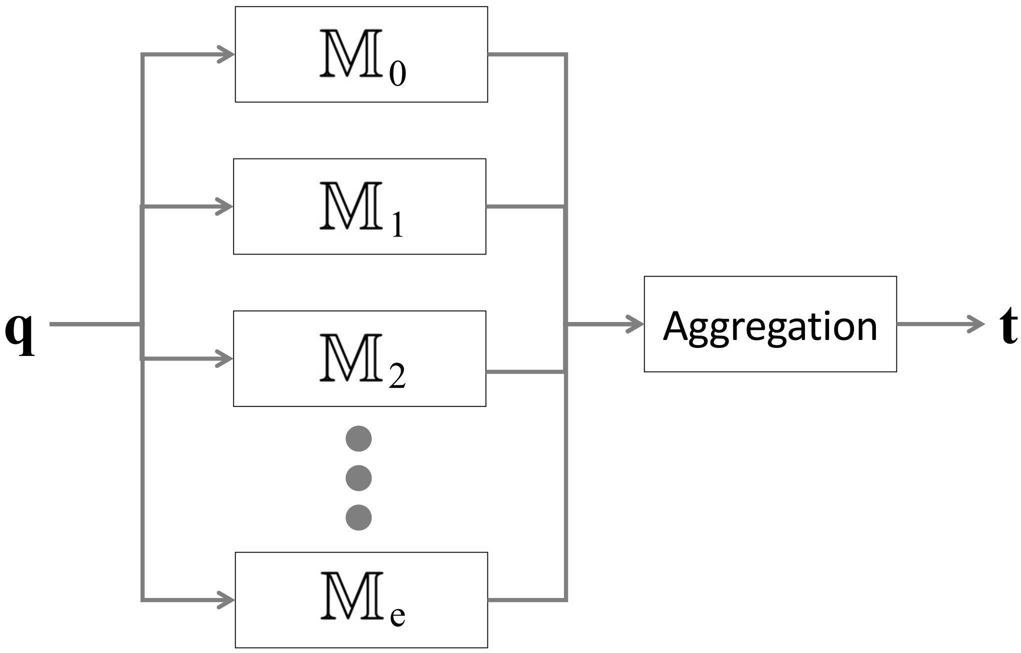
36

Practical Ensembles

There are however a series of pracical ensemble approaches

- Bagging
- **Random forests**
- Boosting
- **Gradient boosting**
- Stacking

37



38

Practical Ensembles

In general we would like our ensembles to have two characteristics

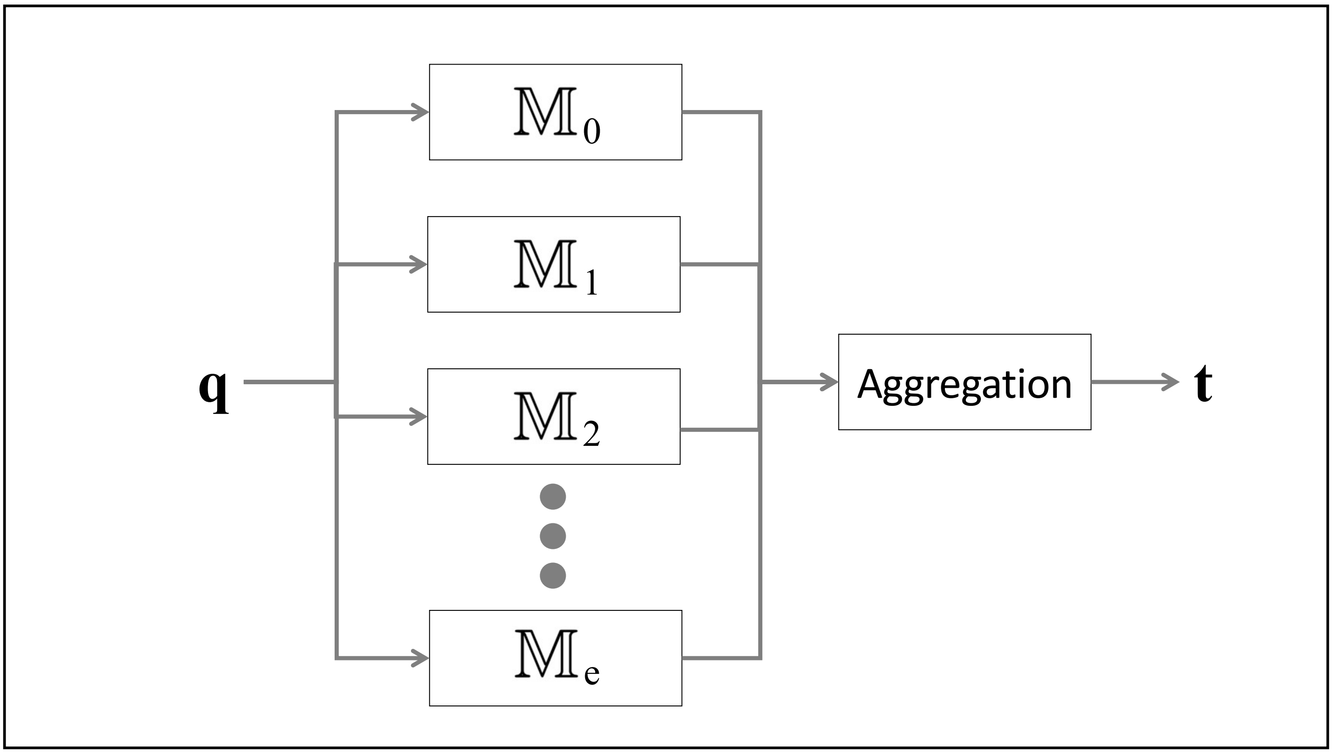
- Individual models in the ensemble should be strong
- The correlation between the models in the ensemble should be weak (diversity)

These two characteristics are in tension with each other

39

RANDOM FORESTS

40



41

Random Forests

It is an extension of Decision Trees that improves accuracy and reduces overfitting by combining multiple trees.

Simple but very powerful ensembling technique

- Trains *e* models in parallel using bootstrapped and sub-space sampled data samples from an overall training set
- Aggregates using majority voting

Bootstrapping is a sampling with replacement technique, meaning some data can be chosen multiple times

therefore each model is trained on a random subset of the training data but some data may appear multiple time while others are left out

Breiman, Leo. "Random forests." *Machine learning* 45.1 (2001): 5-32.

42

sub-space sampled data means that each tree is trained on a random subset of features , not all features are garenteed to be trained

ID	EXERCISE	SMOKER	OBESE	FAMILY	RISK
1	daily	false	false	yes	low
2	weekly	true	false	yes	high
3	daily	false	false	no	low
4	rarely	true	true	yes	high
5	rarely	true	true	no	high

Bagging and Subspace Sampling

ID	EXERCISE	FAMILY	RISK
1	daily	yes	low
2	weekly	yes	high
2	weekly	yes	high
5	rarely	no	high
5	rarely	no	high

Bootstrap Sample A

ID	SMOKER	OBESE	RISK
1	false	false	low
2	true	false	high
2	true	false	high
4	true	true	high
5	true	true	high

Bootstrap Sample B

ID	OBESE	FAMILY	RISK
1	false	yes	low
1	false	yes	low
2	false	yes	high
4	true	yes	high
5	true	no	high

Bootstrap Sample C

43

ID	EXERCISE	FAMILY	RISK
1	daily	yes	low
2	weekly	yes	high
2	weekly	yes	high
5	rarely	no	high
5	rarely	no	high

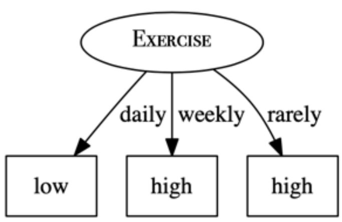
Bootstrap Sample A

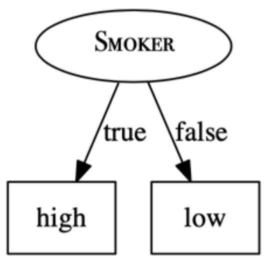
ID	SMOKER	OBESE	RISK
1	false	false	low
2	true	false	high
2	true	false	high
4	true	true	high
5	true	true	high

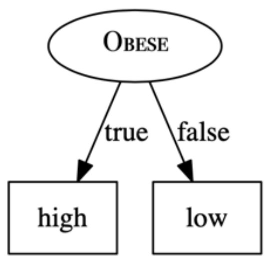
Bootstrap Sample B

ID	OBESE	FAMILY	RISK
1	false	yes	low
1	false	yes	low
2	false	yes	high
4	true	yes	high
5	true	no	high

Bootstrap Sample C



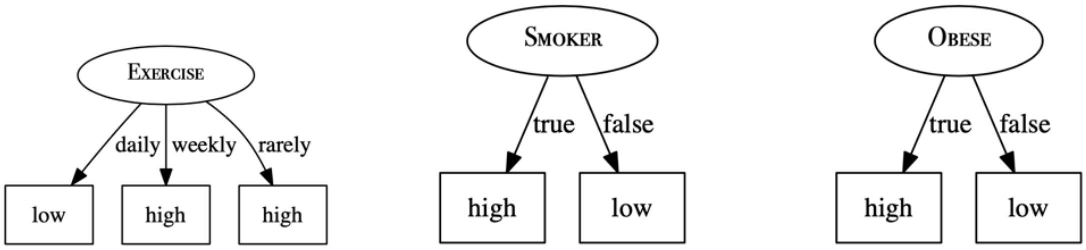




Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

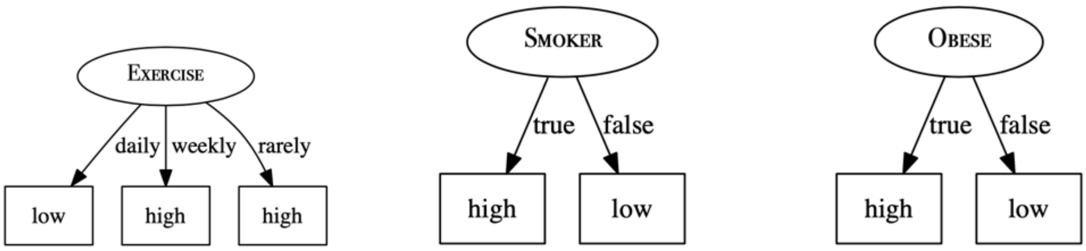
44

EXERCISE='*rarely*', SMOKER='*false*', OBESE='*true*', FAMILY='*yes*'



45

EXERCISE='*rarely*', SMOKER='*false*', OBESE='*true*', FAMILY='*yes*'

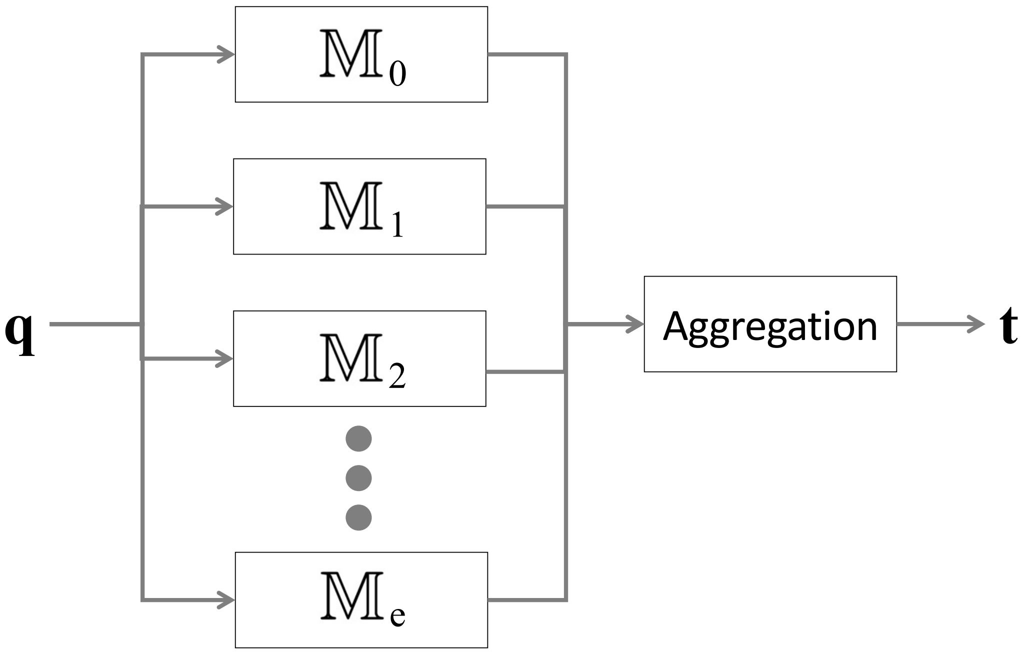


RISK='*high*' RISK='*low*' RISK='*high*'

46

GRADIENT BOOSTING

47



48

Gradient Boosting

Gradient boosting creates an ensemble model by iteratively adding learners - similar to AdaBoost

Gradient boosting is more aggessive fitting each new model directly to the errors of the ensemble (as constituted up to the current iteration) rather then to a weighted dataset which is more subtle

Gradient boostin builds a series of models sequentially and combines their outputs

Gradient Boosting

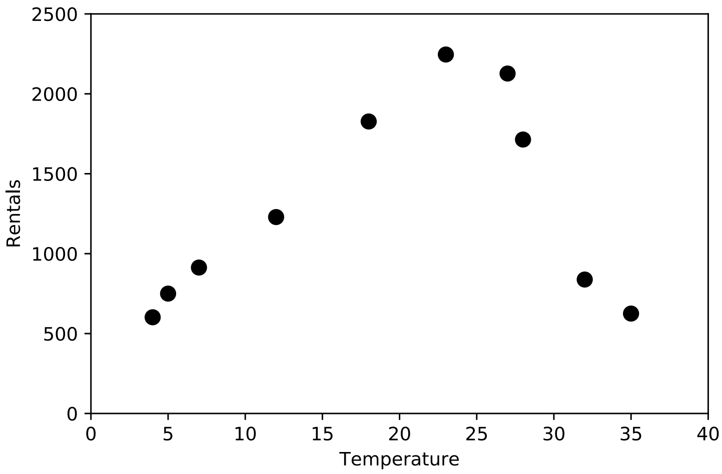
Gradient boosting is best explained in the context of predicting a continuous target

In a regression task we are trying to predict a continuous target and the goal of training is to minimise some measure of error; e.g., the mean squared error:

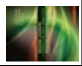
$$MSE = \frac{\sum_{i=1}^{n}(t_i - \mathbb{M}(\mathbf{d}_i))^2}{n}$$

Example: A simple bicycle demand predictions dataset and the workings of the first iterations of training a gradient boosting model.

ID	Temp	Rentals
1	4	602
2	5	750
3	7	913
4	12	1229
5	18	1827
6	23	2246
7	27	2127
8	28	1714
9	32	838
10	35	625



Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy



51

Gradient Boosting

At each iteration gradient boosting assumes we already have a model that can make predictions (this model can be very weak)

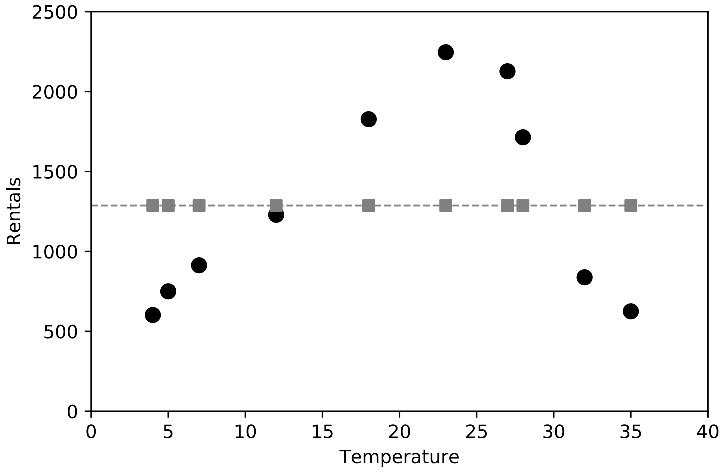
For example, in the first iteration this model may simply predict the mean of the target

$$\mathbb{M}_0(\mathbf{d}) = \frac{1}{n}\sum_i t_i$$

52

Example: A simple bicycle demand predictions dataset and the workings of the first iterations of training a gradient boosting model.

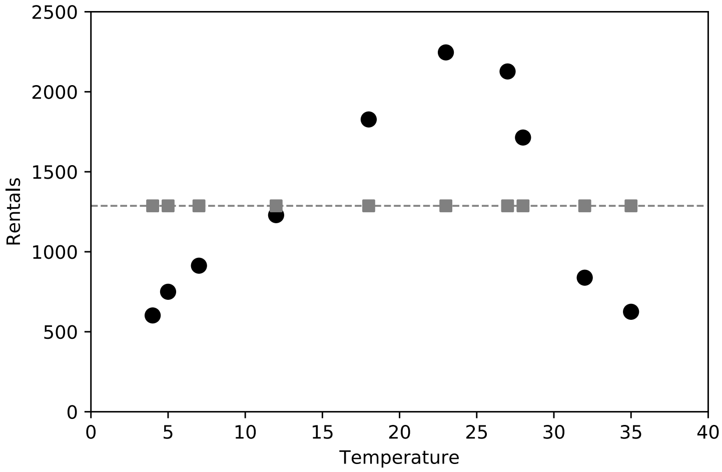
ID	TEMP	RENTALS	$\mathbb{M}_0(\mathbf{d})$
1	4	602	1 287.1
2	5	750	1 287.1
3	7	913	1 287.1
4	12	1229	1 287.1
5	18	1827	1 287.1
6	23	2246	1 287.1
7	27	2127	1 287.1
8	28	1714	1 287.1
9	32	838	1 287.1
10	35	625	1 287.1



53

Example: A simple bicycle demand predictions dataset and the workings of the first iterations of training a gradient boosting model.

ID	TEMP	RENTALS	$\mathbb{M}_0(\mathbf{d})$	$t - \mathbb{M}_0(\mathbf{d})$
1	4	602	1 287.1	-685.1
2	5	750	1 287.1	-537.1
3	7	913	1 287.1	-374.1
4	12	1229	1 287.1	-58.1
5	18	1827	1 287.1	539.9
6	23	2246	1 287.1	958.9
7	27	2127	1 287.1	839.9
8	28	1714	1 287.1	426.9
9	32	838	1 287.1	-449.1
10	35	625	1 287.1	-662.1



Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com) John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

54

Gradient Boosting

Gradient boosting improves this existing model by adding a new model that reduces the error of the existing model

$$\mathbb{M}_1(\mathbf{d}) = \mathbb{M}_0(\mathbf{d}) + \mathbb{M}_{\Delta 1}(\mathbf{d})$$

55

Gradient Boosting

Gradient boosting improves this existing model by adding a new model that reduces the error of the existing model

$$\mathbb{M}_1(\mathbf{d}) = \mathbb{M}_0(\mathbf{d}) + \mathbb{M}_{\Delta 1}(\mathbf{d})$$

And we repeat this multiple times

$$\mathbb{M}_i(\mathbf{d}) = \mathbb{M}_{i-1}(\mathbf{d}) + \mathbb{M}_{\Delta i}(\mathbf{d})$$

56

Gradient Boosting

The question is how to define the model that we add to the existing model

The solution adopted by gradient boosting is based on the intuition that the perfect model to add would be the model that made the predictions for the total ensemble correct:

$$\mathbb{M}_i(\mathbf{d}) = \mathbb{M}_{i-1}(\mathbf{d}) + \boxed{\mathbb{M}_{\Delta i}(\mathbf{d})} = t$$

57

Gradient Boosting

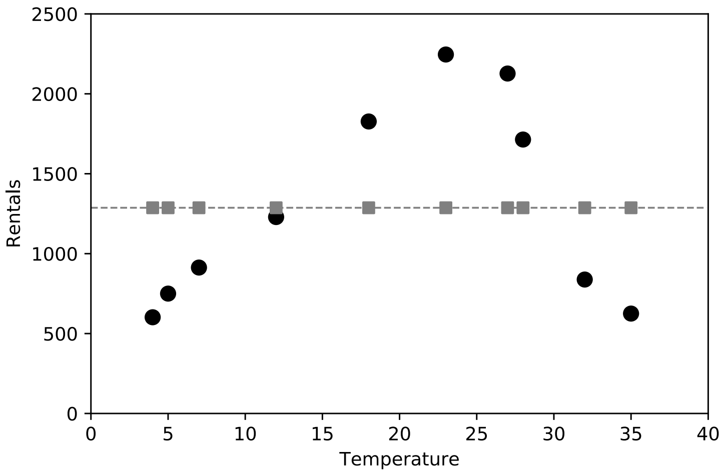
From the above equation we can see that the best model to fit would be the model that predicts the difference between the old models prediction and the true prediction:

$$\mathbb{M}_{\Delta i}(\mathbf{d}) = t - \mathbb{M}_{i-1}(\mathbf{d})$$

58

Example: A simple bicycle demand predictions dataset and the workings of the first iterations of training a gradient boosting model.

ID	TEMP	RENTALS	$\mathbb{M}_0(\mathbf{d})$	$t - \mathbb{M}_0(\mathbf{d})$
1	4	602	1 287.1	-685.1
2	5	750	1 287.1	-537.1
3	7	913	1 287.1	-374.1
4	12	1229	1 287.1	-58.1
5	18	1827	1 287.1	539.9
6	23	2246	1 287.1	958.9
7	27	2127	1 287.1	839.9
8	28	1714	1 287.1	426.9
9	32	838	1 287.1	-449.1
10	35	625	1 287.1	-662.1



Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

59

Gradient Boosting

So gradient boosting trains the new model to add to the ensemble by training the model to predict the errors (in regression terms the residuals) of the old model

We can use any base model in this ensemble, but it is typical to use shallow decision trees – i.e. *decision stumps*

60

Example: A simple bicycle demand predictions dataset and the workings of the first iterations of training a gradient boosting model.

ID	TEMP	RENTALS	$\mathbb{M}_0(\mathbf{d})$	$t - \mathbb{M}_0(\mathbf{d})$	$\mathbb{M}_{\Delta 1}(\mathbf{d})$
1	4	602	1 287.1	-685.1	-460.9
2	5	750	1 287.1	-537.1	-460.9
3	7	913	1 287.1	-374.1	-460.9
4	12	1229	1 287.1	-58.1	-460.9
5	18	1827	1 287.1	539.9	691.4
6	23	2246	1 287.1	958.9	691.4
7	27	2127	1 287.1	839.9	691.4
8	28	1714	1 287.1	426.9	691.4
9	32	838	1 287.1	-449.1	-460.9
10	35	625	1 287.1	-662.1	-460.9

61

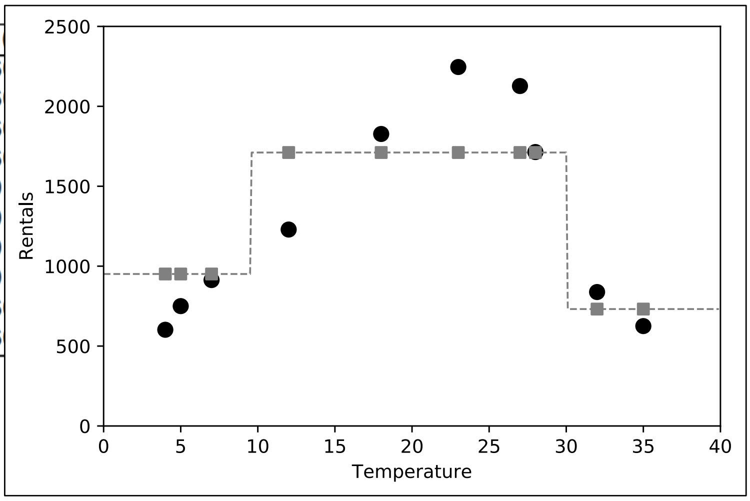
Example: A simple bicycle demand predictions dataset and the workings of the first iterations of training a gradient boosting model.

ID	TEMP	RENTALS	$\mathbb{M}_0(\mathbf{d})$	$t - \mathbb{M}_0(\mathbf{d})$	$\mathbb{M}_{\Delta 1}(\mathbf{d})$	$\mathbb{M}_1(\mathbf{d})$
1	4	602	1 287.1	-685.1	-460.9	826.2
2	5	750	1 287.1	-537.1	-460.9	826.2
3	7	913	1 287.1	-374.1	-460.9	826.2
4	12	1229	1 287.1	-58.1	-460.9	826.2
5	18	1827	1 287.1	539.9	691.4	1 978.5
6	23	2246	1 287.1	958.9	691.4	1 978.5
7	27	2127	1 287.1	839.9	691.4	1 978.5
8	28	1714	1 287.1	426.9	691.4	1 978.5
9	32	838	1 287.1	-449.1	-460.9	826.2
10	35	625	1 287.1	-662.1	-460.9	826.2

62

Example: A simple bicycle demand predictions dataset and the workings of the first iterations of training a gradient boosting model.

ID	TEMP	RENTALS	$\mathbb{M}_0(\mathbf{d})$	$t - \mathbb{M}_0(\mathbf{d})$	$\mathbb{M}_{\Delta 1}($
1	4	602	1 287.1	-685.1	-46
2	5	750	1 287.1	-537.1	-46
3	7	913	1 287.1	-374.1	-46
4	12	1229	1 287.1	-58.1	-46
5	18	1827	1 287.1	539.9	69
6	23	2246	1 287.1	958.9	69
7	27	2127	1 287.1	839.9	69
8	28	1714	1 287.1	426.9	69
9	32	838	1 287.1	-449.1	-46
10	35	625	1 287.1	-662.1	-46



63

Gradient Boosting

$$
\begin{aligned}
\mathbb{M}_4(\mathbf{d}) &= \mathbb{M}_3(\mathbf{d}) + \mathbb{M}_{\Delta 4}(\mathbf{d}) \\
&= (\mathbb{M}_2(\mathbf{d}) + \mathbb{M}_{\Delta 3}(\mathbf{d})) + \mathbb{M}_{\Delta 4}(\mathbf{d}) \\
&= ((\mathbb{M}_1 + \mathbb{M}_{\Delta 2}(\mathbf{d})) + \mathbb{M}_{\Delta 3}(\mathbf{d})) + \mathbb{M}_{\Delta 4}(\mathbf{d}) \\
&= (((\mathbb{M}_0(\mathbf{d}) + \mathbb{M}_{\Delta 1}(\mathbf{d})) + \mathbb{M}_{\Delta 2}(\mathbf{d})) + \mathbb{M}_{\Delta 3}(\mathbf{d})) + \mathbb{M}_{\Delta 4}(\mathbf{d}) \\
&= \mathbb{M}_0(\mathbf{d}) + \mathbb{M}_{\Delta 1}(\mathbf{d}) + \mathbb{M}_{\Delta 2}(\mathbf{d}) + \mathbb{M}_{\Delta 3}(\mathbf{d}) + \mathbb{M}_{\Delta 4}(\mathbf{d})
\end{aligned}
$$

64

Example: A simple bicycle demand predictions dataset and the workings of the first iterations of training a gradient boosting model.

ID	TEMP	RENTALS	$\mathbb{M}_0(\mathbf{d})$	$t - \mathbb{M}_0(\mathbf{d})$	$\mathbb{M}_{\Delta 1}(\mathbf{d})$	$\mathbb{M}_1(\mathbf{d})$	$t - \mathbb{M}_1(\mathbf{d})$
1	4	602	1 287.1	-685.1	-460.9	826.2	-224.2
2	5	750	1 287.1	-537.1	-460.9	826.2	-76.2
3	7	913	1 287.1	-374.1	-460.9	826.2	86.8
4	12	1229	1 287.1	-58.1	-460.9	826.2	402.8
5	18	1827	1 287.1	539.9	691.4	1 978.5	-151.5
6	23	2246	1 287.1	958.9	691.4	1 978.5	267.5
7	27	2127	1 287.1	839.9	691.4	1 978.5	148.5
8	28	1714	1 287.1	426.9	691.4	1 978.5	-264.5
9	32	838	1 287.1	-449.1	-460.9	826.2	11.8
10	35	625	1 287.1	-662.1	-460.9	826.2	-201.2

65

Example: A simple bicycle demand predictions dataset and the workings of the first iterations of training a gradient boosting model.

ID	TEMP	RENTALS	$\mathbb{M}_0(\mathbf{d})$	$t - \mathbb{M}_0(\mathbf{d})$	$\mathbb{M}_{\Delta 1}(\mathbf{d})$	$\mathbb{M}_1(\mathbf{d})$	$t - \mathbb{M}_1(\mathbf{d})$	$\mathbb{M}_{\Delta 2}(\mathbf{d})$
1	4	602	1 287.1	-685.1	-460.9	826.2	-224.2	-167.2
2	5	750	1 287.1	-537.1	-460.9	826.2	-76.2	-167.2
3	7	913	1 287.1	-374.1	-460.9	826.2	86.8	71.6
4	12	1229	1 287.1	-58.1	-460.9	826.2	402.8	71.6
5	18	1827	1 287.1	539.9	691.4	1 978.5	-151.5	71.6
6	23	2246	1 287.1	958.9	691.4	1 978.5	267.5	71.6
7	27	2127	1 287.1	839.9	691.4	1 978.5	148.5	71.6
8	28	1714	1 287.1	426.9	691.4	1 978.5	-264.5	71.6
9	32	838	1 287.1	-449.1	-460.9	826.2	11.8	71.6
10	35	625	1 287.1	-662.1	-460.9	826.2	-201.2	-167.2

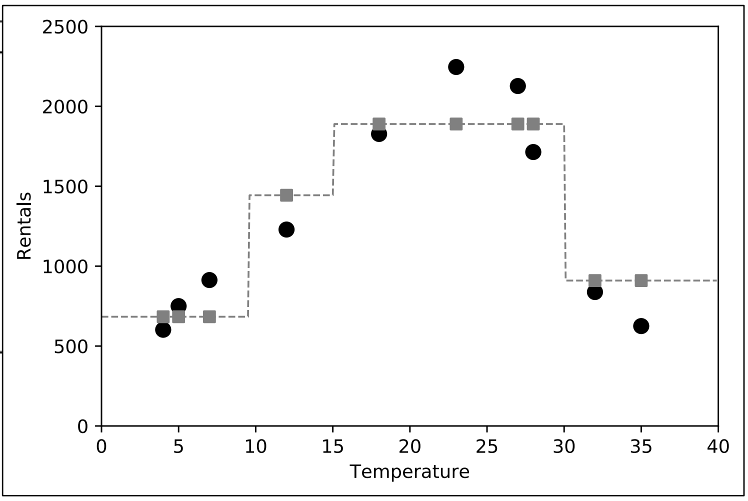
66

Example: A simple bicycle demand predictions dataset and the workings of the first iterations of training a gradient boosting model.

ID	TEMP	RENTALS	$\mathbb{M}_0(\mathbf{d})$	$t - \mathbb{M}_0(\mathbf{d})$	$\mathbb{M}_{\Delta 1}(\mathbf{d})$	$\mathbb{M}_1(\mathbf{d})$	$t - \mathbb{M}_1(\mathbf{d})$	$\mathbb{M}_{\Delta 2}(\mathbf{d})$	$\mathbb{M}_2(\mathbf{d})$
1	4	602	1 287.1	-685.1	-460.9	826.2	-224.2	-167.2	659.0
2	5	750	1 287.1	-537.1	-460.9	826.2	-76.2	-167.2	659.0
3	7	913	1 287.1	-374.1	-460.9	826.2	86.8	71.6	897.8
4	12	1229	1 287.1	-58.1	-460.9	826.2	402.8	71.6	897.8
5	18	1827	1 287.1	539.9	691.4	1 978.5	-151.5	71.6	2 050.1
6	23	2246	1 287.1	958.9	691.4	1 978.5	267.5	71.6	2 050.1
7	27	2127	1 287.1	839.9	691.4	1 978.5	148.5	71.6	2 050.1
8	28	1714	1 287.1	426.9	691.4	1 978.5	-264.5	71.6	2 050.1
9	32	838	1 287.1	-449.1	-460.9	826.2	11.8	71.6	897.8
10	35	625	1 287.1	-662.1	-460.9	826.2	-201.2	-167.2	659.0

Example: A simple bicycle demand predictions dataset and the workings of the first iterations of training a gradient boosting model.

ID	TEMP	RENTALS	$\mathbb{M}_0(\mathbf{d})$	$t - \mathbb{M}_0(\mathbf{d})$	$\mathbb{M}_{\Delta 1}(\mathbf{d})$
1	4	602	1 287.1	-685.1	-460.9
2	5	750	1 287.1	-537.1	-460.9
3	7	913	1 287.1	-374.1	-460.9
4	12	1229	1 287.1	-58.1	-460.9
5	18	1827	1 287.1	539.9	691.4
6	23	2246	1 287.1	958.9	691.4
7	27	2127	1 287.1	839.9	691.4
8	28	1714	1 287.1	426.9	691.4
9	32	838	1 287.1	-449.1	-460.9
10	35	625	1 287.1	-662.1	-460.9



Example: A simple bicycle demand predictions dataset and the workings of the first iterations of training a gradient boosting model.

ID	TEMP	RENTALS	$\mathbb{M}_0(\mathbf{d})$	$t - \mathbb{M}_0(\mathbf{d})$	$\mathbb{M}_{\Delta 1}(\mathbf{d})$	$\mathbb{M}_1(\mathbf{d})$	$t - \mathbb{M}_1(\mathbf{d})$	$\mathbb{M}_{\Delta 2}(\mathbf{d})$	$\mathbb{M}_2(\mathbf{d})$	$t - \mathbb{M}_2(\mathbf{d})$
1	4	602	1 287.1	-685.1	-460.9	826.2	-224.2	-167.2	659.0	-57.0
2	5	750	1 287.1	-537.1	-460.9	826.2	-76.2	-167.2	659.0	91.0
3	7	913	1 287.1	-374.1	-460.9	826.2	86.8	71.6	897.8	15.2
4	12	1229	1 287.1	-58.1	-460.9	826.2	402.8	71.6	897.8	331.2
5	18	1827	1 287.1	539.9	691.4	1 978.5	-151.5	71.6	2 050.1	-223.1
6	23	2246	1 287.1	958.9	691.4	1 978.5	267.5	71.6	2 050.1	195.9
7	27	2127	1 287.1	839.9	691.4	1 978.5	148.5	71.6	2 050.1	76.9
8	28	1714	1 287.1	426.9	691.4	1 978.5	-264.5	71.6	2 050.1	-336.1
9	32	838	1 287.1	-449.1	-460.9	826.2	11.8	71.6	897.8	-59.8
10	35	625	1 287.1	-662.1	-460.9	826.2	-201.2	-167.2	659.0	-34.0

69

Example: A simple bicycle demand predictions dataset and the workings of the first iterations of training a gradient boosting model.

ID	TEMP	RENTALS	$\mathbb{M}_0(\mathbf{d})$	$t - \mathbb{M}_0(\mathbf{d})$	$\mathbb{M}_{\Delta 1}(\mathbf{d})$	$\mathbb{M}_1(\mathbf{d})$	$t - \mathbb{M}_1(\mathbf{d})$	$\mathbb{M}_{\Delta 2}(\mathbf{d})$	$\mathbb{M}_2(\mathbf{d})$	$t - \mathbb{M}_2(\mathbf{d})$	$\mathbb{M}_{\Delta 3}(\mathbf{d})$
1	4	602	1 287.1	-685.1	-460.9	826.2	-224.2	-167.2	659.0	-57.0	-34.1
2	5	750	1 287.1	-537.1	-460.9	826.2	-76.2	-167.2	659.0	91.0	-34.1
3	7	913	1 287.1	-374.1	-460.9	826.2	86.8	71.6	897.8	15.2	-34.1
4	12	1229	1 287.1	-58.1	-460.9	826.2	402.8	71.6	897.8	331.2	-34.1
5	18	1827	1 287.1	539.9	691.4	1 978.5	-151.5	71.6	2 050.1	-223.1	-34.1
6	23	2246	1 287.1	958.9	691.4	1 978.5	267.5	71.6	2 050.1	195.9	136.4
7	27	2127	1 287.1	839.9	691.4	1 978.5	148.5	71.6	2 050.1	76.9	136.4
8	28	1714	1 287.1	426.9	691.4	1 978.5	-264.5	71.6	2 050.1	-336.1	-34.1
9	32	838	1 287.1	-449.1	-460.9	826.2	11.8	71.6	897.8	-59.8	-34.1
10	35	625	1 287.1	-662.1	-460.9	826.2	-201.2	-167.2	659.0	-34.0	-34.1

70

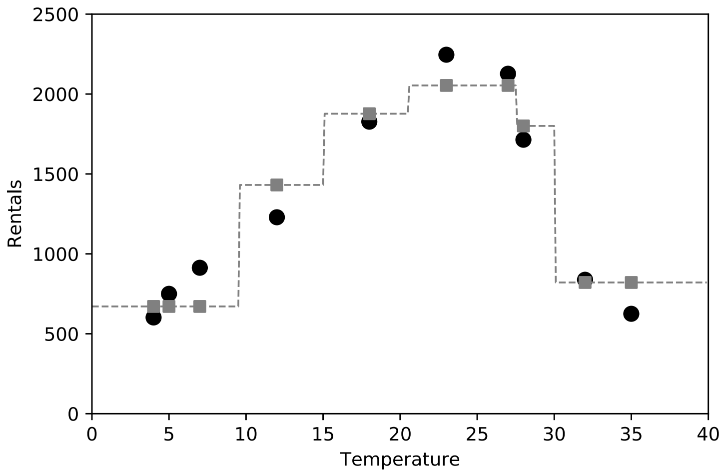
Example: A simple bicycle demand predictions dataset and the workings of the first iterations of training a gradient boosting model.

ID	TEMP	RENTALS	$\mathbb{M}_0(\mathbf{d})$	$t - \mathbb{M}_0(\mathbf{d})$	$\mathbb{M}_{\Delta 1}(\mathbf{d})$	$\mathbb{M}_1(\mathbf{d})$	$t - \mathbb{M}_1(\mathbf{d})$	$\mathbb{M}_{\Delta 2}(\mathbf{d})$	$\mathbb{M}_2(\mathbf{d})$	$t - \mathbb{M}_2(\mathbf{d})$	$\mathbb{M}_{\Delta 3}(\mathbf{d})$	$\mathbb{M}_3(\mathbf{d})$
1	4	602	1 287.1	-685.1	-460.9	826.2	-224.2	-167.2	659.0	-57.0	-34.1	624.9
2	5	750	1 287.1	-537.1	-460.9	826.2	-76.2	-167.2	659.0	91.0	-34.1	624.9
3	7	913	1 287.1	-374.1	-460.9	826.2	86.8	71.6	897.8	15.2	-34.1	863.7
4	12	1229	1 287.1	-58.1	-460.9	826.2	402.8	71.6	897.8	331.2	-34.1	863.7
5	18	1827	1 287.1	539.9	691.4	1 978.5	-151.5	71.6	2 050.1	-223.1	-34.1	2 016.1
6	23	2246	1 287.1	958.9	691.4	1 978.5	267.5	71.6	2 050.1	195.9	136.4	2 186.5
7	27	2127	1 287.1	839.9	691.4	1 978.5	148.5	71.6	2 050.1	76.9	136.4	2 186.5
8	28	1714	1 287.1	426.9	691.4	1 978.5	-264.5	71.6	2 050.1	-336.1	-34.1	2 016.1
9	32	838	1 287.1	-449.1	-460.9	826.2	11.8	71.6	897.8	-59.8	-34.1	863.7
10	35	625	1 287.1	-662.1	-460.9	826.2	-201.2	-167.2	659.0	-34.0	-34.1	624.9

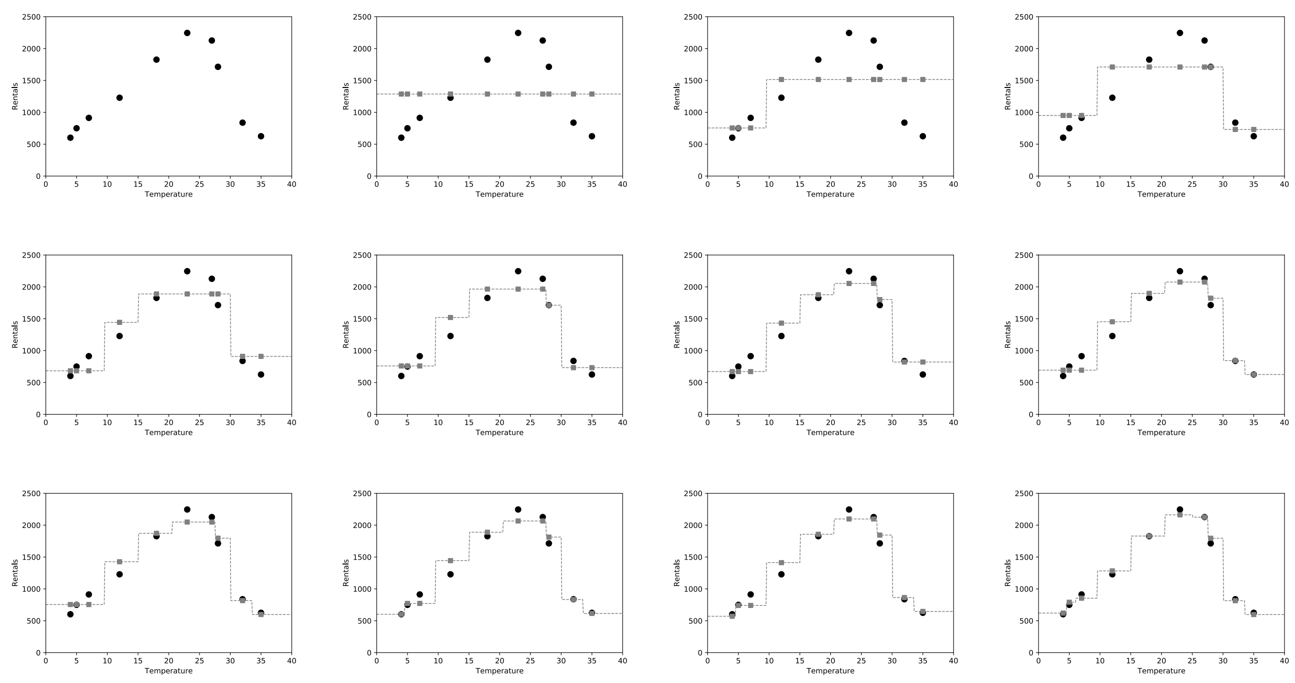
71

Example: A simple bicycle demand predictions dataset and the workings of the first iterations of training a gradient boosting model.

ID	TEMP	RENTALS	$\mathbb{M}_0(\mathbf{d})$	$t - \mathbb{M}_0(\mathbf{d})$	$\mathbb{M}_{\Delta 1}(\mathbf{d})$
1	4	602	1 287.1	-685.1	-460.9
2	5	750	1 287.1	-537.1	-460.9
3	7	913	1 287.1	-374.1	-460.9
4	12	1229	1 287.1	-58.1	-460.9
5	18	1827	1 287.1	539.9	691.4
6	23	2246	1 287.1	958.9	691.4
7	27	2127	1 287.1	839.9	691.4
8	28	1714	1 287.1	426.9	691.4
9	32	838	1 287.1	-449.1	-460.9
10	35	625	1 287.1	-662.1	-460.9



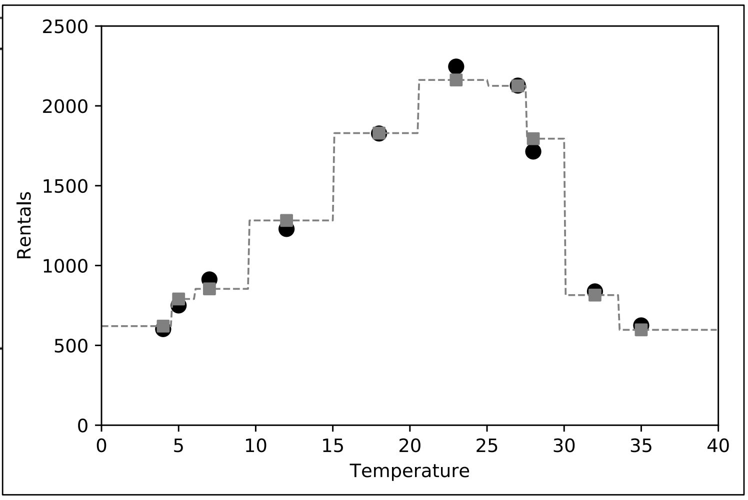
72



73

Example: A simple bicycle demand predictions dataset and the workings of the first iterations of training a gradient boosting model.

ID	TEMP	RENTALS	$\mathbb{M}_0(\mathbf{d})$	$t - \mathbb{M}_0(\mathbf{d})$	$\mathbb{M}_{\Delta 1}(\mathbf{d})$
1	4	602	1 287.1	-685.1	-460.9
2	5	750	1 287.1	-537.1	-460.9
3	7	913	1 287.1	-374.1	-460.9
4	12	1229	1 287.1	-58.1	-460.9
5	18	1827	1 287.1	539.9	691.4
6	23	2246	1 287.1	958.9	691.4
7	27	2127	1 287.1	839.9	691.4
8	28	1714	1 287.1	426.9	691.4
9	32	838	1 287.1	-449.1	-460.9
10	35	625	1 287.1	-662.1	-460.9



$\mathbb{M}_{20}$

74

Gradient Boosting Algorithm

Algorithm: GB($\mathcal{D}$,E) returns $\mathbb{M}$

let $\mathbb{M}_0 = \frac{1}{n}\sum_i t_i$

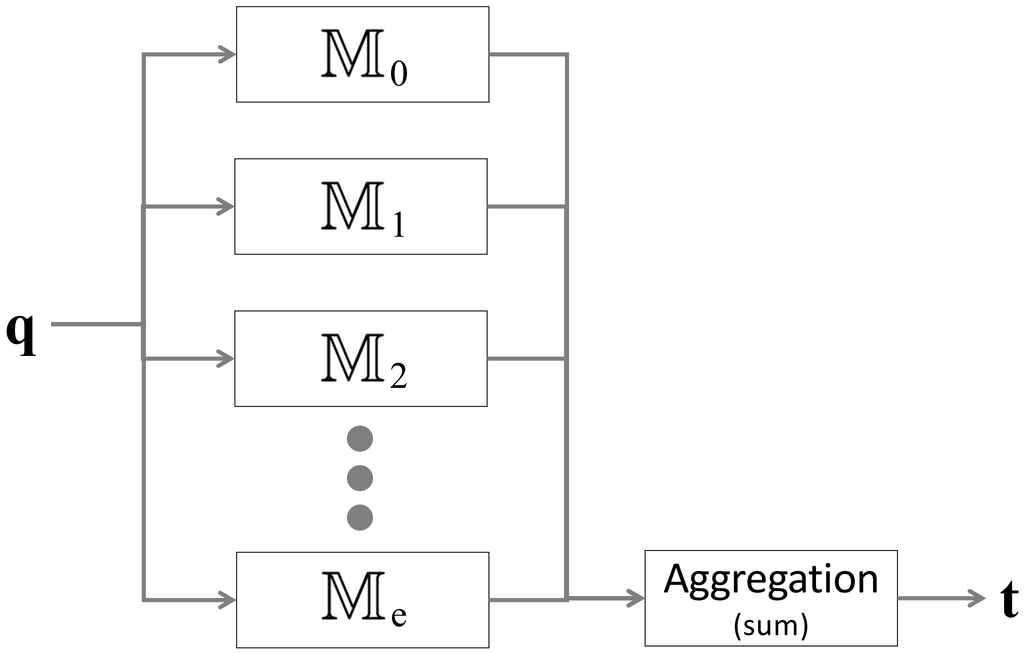
for i = 1 to E

Let $\Delta_i = t - \mathbb{M}_{i-1}(\mathbf{d})$

Train $\mathbb{M}_{\Delta i}$ to predict Δ_i

Let $\mathbb{M}_i = \mathbb{M}_{i-1} + \mathbb{M}_{\Delta i}$

75



76

Gradient Boosting

$$
\begin{aligned}
\mathbb{M}_4(\mathbf{d}) &= \mathbb{M}_3(\mathbf{d}) + \mathbb{M}_{\Delta 4}(\mathbf{d}) \\
&= (\mathbb{M}_2(\mathbf{d}) + \mathbb{M}_{\Delta 3}(\mathbf{d})) + \mathbb{M}_{\Delta 4}(\mathbf{d}) \\
&= ((\mathbb{M}_1 + \mathbb{M}_{\Delta 2}(\mathbf{d})) + \mathbb{M}_{\Delta 3}(\mathbf{d})) + \mathbb{M}_{\Delta 4}(\mathbf{d}) \\
&= (((\mathbb{M}_0(\mathbf{d}) + \mathbb{M}_{\Delta 1}(\mathbf{d})) + \mathbb{M}_{\Delta 2}(\mathbf{d})) + \mathbb{M}_{\Delta 3}(\mathbf{d})) + \mathbb{M}_{\Delta 4}(\mathbf{d}) \\
&= \mathbb{M}_0(\mathbf{d}) + \mathbb{M}_{\Delta 1}(\mathbf{d}) + \mathbb{M}_{\Delta 2}(\mathbf{d}) + \mathbb{M}_{\Delta 3}(\mathbf{d}) + \mathbb{M}_{\Delta 4}(\mathbf{d})
\end{aligned}
$$

77

Why *Gradient* Boosting?

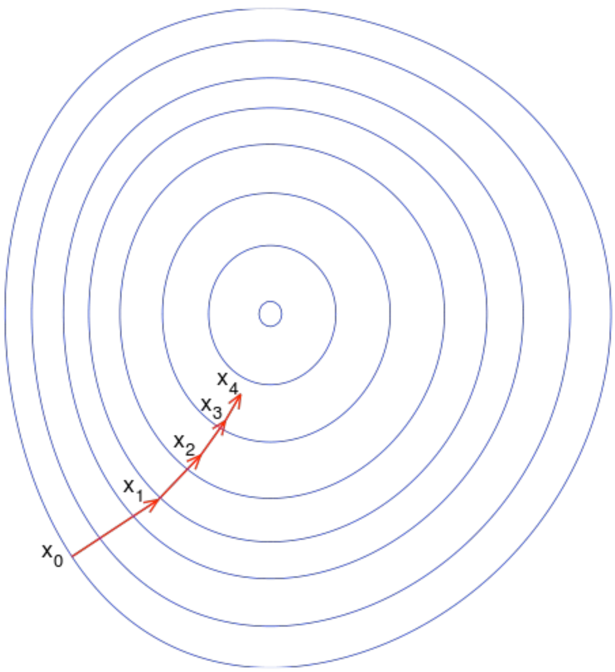
Gradient boosting is called gradient boosting because we can treat the residuals

$$t_i - \mathbb{M}_{n-1}(\mathbf{d}_i)$$

as the negative gradients of the squared error loss function

So, under the hood gradient boosting is essentially doing gradient descent on an error surface

78



https://en.wikipedia.org/wiki/Gradient_descent#/media/File:Gradient_descent.svg

79

Gradient Boosting Variants

There are lots of variants of gradient boosting

- Different kinds of loss functions are common (least squares, huber, ...)
- Gradient boosting can be implemented with any kinds of base models (small decision trees, ~5 levels, are common)
- Stochastic gradient boosting adds subsampling to each iteration and has been shown to prevent overfitting

80

Gradient Boosting Variants

- Learning rate is often added which decreases the influence of each subsequent tree in a model
- Modifying the algorithm for classification is not difficult - changes in loss functions
- XGBoost is a nice, powerful, scalable implementation of gradient boosting that is in widespread use

81

SUMMARY

82

Summary

Supervised learning involves bulding prediction models that learn patterns between a set of descriptive fetaures and a target feature based on a large labelled dataset

Training a model can be viewed as a search process

Inductive bias is required for this search process to converge

83

Summary

Ensembles are amongst the most powerful supervised learning techniques

Random forests, in particular, are very simple but very effective

84

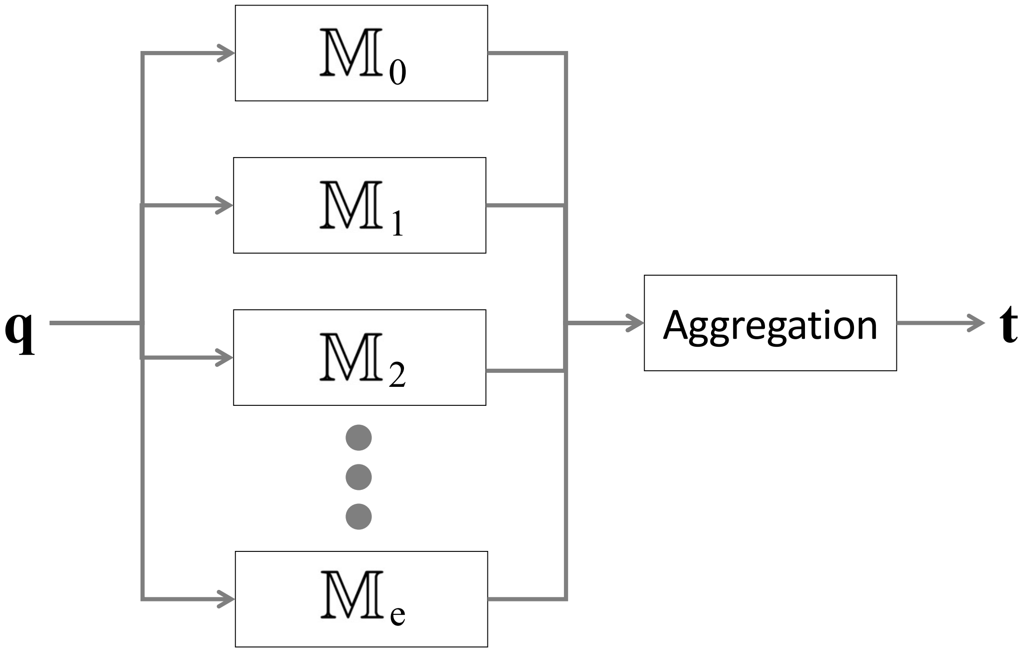
Questions

?

85

BAGGING (OPTIONAL)

86



87

Bagging

Very simple ensemble training technique

- Trains *e* models in parallel using bootstrapped data samples from an overall training set (100% sampling with replacement)
- Aggregates using majority voting
- *Boostrapped aggregating = bagging*

Breiman, Leo. "Bagging predictors." *Machine learning* 24.2 (1996): 123-140.

88

Dataset

ID	EXERCISE	SMOKER	OBESE	FAMILY	RISK
1	daily	false	false	yes	low
2	weekly	true	false	yes	high
3	daily	false	false	no	low
4	rarely	true	true	yes	high
5	rarely	true	true	no	high

89

Dataset

ID	EXERCISE	SMOKER	OBESE	FAMILY	RISK
1	daily	false	false	yes	low
2	weekly	true	false	yes	high
3	daily	false	false	no	low
4	rarely	true	true	yes	high
5	rarely	true	true	no	high

Bootstrap Sample A

ID	EXERCISE	SMOKER	OBESE	FAMILY	RISK
1	daily	false	false	yes	low
2	weekly	true	false	yes	high
2	weekly	true	false	yes	high
5	rarely	true	true	no	high
5	rarely	true	true	no	high

90

Bootstrap Sample A

ID	Exercise	Smoker	Obese	Family	Risk
1	daily	false	false	yes	low
2	weekly	true	false	yes	high
2	weekly	true	false	yes	high
5	rarely	true	true	no	high
5	rarely	true	true	no	high

Bootstrap Sample B

ID	Exercise	Smoker	Obese	Family	Risk
1	daily	false	false	yes	low
2	weekly	true	false	yes	high
2	weekly	true	false	yes	high
3	daily	false	false	no	low
4	rarely	true	true	yes	high

Bootstrap Sample C

ID	Exercise	Smoker	Obese	Family	Risk
2	weekly	true	false	yes	high
2	weekly	true	false	yes	high
3	daily	false	false	no	low
3	daily	false	false	no	low
4	rarely	true	true	yes	high

91

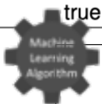
Bootstrap Sample A

ID	Exercise	Smoker	Obese	Family	Risk
1	daily	false	false	yes	low
2	weekly	true	false	yes	high
2	weekly	true	false	yes	high
5	rarely	true	true	no	high
5	rarely	true	true	no	high

$\mathbb{M}_0$

Bootstrap Sample B

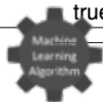
ID	Exercise	Smoker	Obese	Family	Risk
1	daily	false	false	yes	low
2	weekly	true	false	yes	high
2	weekly	true	false	yes	high
3	daily	false	false	no	low
4	rarely	true	true	yes	high



$\mathbb{M}_1$

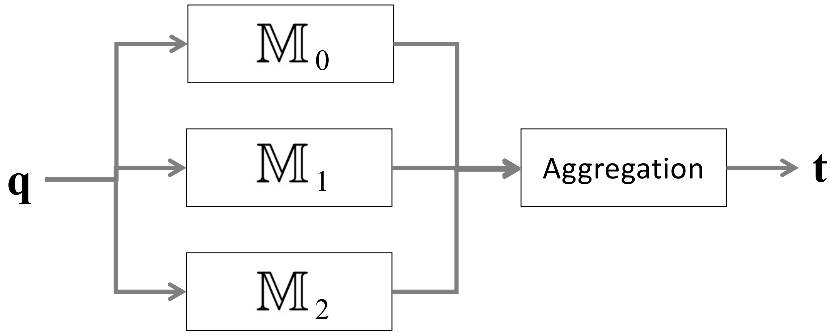
Bootstrap Sample C

ID	Exercise	Smoker	Obese	Family	Risk
2	weekly	true	false	yes	high
2	weekly	true	false	yes	high
3	daily	false	false	no	low
3	daily	false	false	no	low
4	rarely	true	true	yes	high



$\mathbb{M}_2$

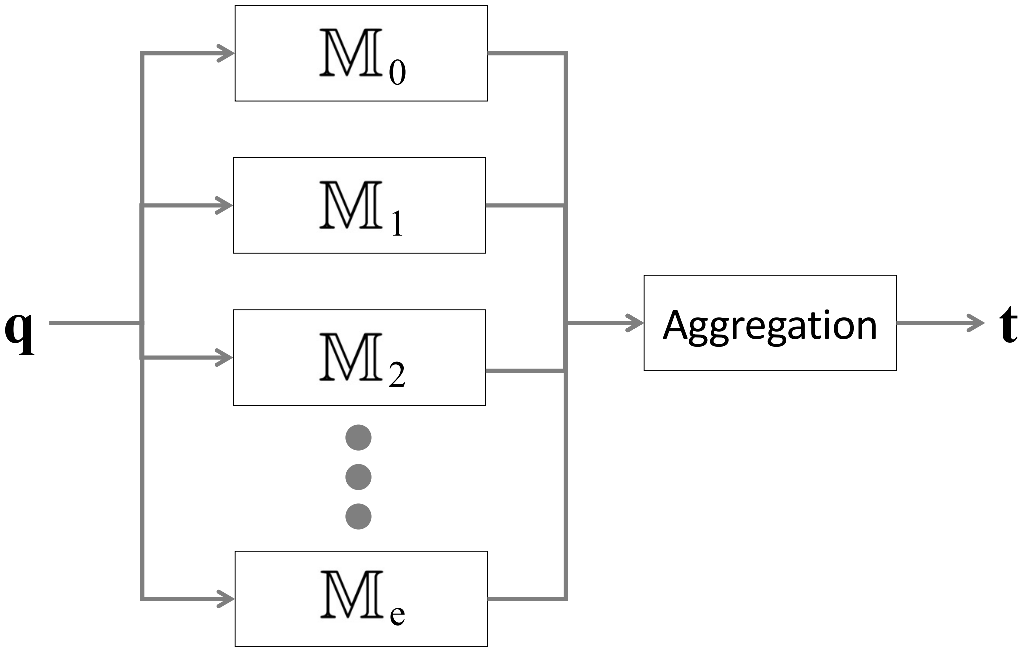
92



93

BOOSTING (OPTIONAL)

94



95

Boosting

Boosting works by iteratively creating models and adding them to the ensemble

- Each new model added to the ensemble is biased to pay more attention to instances that previous models miss-classified
- This is done by incrementally adapting the dataset used to train the models
- The iteration stops when a predefined number of models have been added

96

Boosting

Boosting uses a weighted dataset

- Each instance has an associated weight $\mathbf{w}_i \geq 0$,
- Initially set to $1/n$ where n is the number of instances in the dataset
- After each model is added to the ensemble it is tested on the training data and the weights are adjusted
- These weights are used as a distribution over which the full dataset is sampled for each training dataset

Boosting

During each training iteration the algorithm:

- Induces a model and calculates the total error, ε, by summing the weights of the training instances for which the predictions made by the model are incorrect.

Boosting

During each training iteration the algorithm:

- Increases the weights for the instances misclassified using:

$$\mathbf{w}[i] \leftarrow \mathbf{w}[i] \times \left(\frac{1}{2 \times \epsilon}\right)$$

- Decreases the weights for the instances correctly classified:

$$\mathbf{w}[i] \leftarrow \mathbf{w}[i] \times \left(\frac{1}{2 \times (1 - \epsilon)}\right)$$

99

Boosting

During each training iteration the algorithm:

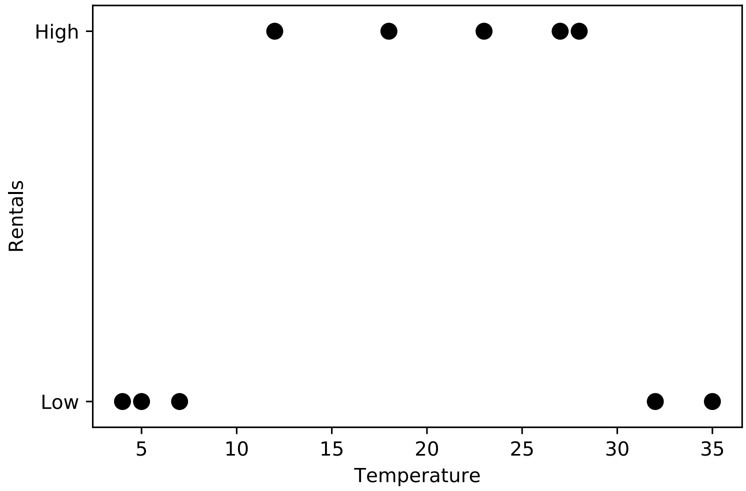
- Calculate a confidence factor, α, for the model such that α increases as ε decreases:

$$\alpha = \frac{1}{2} \times log_e \left(\frac{1 - \epsilon}{\epsilon}\right)$$

100

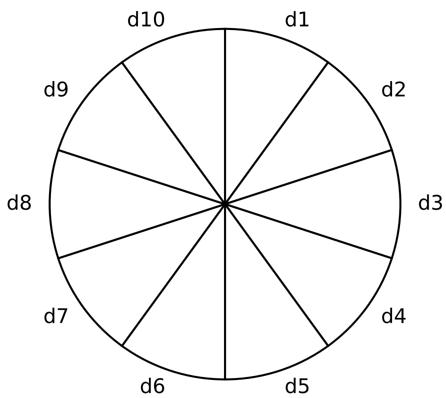
Example: A simple bicycle demand predictions dataset and the workings of the first three iterations of training an ensemble model using boosting to predict RENTALS given TEMP

ID	TEMP	RENTALS
1	4	'Low'
2	5	'Low'
3	7	'Low'
4	12	'High'
5	18	'High'
6	23	'High'
7	27	'High'
8	28	'High'
9	32	'Low'
10	35	'Low'



Example : A simple bicycle demand predictions dataset and the workings of the first three iterations of training an ensemble model using boosting to predict RENTALS given TEMP

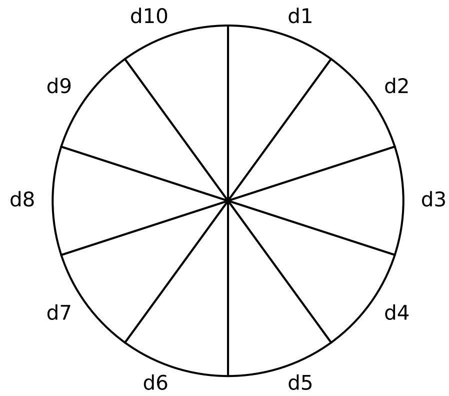
			Iteration 0		
ID	TEMP	RENTALS	Dist.	Freq.	$\mathbb{M}_0(\mathbf{d})$
1	4	'Low'	0.100		
2	5	'Low'	0.100		
3	7	'Low'	0.100		
4	12	'High'	0.100		
5	18	'High'	0.100		
6	23	'High'	0.100		
7	27	'High'	0.100		
8	28	'High'	0.100		
9	32	'Low'	0.100		
10	35	'Low'	0.100		



Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

Example: A simple bicycle demand predictions dataset and the workings of the first three iterations of training an ensemble model using boosting to predict RENTALS given TEMP

			Iteration 0		
ID	TEMP	RENTALS	Dist.	Freq.	$\mathbb{M}_0(\mathbf{d})$
1	4	'Low'	0.100	2	
2	5	'Low'	0.100	1	
3	7	'Low'	0.100	0	
4	12	'High'	0.100	1	
5	18	'High'	0.100	1	
6	23	'High'	0.100	1	
7	27	'High'	0.100	1	
8	28	'High'	0.100	1	
9	32	'Low'	0.100	2	
10	35	'Low'	0.100	0	

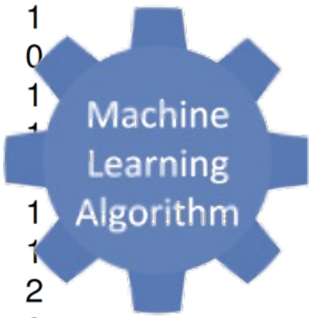


Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com) John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

103

Example: A simple bicycle demand predictions dataset and the workings of the first three iterations of training an ensemble model using boosting to predict RENTALS given TEMP

			Iteration 0		
ID	TEMP	RENTALS	Dist.	Freq.	$\mathbb{M}_0(\mathbf{d})$
1	4	'Low'	0.100	2	
2	5	'Low'	0.100	1	
3	7	'Low'	0.100	0	
4	12	'High'	0.100	1	
5	18	'High'	0.100	1	
6	23	'High'	0.100		
7	27	'High'	0.100	1	
8	28	'High'	0.100	1	
9	32	'Low'	0.100	2	
10	35	'Low'	0.100	0	

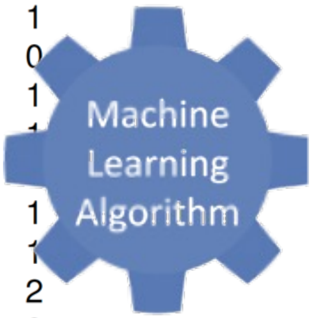


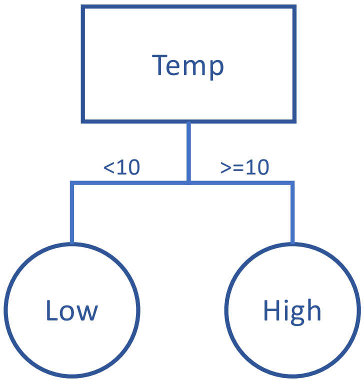
Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com) John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

104

Example: A simple bicycle demand predictions dataset and the workings of the first three iterations of training an ensemble model using boosting to predict RENTALS given TEMP

			Iteration 0		
ID	TEMP	RENTALS	Dist.	Freq.	$\mathbb{M}_0(\mathbf{d})$
1	4	'Low'	0.100	2	
2	5	'Low'	0.100	1	
3	7	'Low'	0.100	0	
4	12	'High'	0.100	1	
5	18	'High'	0.100	[illegible]	
6	23	'High'	0.100	[illegible]	
7	27	'High'	0.100	1	
8	28	'High'	0.100	1	
9	32	'Low'	0.100	2	
10	35	'Low'	0.100	0	

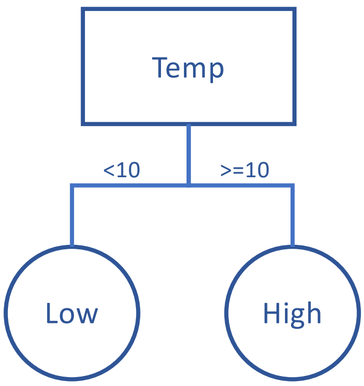




105

Example: A simple bicycle demand predictions dataset and the workings of the first three iterations of training an ensemble model using boosting to predict RENTALS given TEMP

			Iteration 0		
ID	TEMP	RENTALS	Dist.	Freq.	$\mathbb{M}_0(\mathbf{d})$
1	4	'Low'	0.100	2	'Low'
2	5	'Low'	0.100	1	'Low'
3	7	'Low'	0.100	0	'Low'
4	12	'High'	0.100	1	'High'
5	18	'High'	0.100	1	'High'
6	23	'High'	0.100	1	'High'
7	27	'High'	0.100	1	'High'
8	28	'High'	0.100	1	'High'
9	32	'Low'	0.100	2	'High'
10	35	'Low'	0.100	0	'High'



Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

106

Example: A simple bicycle demand predictions dataset and the workings of the first three iterations of training an ensemble model using boosting to predict RENTALS given TEMP

			Iteration 0		
ID	TEMP	RENTALS	Dist.	Freq.	$\mathbb{M}_0(\mathbf{d})$
1	4	'Low'	0.100	2	'Low'
2	5	'Low'	0.100	1	'Low'
3	7	'Low'	0.100	0	'Low'
4	12	'High'	0.100	1	'High'
5	18	'High'	0.100	1	'High'
6	23	'High'	0.100	1	'High'
7	27	'High'	0.100	1	'High'
8	28	'High'	0.100	1	'High'
9	32	'Low'	0.100	2	'High'
10	35	'Low'	0.100	0	'High'

ϵ = (0.100 + 0.100)
= 0.200

Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

107

Example: A simple bicycle demand predictions dataset and the workings of the first three iterations of training an ensemble model using boosting to predict RENTALS given TEMP

			Iteration 0		
ID	TEMP	RENTALS	Dist.	Freq.	$\mathbb{M}_0(\mathbf{d})$
1	4	'Low'	0.100	2	'Low'
2	5	'Low'	0.100	1	'Low'
3	7	'Low'	0.100	0	'Low'
4	12	'High'	0.100	1	'High'
5	18	'High'	0.100	1	'High'
6	23	'High'	0.100	1	'High'
7	27	'High'	0.100	1	'High'
8	28	'High'	0.100	1	'High'
9	32	'Low'	0.100	2	'High'
10	35	'Low'	0.100	0	'High'

ϵ = (0.100 + 0.100)
= 0.200

α = 0.5 * $\log_e$((1-0.2)/0.2)
= 0.6931

Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

108

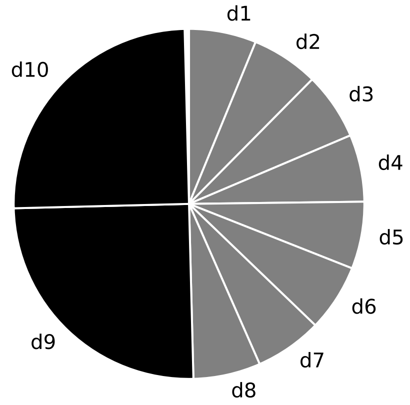
$$\mathbf{w}[1] \leftarrow 0.100 \times \left(\frac{1}{2 \times (1 - 0.200)}\right) \leftarrow 0.0625$$

$$\mathbf{w}[9] \leftarrow 0.100 \times \left(\frac{1}{2 \times 0.200}\right) \leftarrow 0.250$$

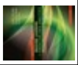
109

Example: A simple bicycle demand predictions dataset and the workings of the first three iterations of training an ensemble model using boosting to predict RENTALS given TEMP

			Iteration 1		
ID	TEMP	RENTALS	Dist.	Freq.	$\mathbb{M}_1(\mathbf{d})$
1	4	*'Low'*	0.062		
2	5	*'Low'*	0.062		
3	7	*'Low'*	0.062		
4	12	*'High'*	0.062		
5	18	*'High'*	0.062		
6	23	*'High'*	0.062		
7	27	*'High'*	0.062		
8	28	*'High'*	0.062		
9	32	*'Low'*	0.250		
10	35	*'Low'*	0.250		



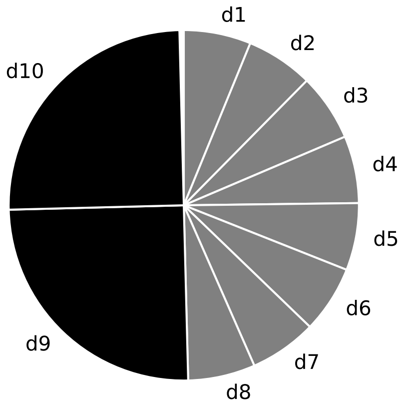
Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy



110

Example: A simple bicycle demand predictions dataset and the workings of the first three iterations of training an ensemble model using boosting to predict RENTALS given TEMP

			Iteration 1		
ID	TEMP	RENTALS	Dist.	Freq.	$\mathbb{M}_1(\mathbf{d})$
1	4	'Low'	0.062	0	
2	5	'Low'	0.062	1	
3	7	'Low'	0.062	1	
4	12	'High'	0.062	2	
5	18	'High'	0.062	0	
6	23	'High'	0.062	0	
7	27	'High'	0.062	1	
8	28	'High'	0.062	1	
9	32	'Low'	0.250	3	
10	35	'Low'	0.250	1	

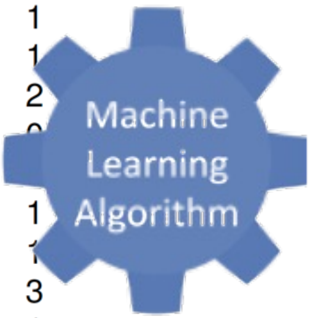


Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

111

Example: A simple bicycle demand predictions dataset and the workings of the first three iterations of training an ensemble model using boosting to predict RENTALS given TEMP

			Iteration 1		
ID	TEMP	RENTALS	Dist.	Freq.	$\mathbb{M}_1(\mathbf{d})$
1	4	'Low'	0.062	0	
2	5	'Low'	0.062	1	
3	7	'Low'	0.062	1	
4	12	'High'	0.062	2	
5	18	'High'	0.062	[illegible]	
6	23	'High'	0.062	[illegible]	
7	27	'High'	0.062	1	
8	28	'High'	0.062	1	
9	32	'Low'	0.250	3	
10	35	'Low'	0.250	1	

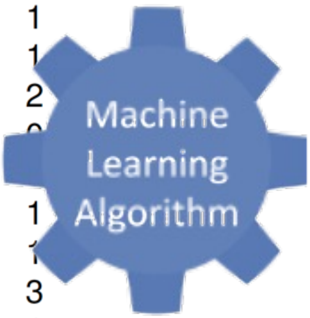


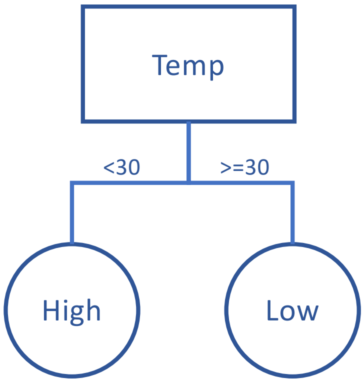
Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

112

Example: A simple bicycle demand predictions dataset and the workings of the first three iterations of training an ensemble model using boosting to predict RENTALS given TEMP

			Iteration 1		
ID	TEMP	RENTALS	Dist.	Freq.	$\mathbb{M}_1(\mathbf{d})$
1	4	'Low'	0.062	0	
2	5	'Low'	0.062	1	
3	7	'Low'	0.062	1	
4	12	'High'	0.062	2	
5	18	'High'	0.062	[illegible]	
6	23	'High'	0.062	[illegible]	
7	27	'High'	0.062	1	
8	28	'High'	0.062	1	
9	32	'Low'	0.250	3	
10	35	'Low'	0.250	1	



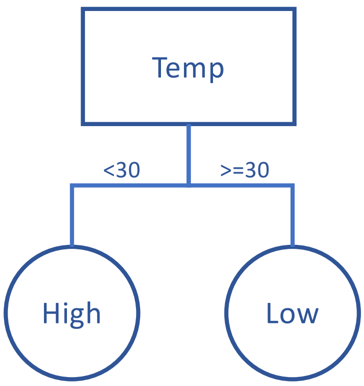


Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

113

Example: A simple bicycle demand predictions dataset and the workings of the first three iterations of training an ensemble model using boosting to predict RENTALS given TEMP

			Iteration 1		
ID	TEMP	RENTALS	Dist.	Freq.	$\mathbb{M}_1(\mathbf{d})$
1	4	'Low'	0.062	0	'High'
2	5	'Low'	0.062	1	'High'
3	7	'Low'	0.062	1	'High'
4	12	'High'	0.062	2	'High'
5	18	'High'	0.062	0	'High'
6	23	'High'	0.062	0	'High'
7	27	'High'	0.062	1	'High'
8	28	'High'	0.062	1	'High'
9	32	'Low'	0.250	3	'Low'
10	35	'Low'	0.250	1	'Low'



Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

114

Example: A simple bicycle demand predictions dataset and the workings of the first three iterations of training an ensemble model using boosting to predict RENTALS given TEMP

			Iteration 1		
ID	TEMP	RENTALS	Dist.	Freq.	$\mathbb{M}_1(\mathbf{d})$
1	4	*'Low'*	0.062	0	*'High'*
2	5	*'Low'*	0.062	1	*'High'*
3	7	*'Low'*	0.062	1	*'High'*
4	12	*'High'*	0.062	2	*'High'*
5	18	*'High'*	0.062	0	*'High'*
6	23	*'High'*	0.062	0	*'High'*
7	27	*'High'*	0.062	1	*'High'*
8	28	*'High'*	0.062	1	*'High'*
9	32	*'Low'*	0.250	3	*'Low'*
10	35	*'Low'*	0.250	1	*'Low'*

ϵ = (0.062 + 0.062 + 0.062)
= 0.186

115

Example: A simple bicycle demand predictions dataset and the workings of the first three iterations of training an ensemble model using boosting to predict RENTALS given TEMP

			Iteration 1		
ID	TEMP	RENTALS	Dist.	Freq.	$\mathbb{M}_1(\mathbf{d})$
1	4	*'Low'*	0.062	0	*'High'*
2	5	*'Low'*	0.062	1	*'High'*
3	7	*'Low'*	0.062	1	*'High'*
4	12	*'High'*	0.062	2	*'High'*
5	18	*'High'*	0.062	0	*'High'*
6	23	*'High'*	0.062	0	*'High'*
7	27	*'High'*	0.062	1	*'High'*
8	28	*'High'*	0.062	1	*'High'*
9	32	*'Low'*	0.250	3	*'Low'*
10	35	*'Low'*	0.250	1	*'Low'*

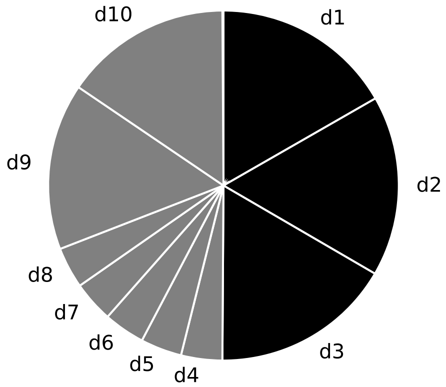
ϵ = (0.062 + 0.062 + 0.062)
= 0.186

α = 0.5 * $\log_e$((1-0.186)/0.186)
= 0.7381

116

Example: A simple bicycle demand predictions dataset and the workings of the first three iterations of training an ensemble model using boosting to predict RENTALS given TEMP

			Iteration 2		
ID	TEMP	RENTALS	Dist.	Freq.	$\mathbb{M}_2(\mathbf{d})$
1	4	'Low'	0.167		
2	5	'Low'	0.167		
3	7	'Low'	0.167		
4	12	'High'	0.038		
5	18	'High'	0.038		
6	23	'High'	0.038		
7	27	'High'	0.038		
8	28	'High'	0.038		
9	32	'Low'	0.154		
10	35	'Low'	0.154		

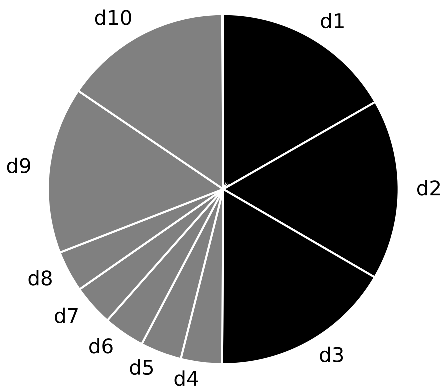


Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

117

Example: A simple bicycle demand predictions dataset and the workings of the first three iterations of training an ensemble model using boosting to predict RENTALS given TEMP

			Iteration 2		
ID	TEMP	RENTALS	Dist.	Freq.	$\mathbb{M}_2(\mathbf{d})$
1	4	'Low'	0.167	2	
2	5	'Low'	0.167	1	
3	7	'Low'	0.167	3	
4	12	'High'	0.038	0	
5	18	'High'	0.038	0	
6	23	'High'	0.038	0	
7	27	'High'	0.038	0	
8	28	'High'	0.038	1	
9	32	'Low'	0.154	1	
10	35	'Low'	0.154	2	

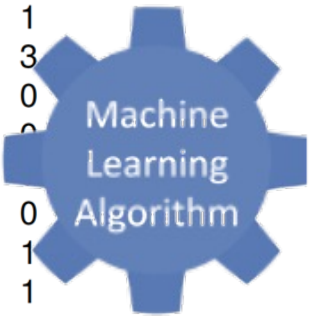


Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

118

Example: A simple bicycle demand predictions dataset and the workings of the first three iterations of training an ensemble model using boosting to predict RENTALS given TEMP

			Iteration 2		
ID	TEMP	RENTALS	Dist.	Freq.	$\mathbb{M}_2(\mathbf{d})$
1	4	'Low'	0.167	2	
2	5	'Low'	0.167	1	
3	7	'Low'	0.167	3	
4	12	'High'	0.038	0	
5	18	'High'	0.038	0	
6	23	'High'	0.038	[illegible]	
7	27	'High'	0.038	0	
8	28	'High'	0.038	1	
9	32	'Low'	0.154	1	
10	35	'Low'	0.154	2	

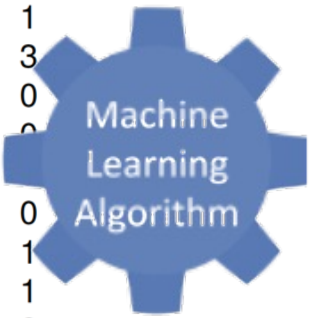


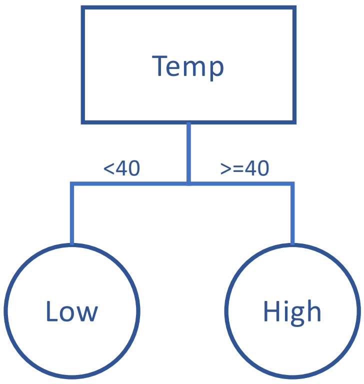
Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

119

Example: A simple bicycle demand predictions dataset and the workings of the first three iterations of training an ensemble model using boosting to predict RENTALS given TEMP

			Iteration 2		
ID	TEMP	RENTALS	Dist.	Freq.	$\mathbb{M}_2(\mathbf{d})$
1	4	'Low'	0.167	2	
2	5	'Low'	0.167	1	
3	7	'Low'	0.167	3	
4	12	'High'	0.038	0	
5	18	'High'	0.038	0	
6	23	'High'	0.038	[illegible]	
7	27	'High'	0.038	0	
8	28	'High'	0.038	1	
9	32	'Low'	0.154	1	
10	35	'Low'	0.154	2	



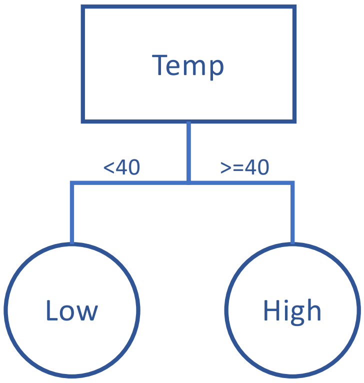


Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

120

Example: A simple bicycle demand predictions dataset and the workings of the first three iterations of training an ensemble model using boosting to predict RENTALS given TEMP

			Iteration 2		
ID	TEMP	RENTALS	Dist.	Freq.	$\mathbb{M}_2(\mathbf{d})$
1	4	'Low'	0.167	2	'Low'
2	5	'Low'	0.167	1	'Low'
3	7	'Low'	0.167	3	'Low'
4	12	'High'	0.038	0	'Low'
5	18	'High'	0.038	0	'Low'
6	23	'High'	0.038	0	'Low'
7	27	'High'	0.038	0	'Low'
8	28	'High'	0.038	1	'Low'
9	32	'Low'	0.154	1	'Low'
10	35	'Low'	0.154	2	'Low'



Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies (www.machinelearningbook.com)
John D. Kelleher, Brian Mac Namee and Aoife D'Arcy

Example: A simple bicycle demand predictions dataset and the workings of the first three iterations of training an ensemble model using boosting to predict RENTALS given TEMP

			Iteration 2		
ID	TEMP	RENTALS	Dist.	Freq.	$\mathbb{M}_2(\mathbf{d})$
1	4	'Low'	0.167	2	'Low'
2	5	'Low'	0.167	1	'Low'
3	7	'Low'	0.167	3	'Low'
4	12	'High'	0.038	0	'Low'
5	18	'High'	0.038	0	'Low'
6	23	'High'	0.038	0	'Low'
7	27	'High'	0.038	0	'Low'
8	28	'High'	0.038	1	'Low'
9	32	'Low'	0.154	1	'Low'
10	35	'Low'	0.154	2	'Low'

$$\epsilon = (0.038 + 0.038 + 0.038 + 0.038 + 0.038) = 0.19$$

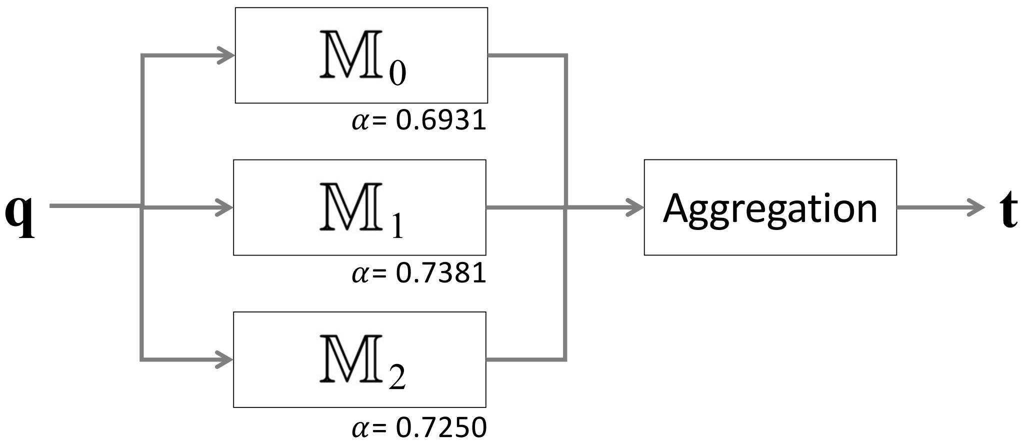
$$\alpha = 0.5 * \log_e((1-0.19)/0.19) = 0.7250$$

Example: A simple bicycle demand predictions dataset and the workings of the first three iterations of training an ensemble model using boosting to predict RENTALS given TEMP

			Iteration 2		
ID	TEMP	RENTALS	Dist.	Freq.	$\mathbb{M}_2(\mathbf{d})$
1	4	'Low'	0.167	2	'Low'
2	5	'Low'	0.167	1	'Low'
3	7	'Low'	0.167	3	'Low'
4	12	'High'	0.038	0	'Low'
5	18	'High'	0.038	0	'Low'
6	23	'High'	0.038	0	'Low'
7	27	'High'	0.038	0	'Low'
8	28	'High'	0.038	1	'Low'
9	32	'Low'	0.154	1	'Low'
10	35	'Low'	0.154	2	'Low'

$$\epsilon = (0.038 + 0.038 + 0.038 + 0.038 + 0.038) = 0.19$$

123



124

Boosting

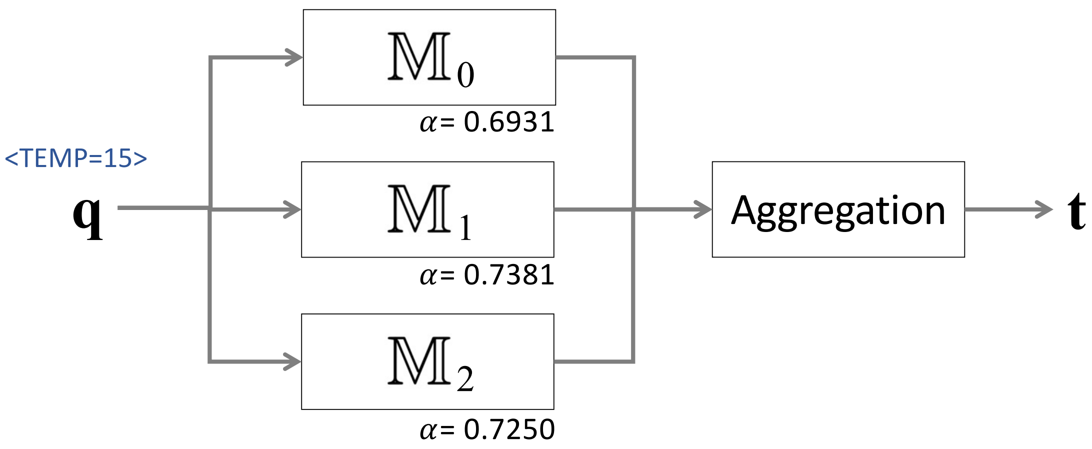
Predictions are made using a weighted aggregate of the individual models

- Weights are based on confidence factors

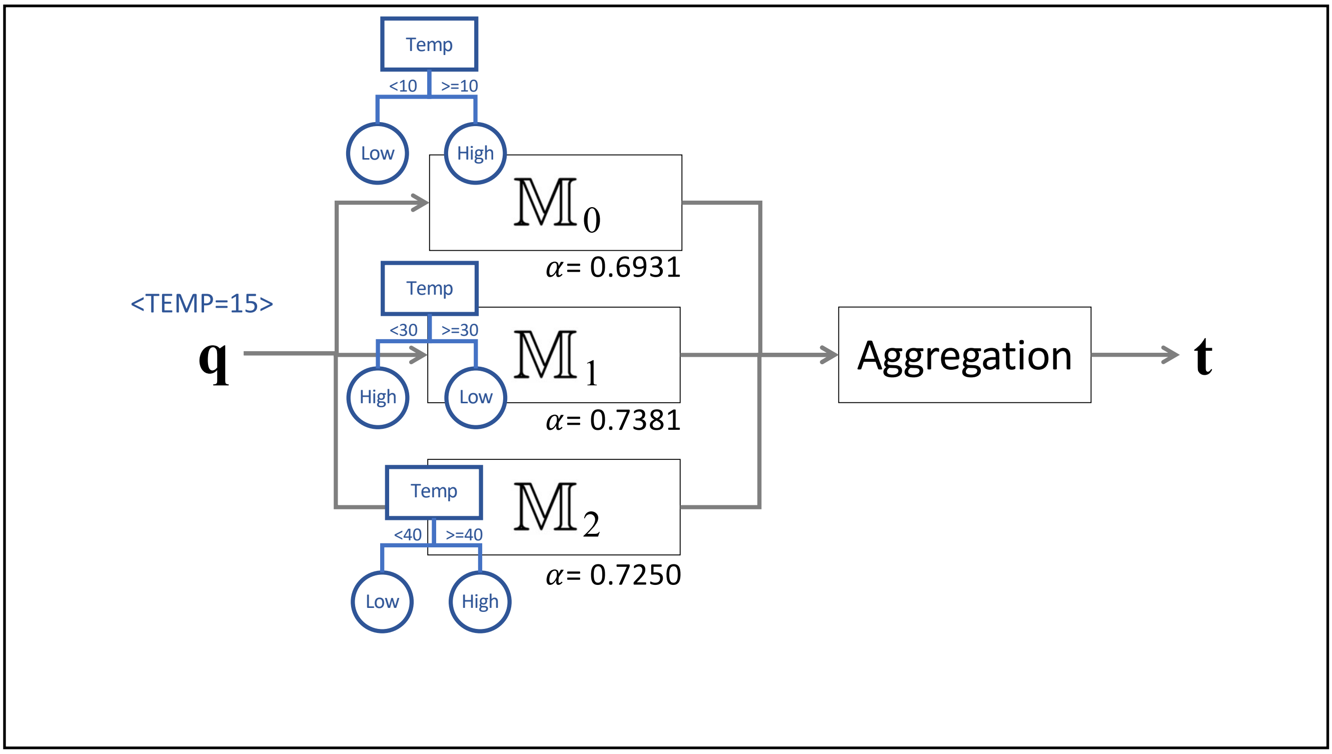
$$t = sign\left(\sum_{\mathbb{M}_i \in \mathbb{M}} \alpha_i \mathbb{M}_i(q)\right)$$

- Assumes binary outputs of +1 or -1

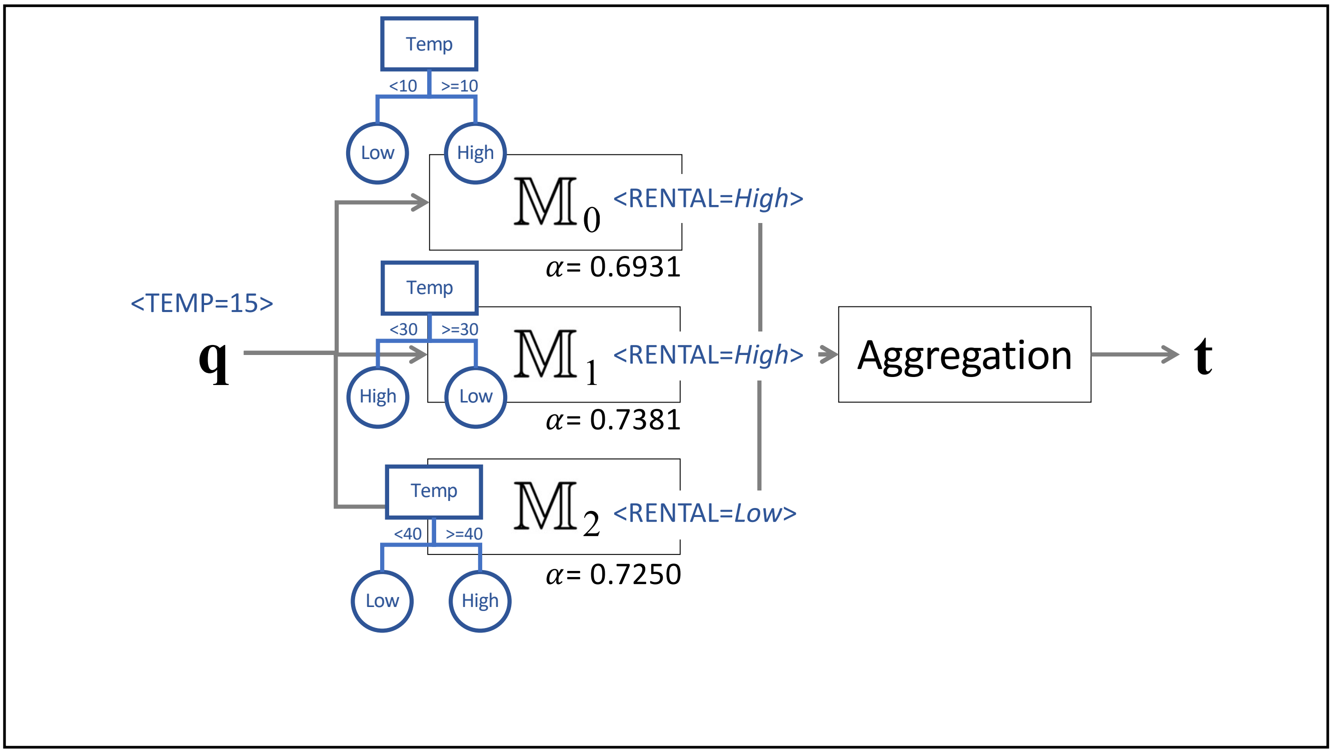
125



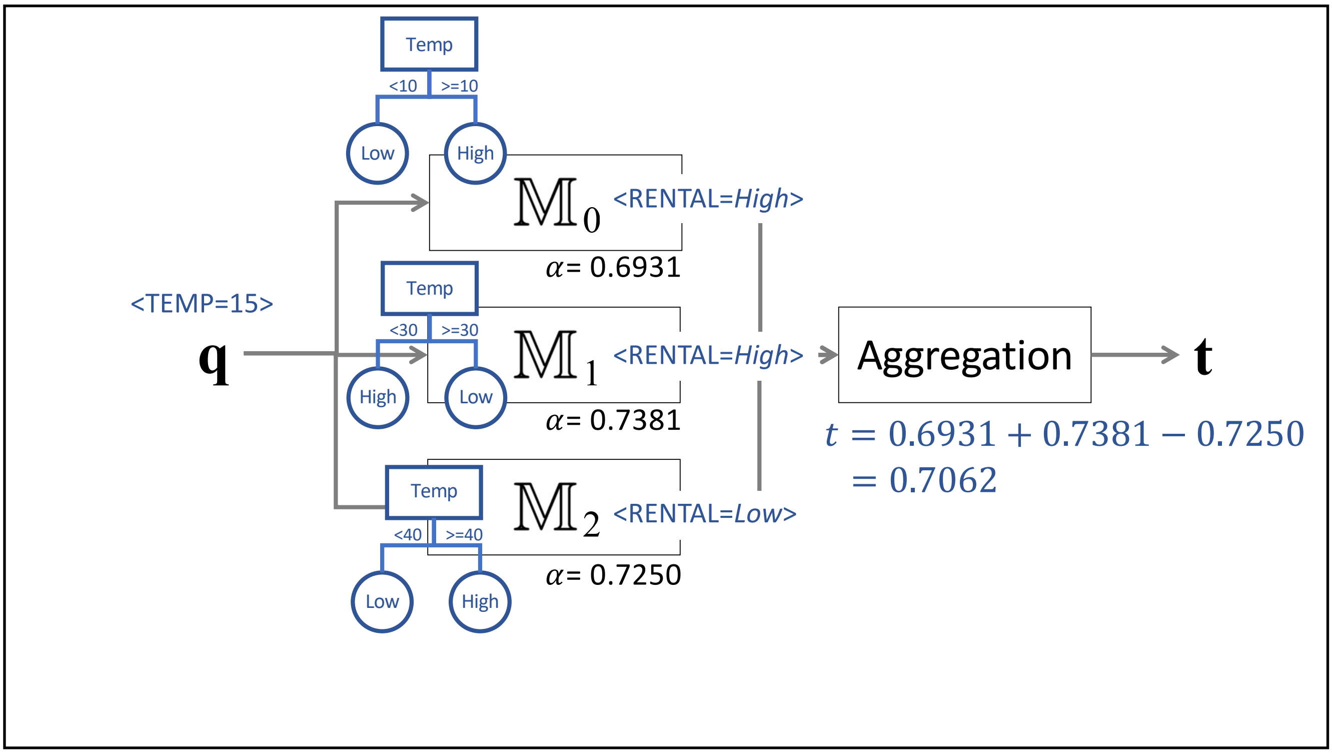
126


Temp
<10
>=10
Low
High
M0
α= 0.6931
<TEMP=15>
q
Temp
<30
>=30
High
Low
M1
α= 0.7381
Aggregation
t
Temp
<40
>=40
Low
High
M2
α= 0.7250

127

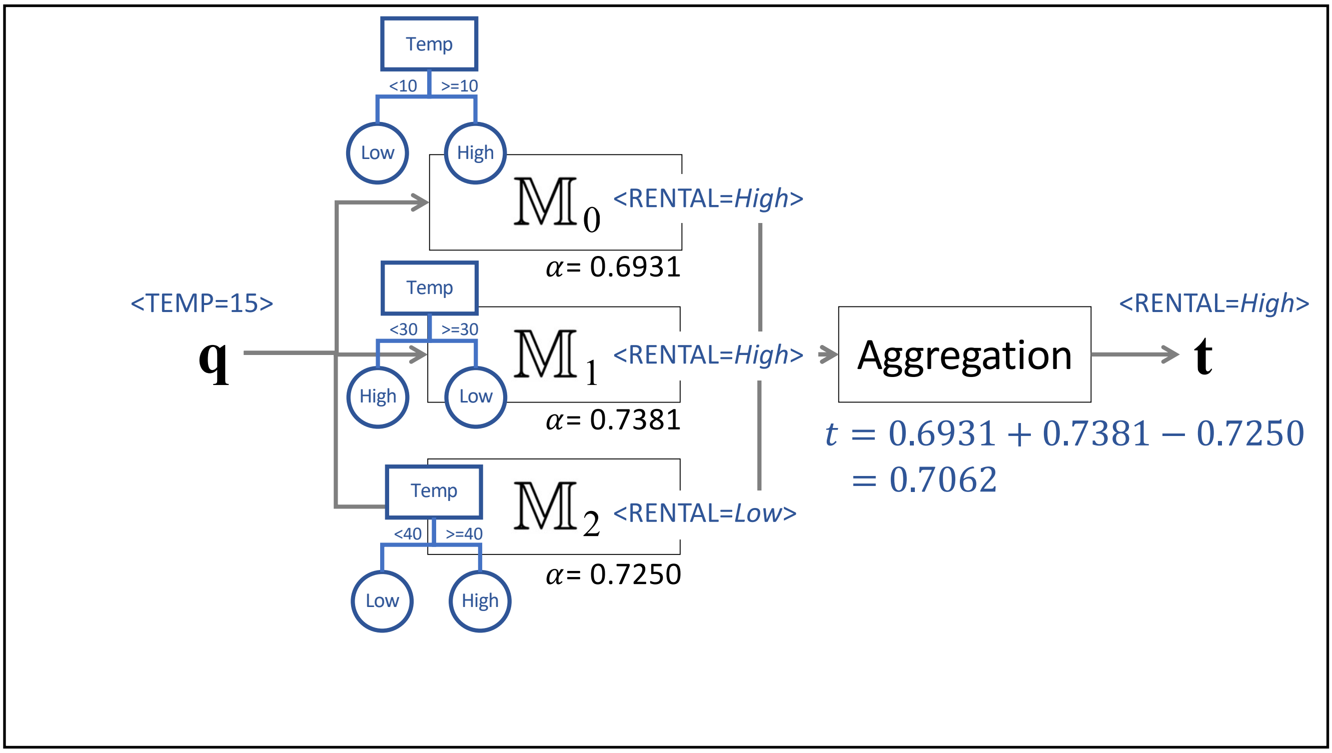

Temp
<10
>=10
Low
High
M0
<RENTAL=High>
α= 0.6931
<TEMP=15>
q
Temp
<30
>=30
High
Low
M1
<RENTAL=High>
α= 0.7381
Aggregation
t
Temp
<40
>=40
Low
High
M2
<RENTAL=Low>
α= 0.7250

128



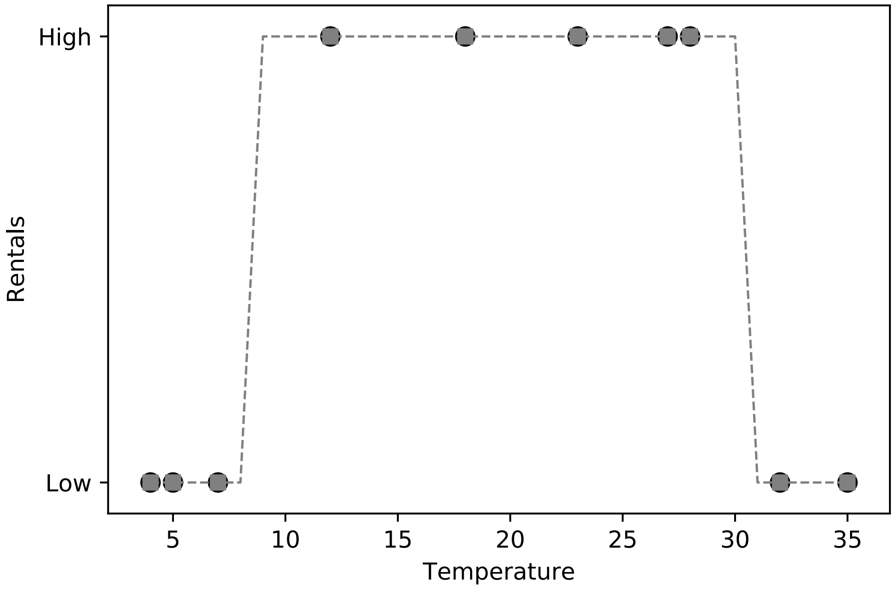
Temp
<10
>=10
Low
High
M0
<RENTAL=High>
α= 0.6931
Temp
<30
>=30
High
Low
M1
<RENTAL=High>
α= 0.7381
Temp
<40
>=40
Low
High
M2
<RENTAL=Low>
α= 0.7250
<TEMP=15>
q
Aggregation
t
t = 0.6931 + 0.7381 – 0.7250
= 0.7062

129



Temp
<10
>=10
Low
High
M0
<RENTAL=High>
α= 0.6931
Temp
<30
>=30
High
Low
M1
<RENTAL=High>
α= 0.7381
Temp
<40
>=40
Low
High
M2
<RENTAL=Low>
α= 0.7250
<TEMP=15>
q
Aggregation
<RENTAL=High>
t
t = 0.6931 + 0.7381 – 0.7250
= 0.7062

130



131

STACKING (OPTIONAL)

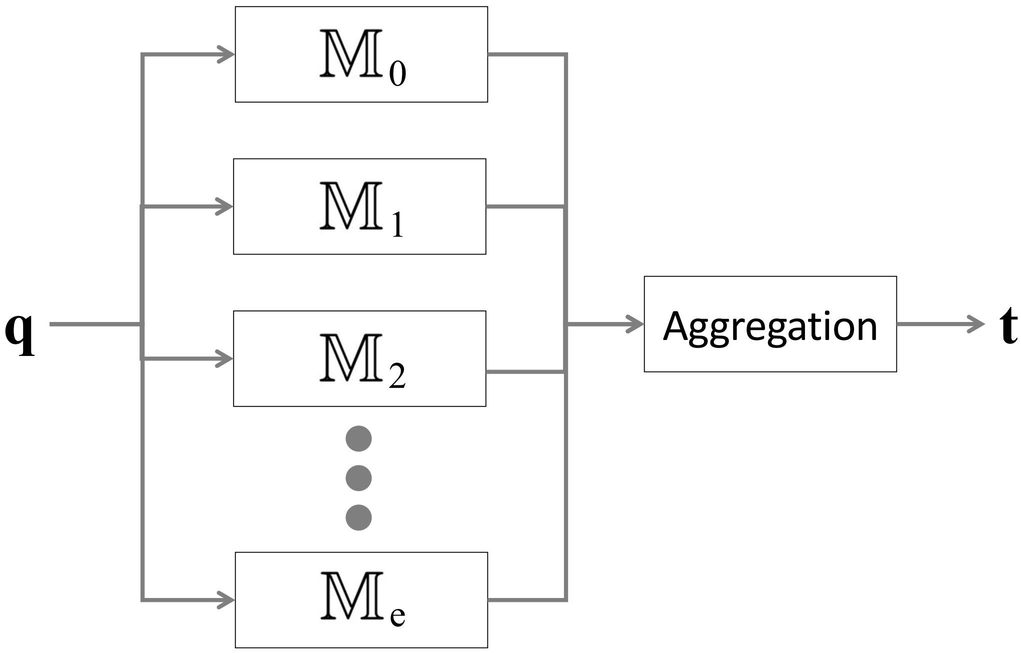
132

Stacking

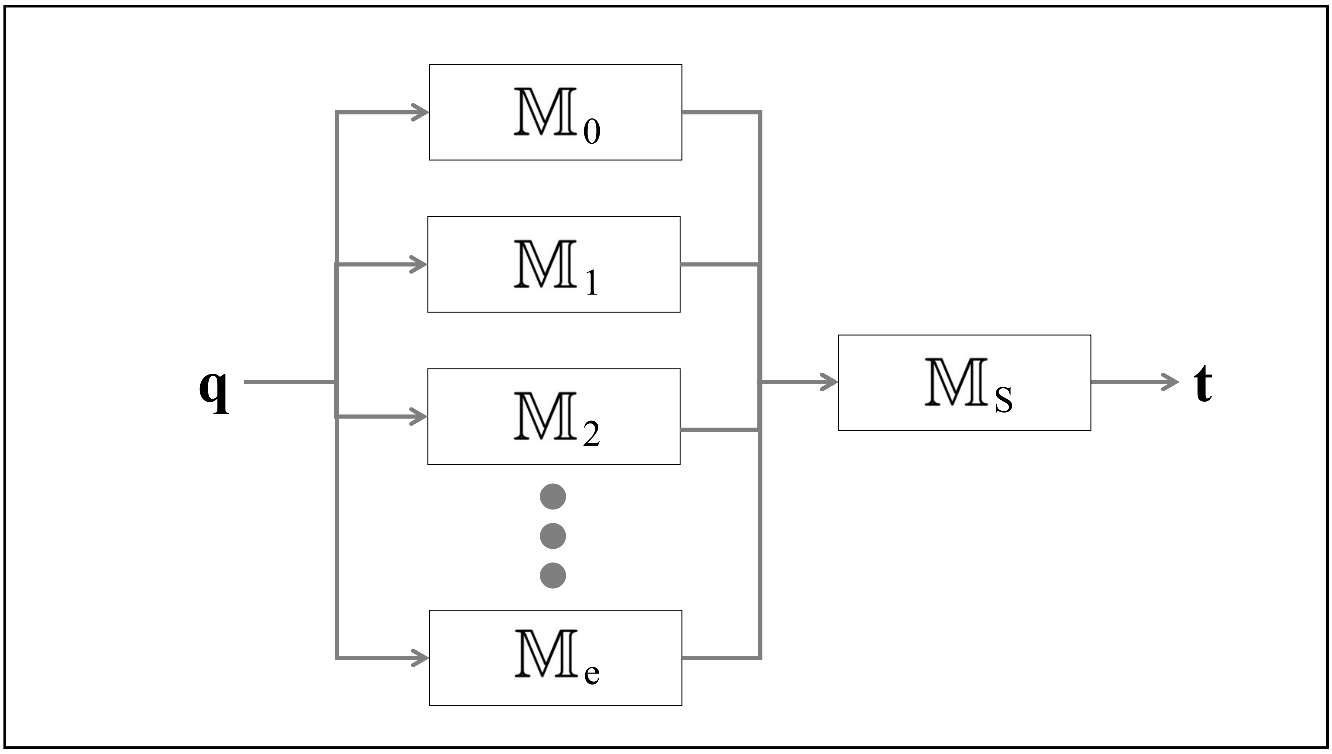
Stacking ensembles use a machine learning model to combine the outputs of the base models in an ensemble

- Can be more effective than simple majority voting or weighted voting
- Requires new datasets to be generated

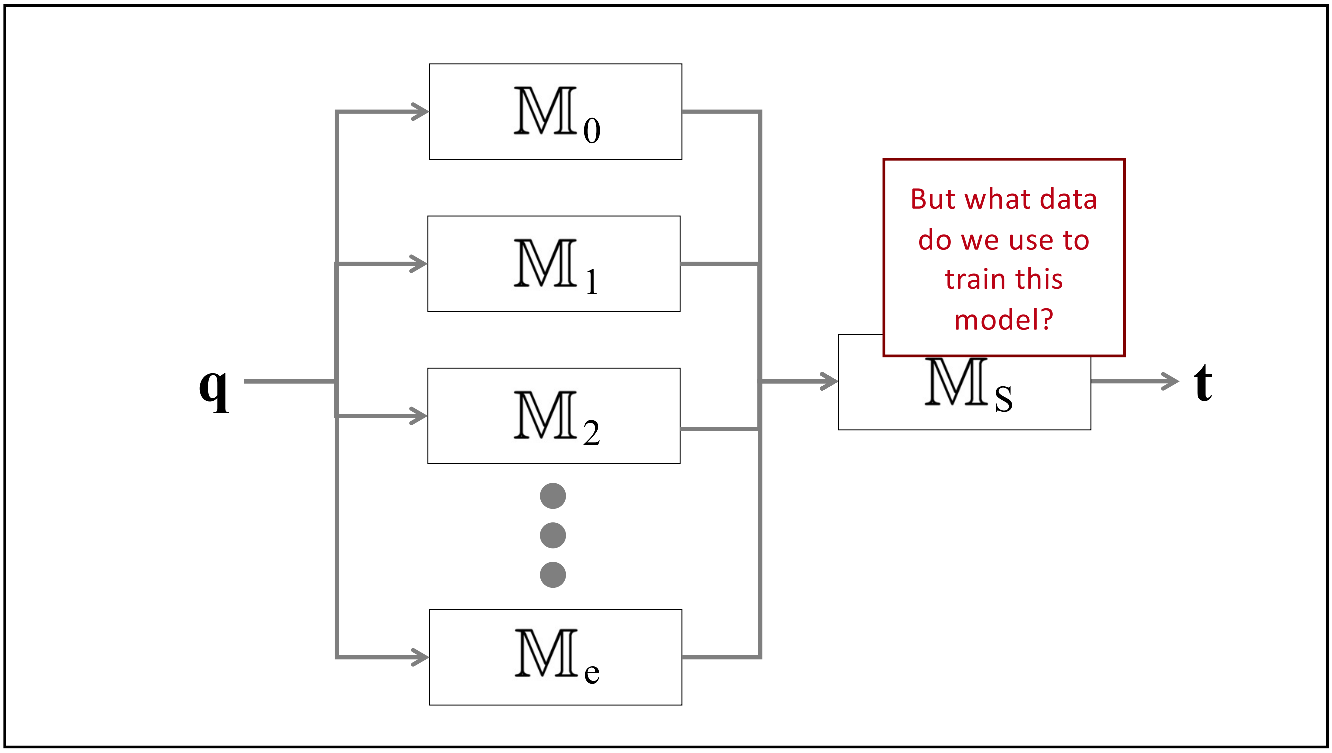
133



134


q
M_0
M_1
M_2
M_e
M_S
t

135


q
M_0
M_1
M_2
M_e
But what data do we use to train this model?
M_S
t

136

	$\mathbb{M}_0$	$\mathbb{M}_1$	$\mathbb{M}_2$	$\mathbb{M}_3$	•••	$\mathbb{M}_e$	Target
d_0	True	False	True	True		False	True
d_1	False	False	False	False		True	False
				⋮			
d_n	True	True	True	False		False	False

137

	$\mathbb{M}_0$	$\mathbb{M}_1$	$\mathbb{M}_2$	$\mathbb{M}_3$	•••	$\mathbb{M}_e$	Target
d_0	0.81	0.22	0.76	0.91		0.11	True
d_1	0.38	0.41	0.29	0.38		0.55	False
				⋮			
d_n	0.99	0.76	0.54	0.44		0.38	False

138

Stacking

If exactly the same data used to train the base learners is also used to train the stacking model there is a serious risk of overfitting

Common to use a k-fold cross validation scheme to gerneate the stacked level training set

139

$\mathbb{M}_0$

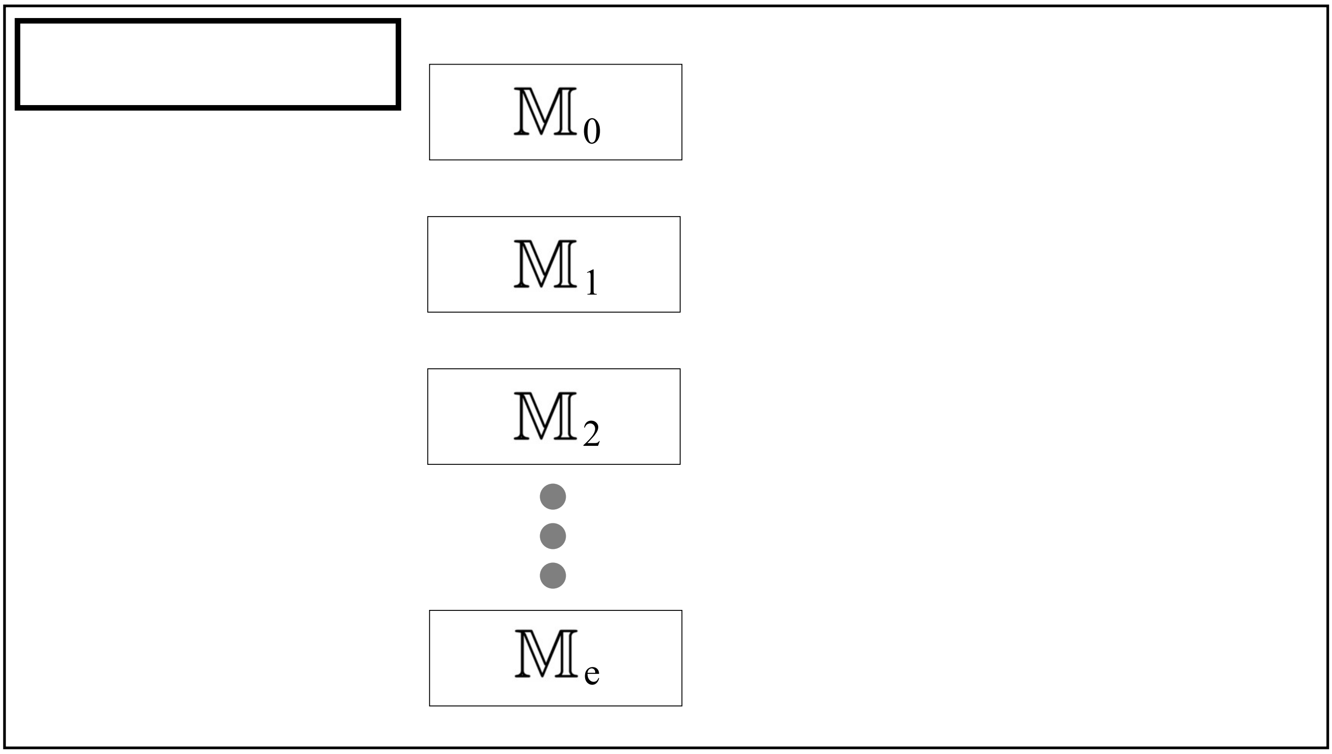
$\mathbb{M}_1$

$\mathbb{M}_2$

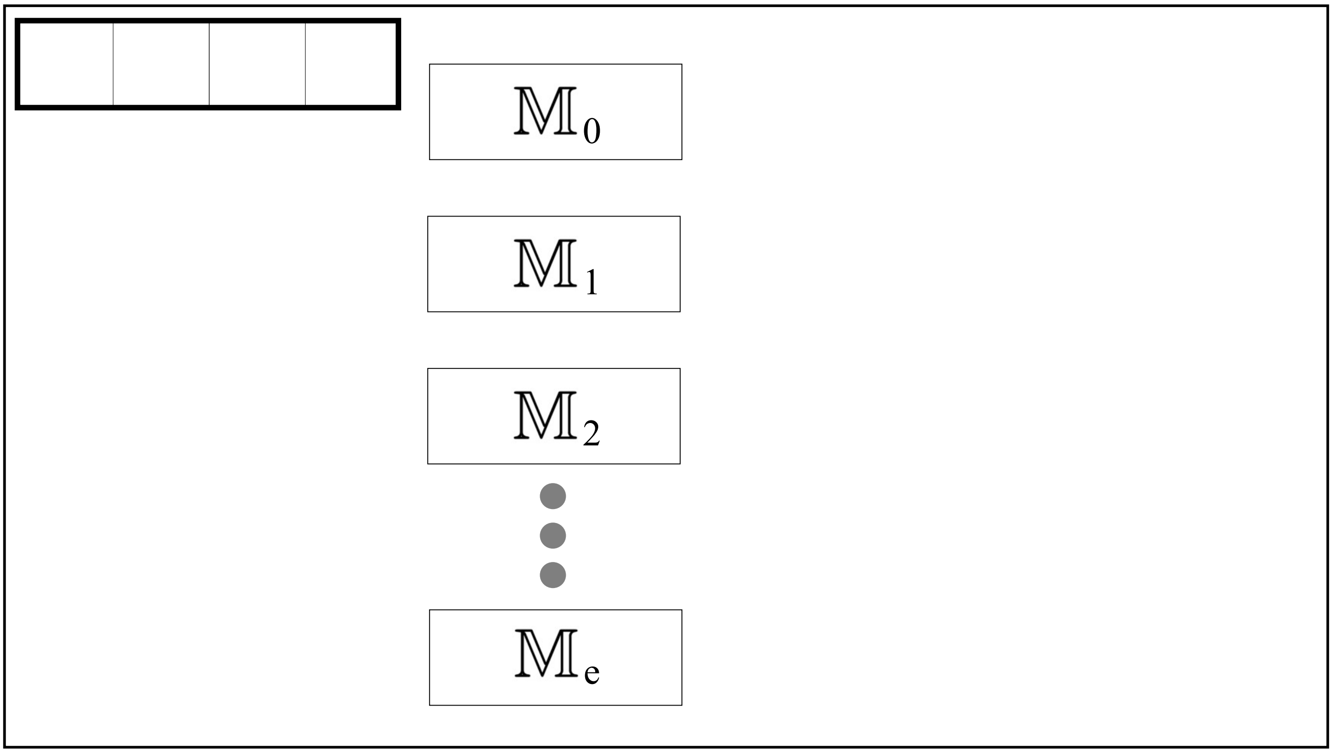
$\vdots$

$\mathbb{M}_e$

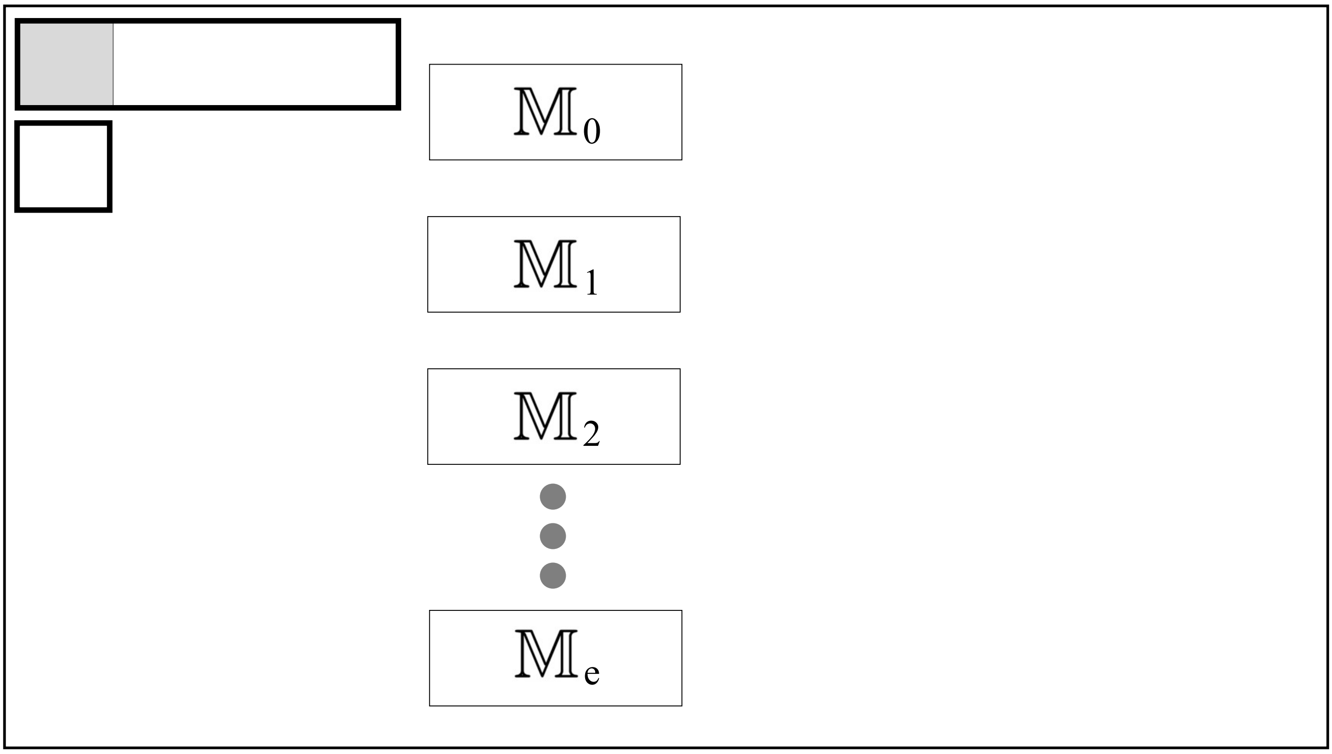
140


$\mathbb{M}_0$
$\mathbb{M}_1$
$\mathbb{M}_2$
$\mathbb{M}_e$

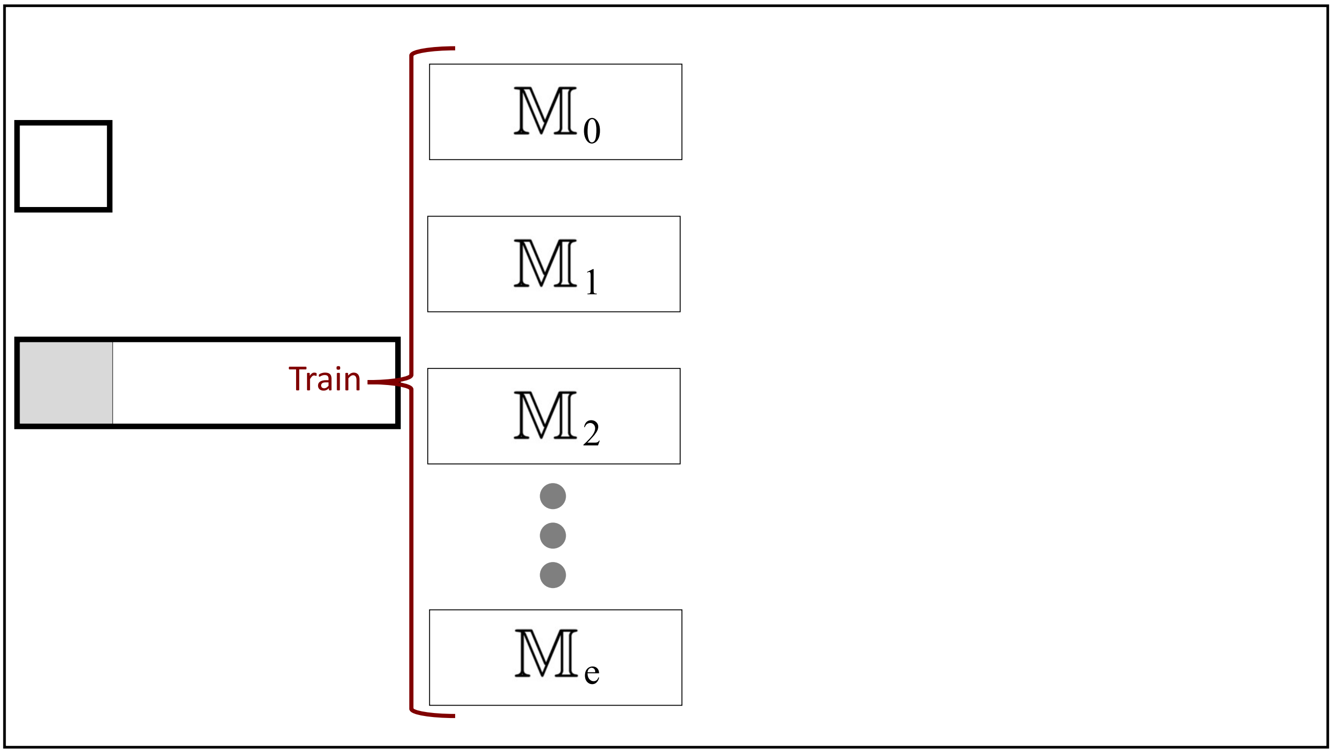
141


$\mathbb{M}_0$
$\mathbb{M}_1$
$\mathbb{M}_2$
$\mathbb{M}_e$

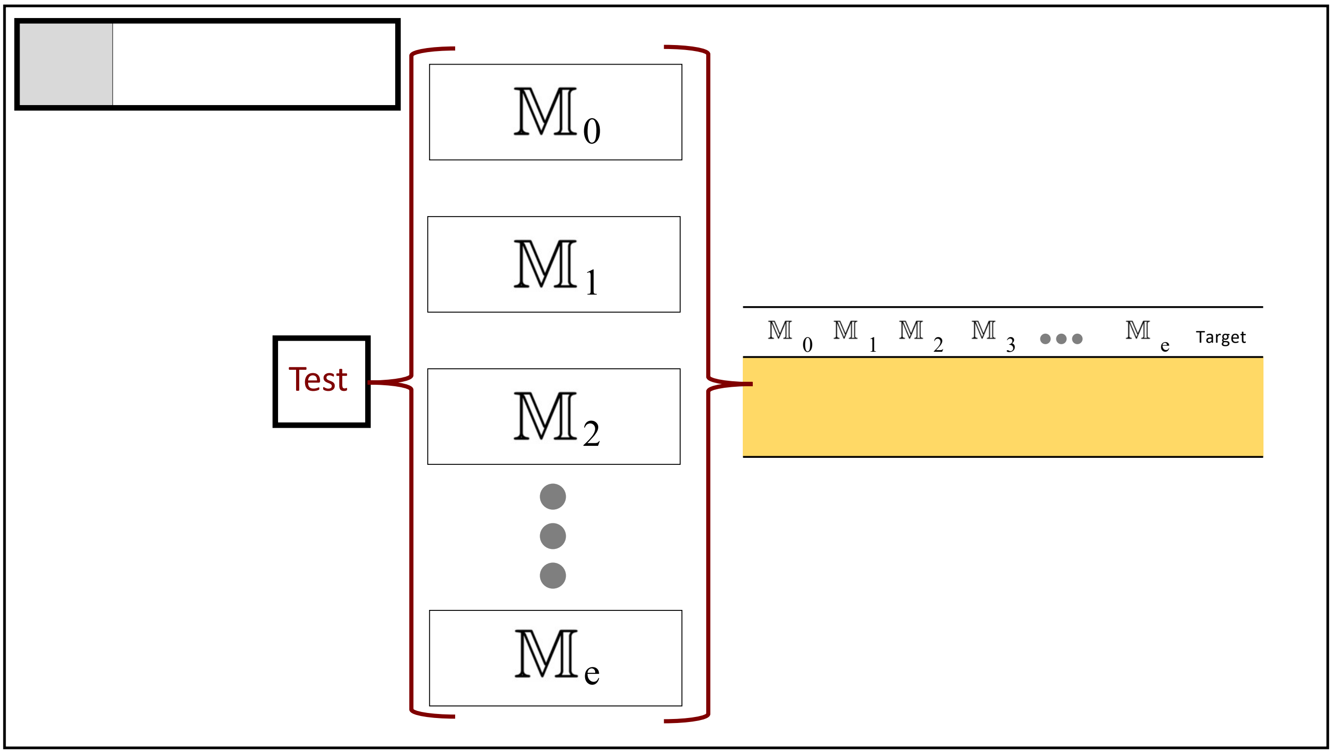
142



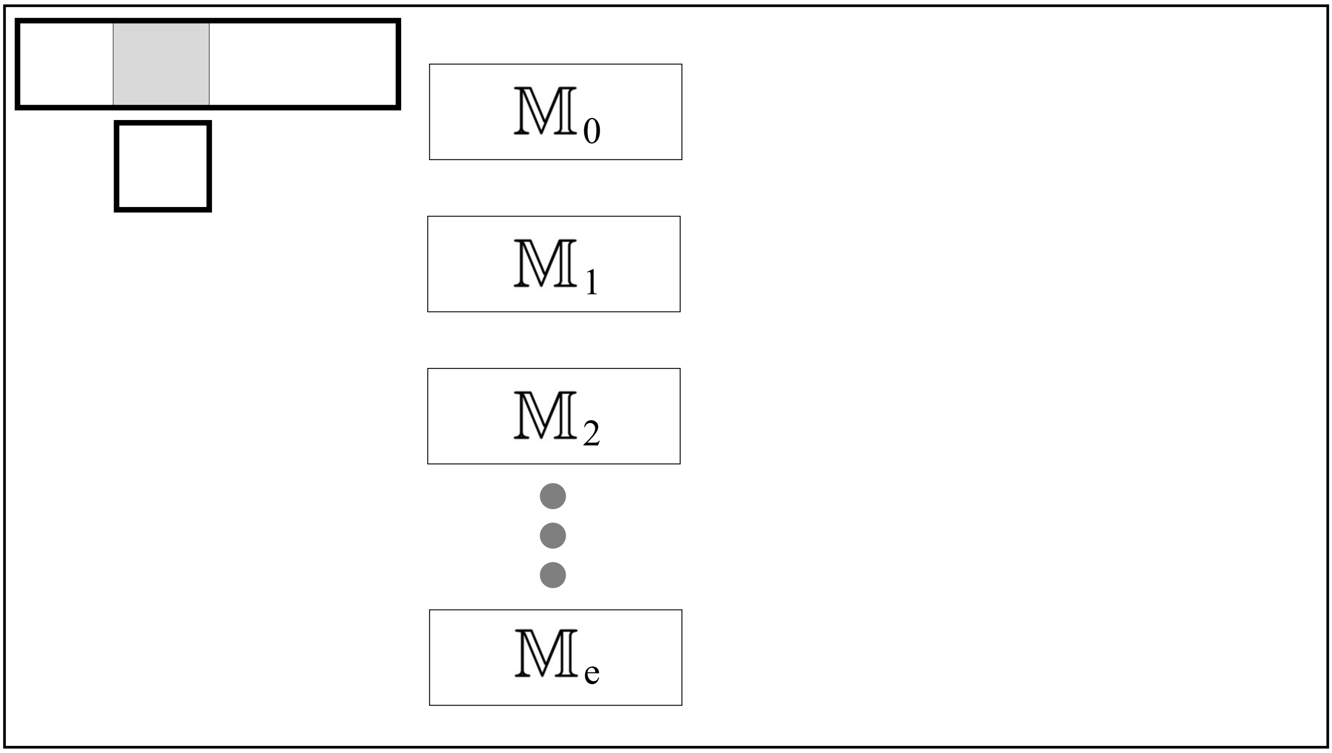
143



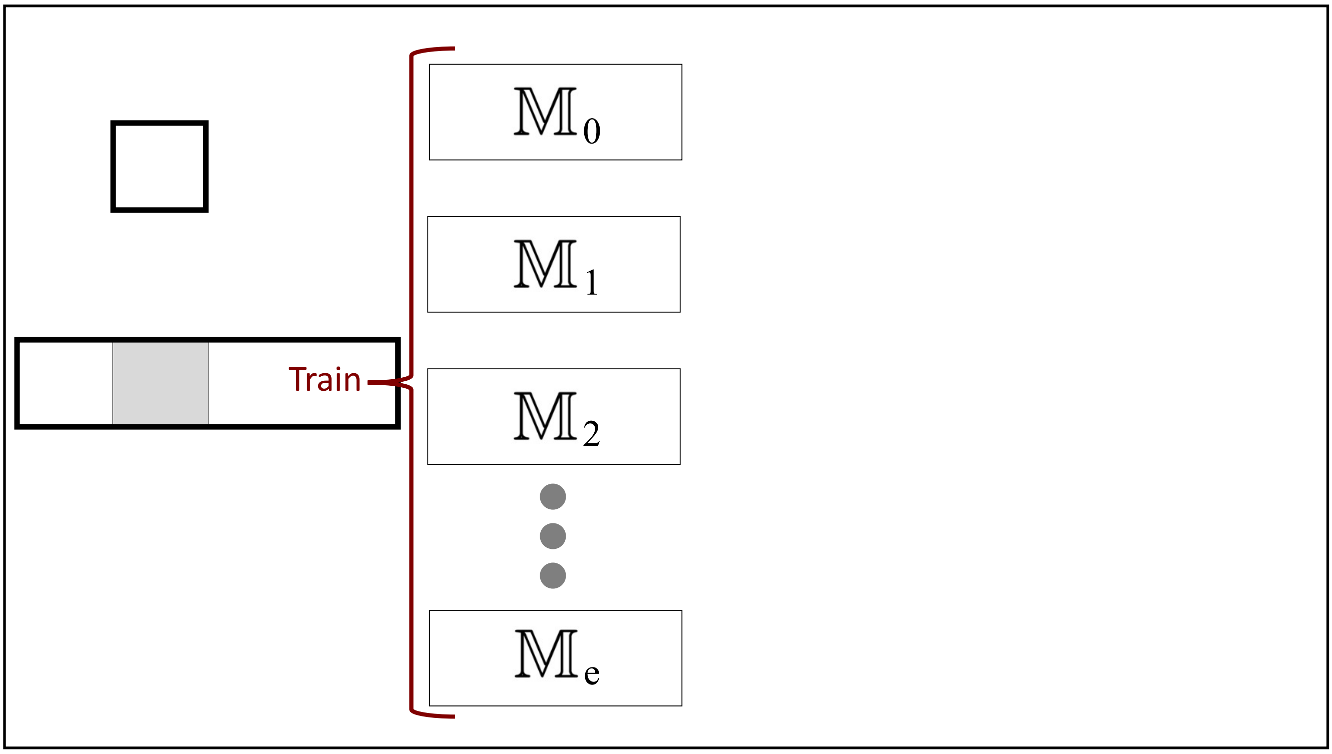
144



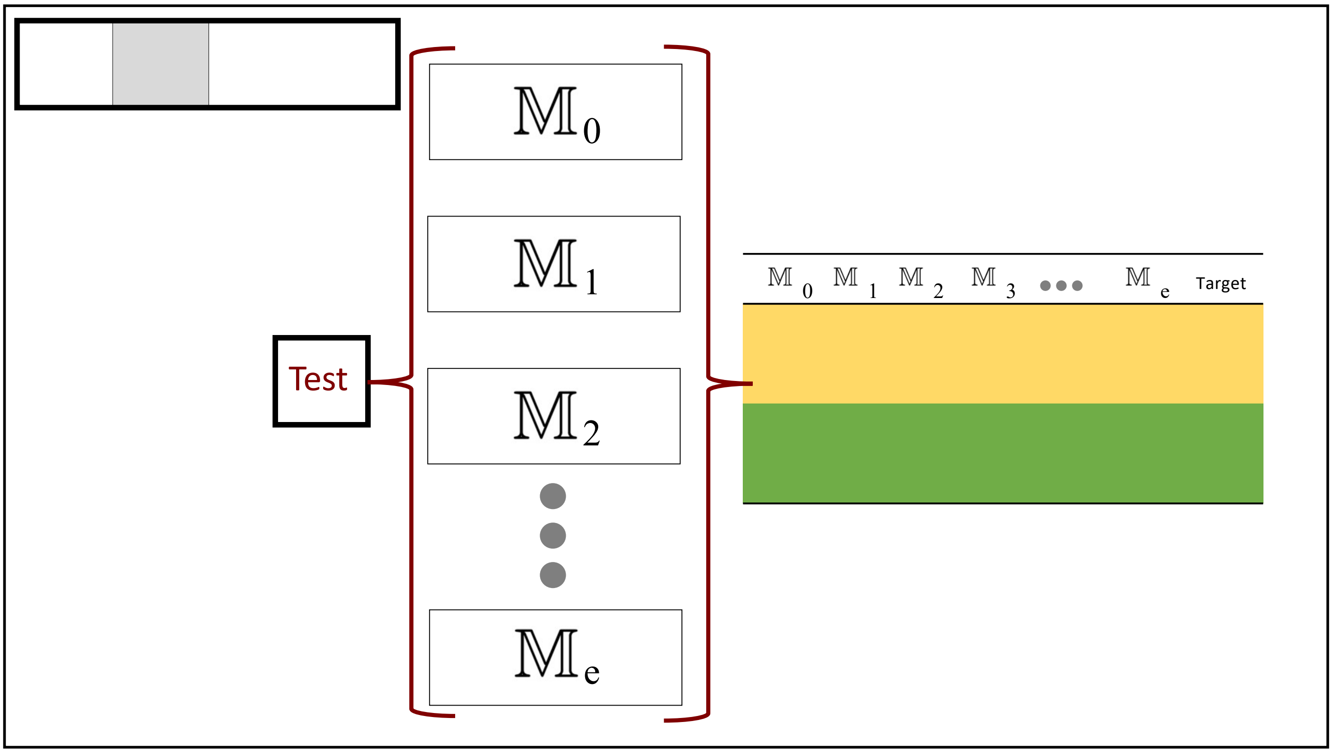
145



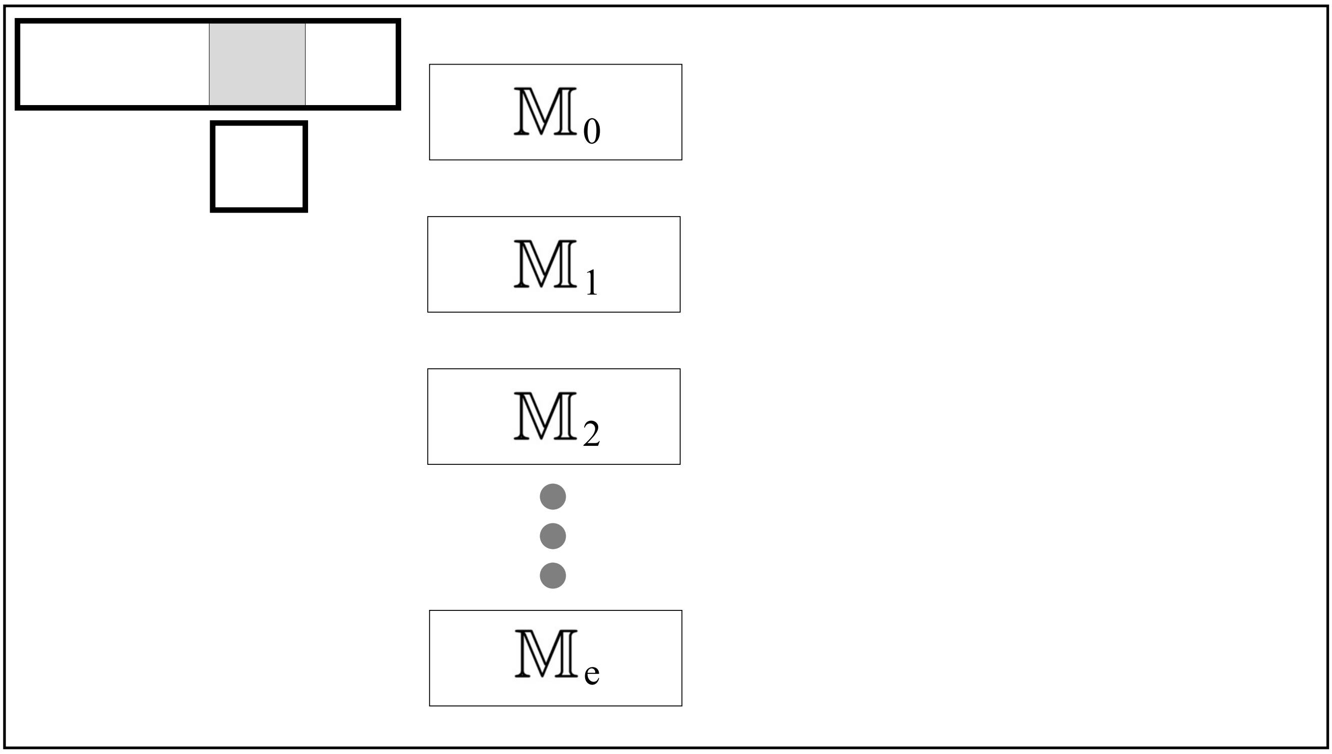
146



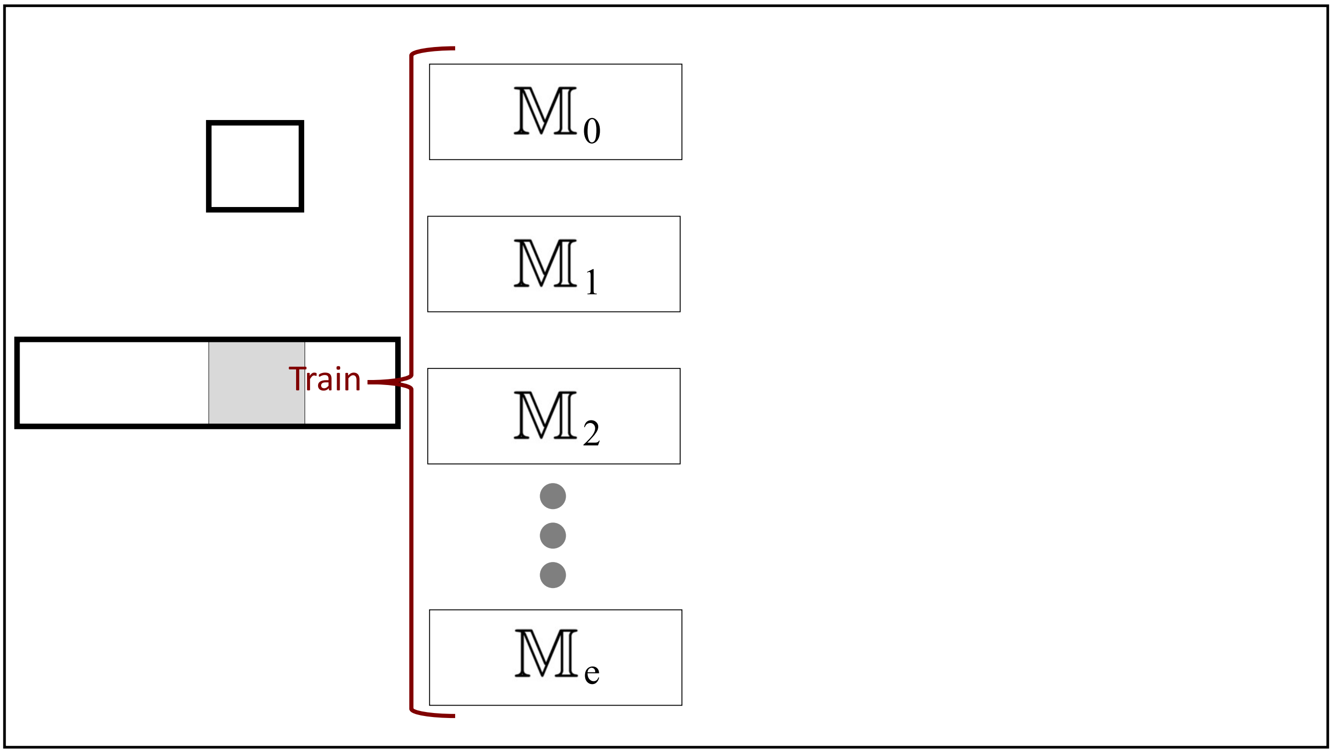
147



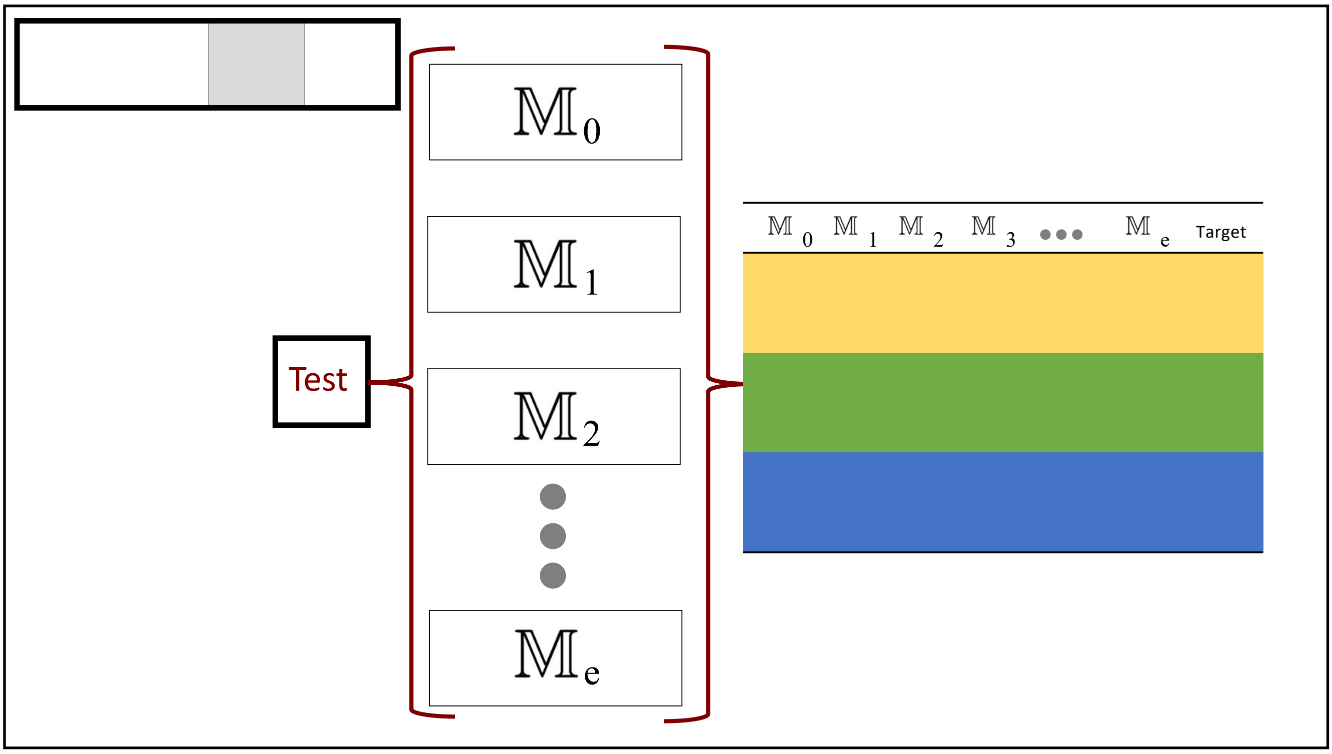
148



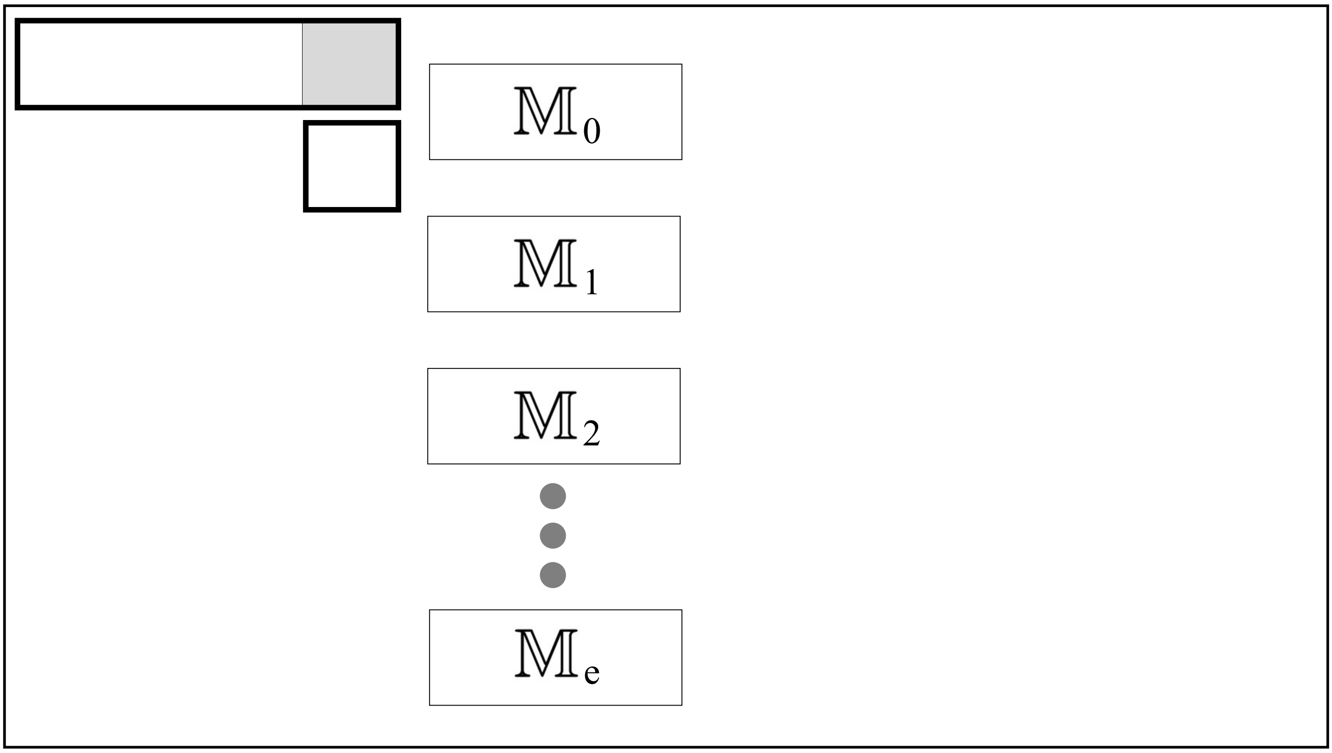
149



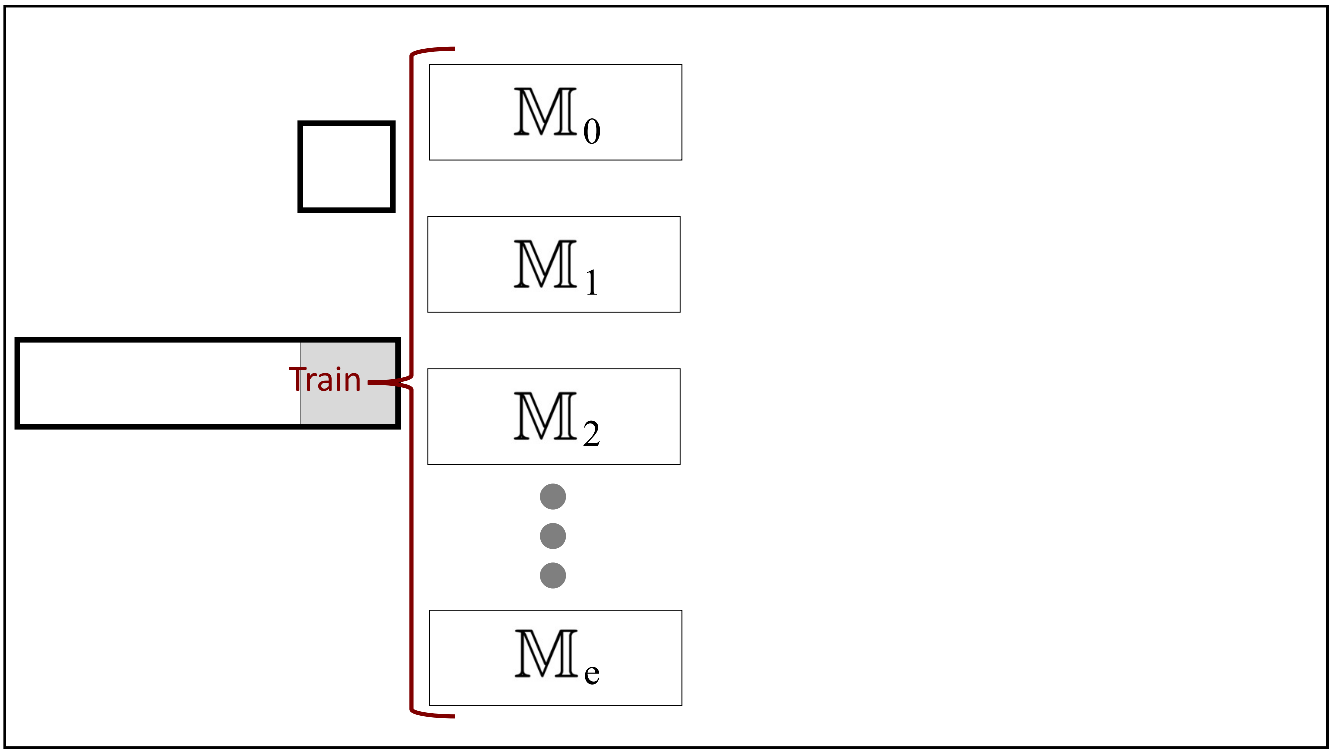
150



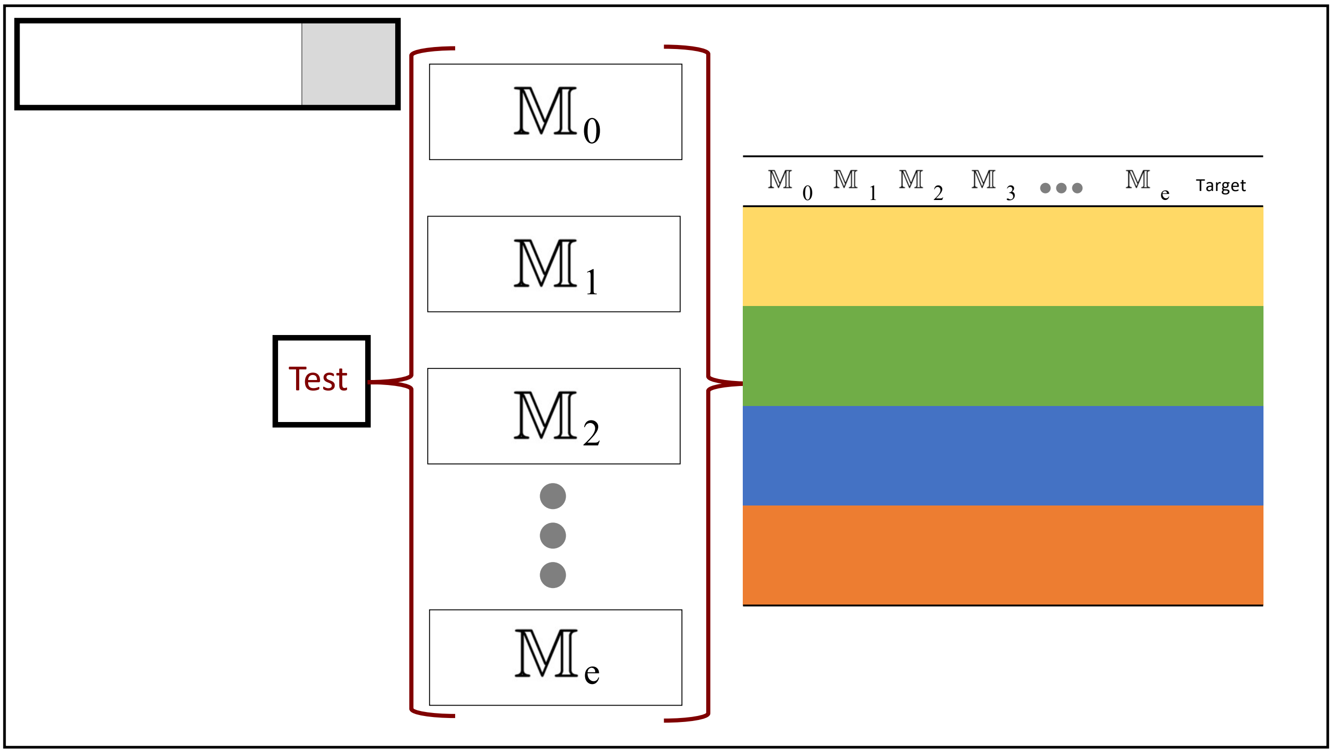
151



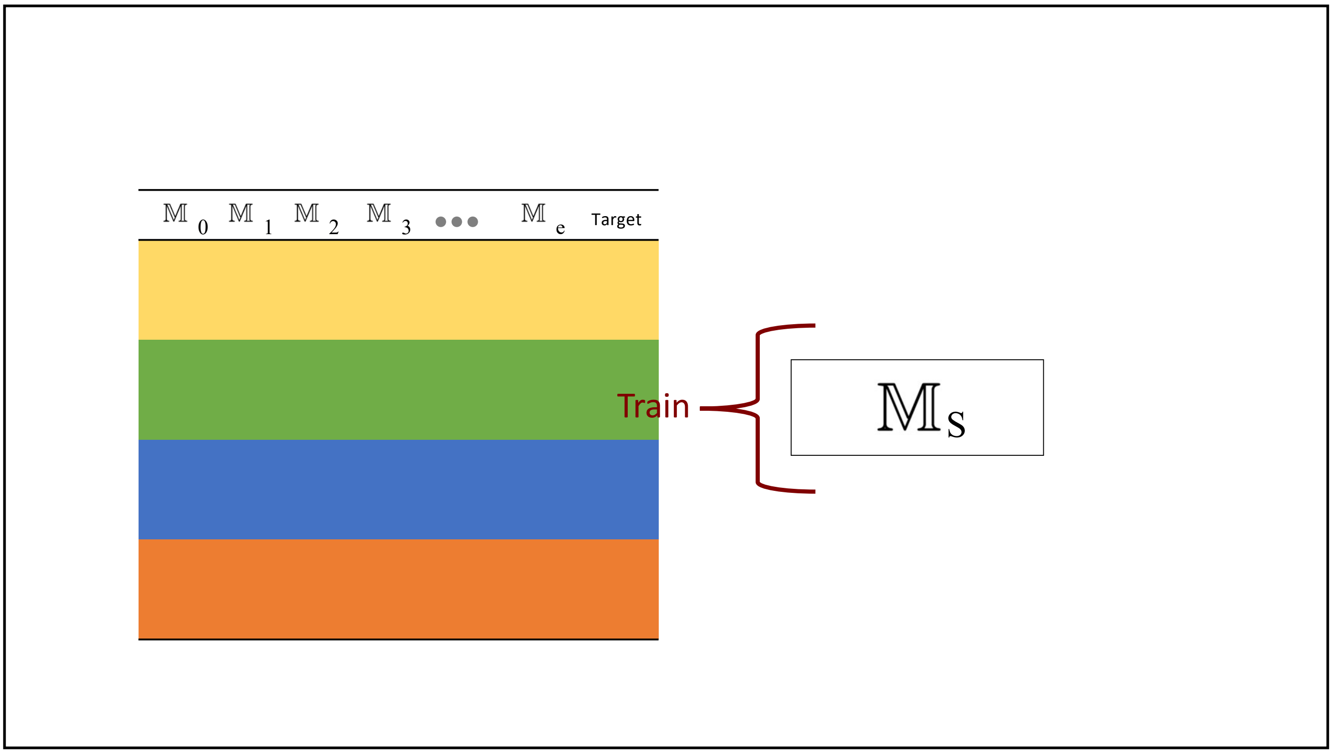
152



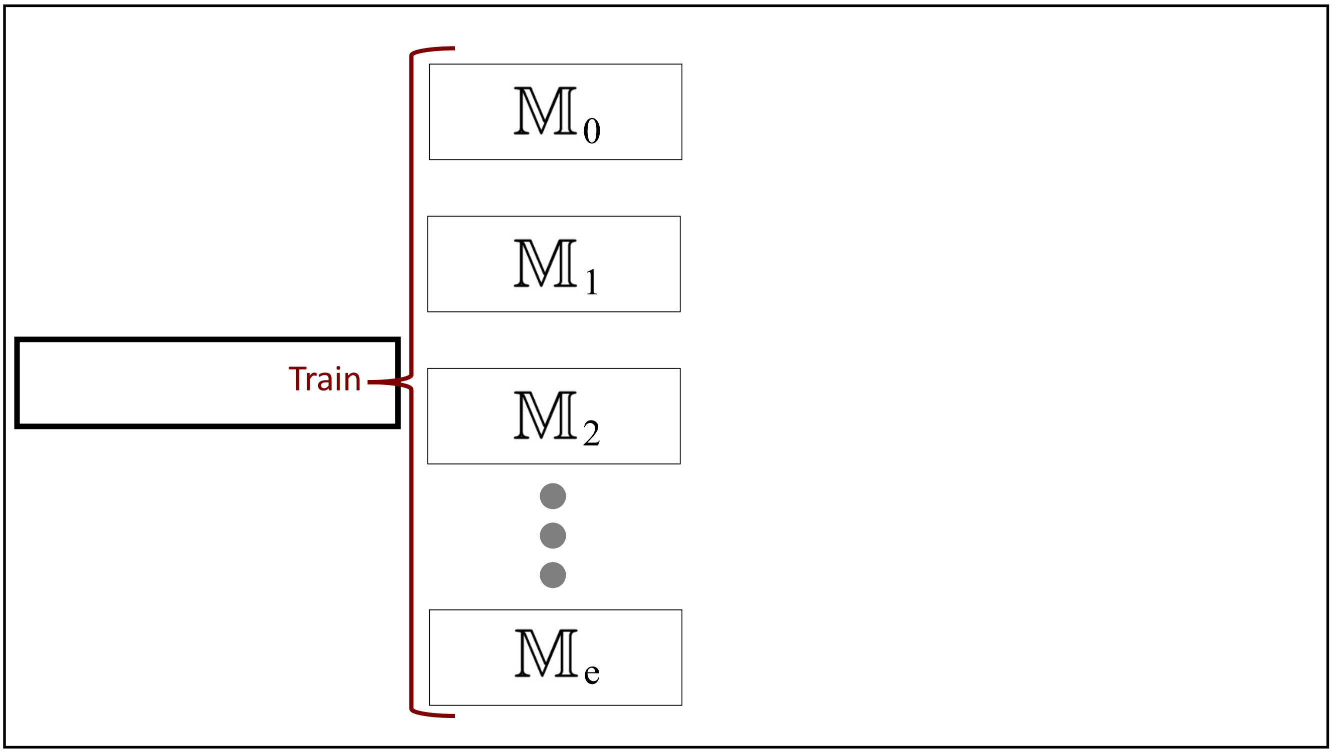
153



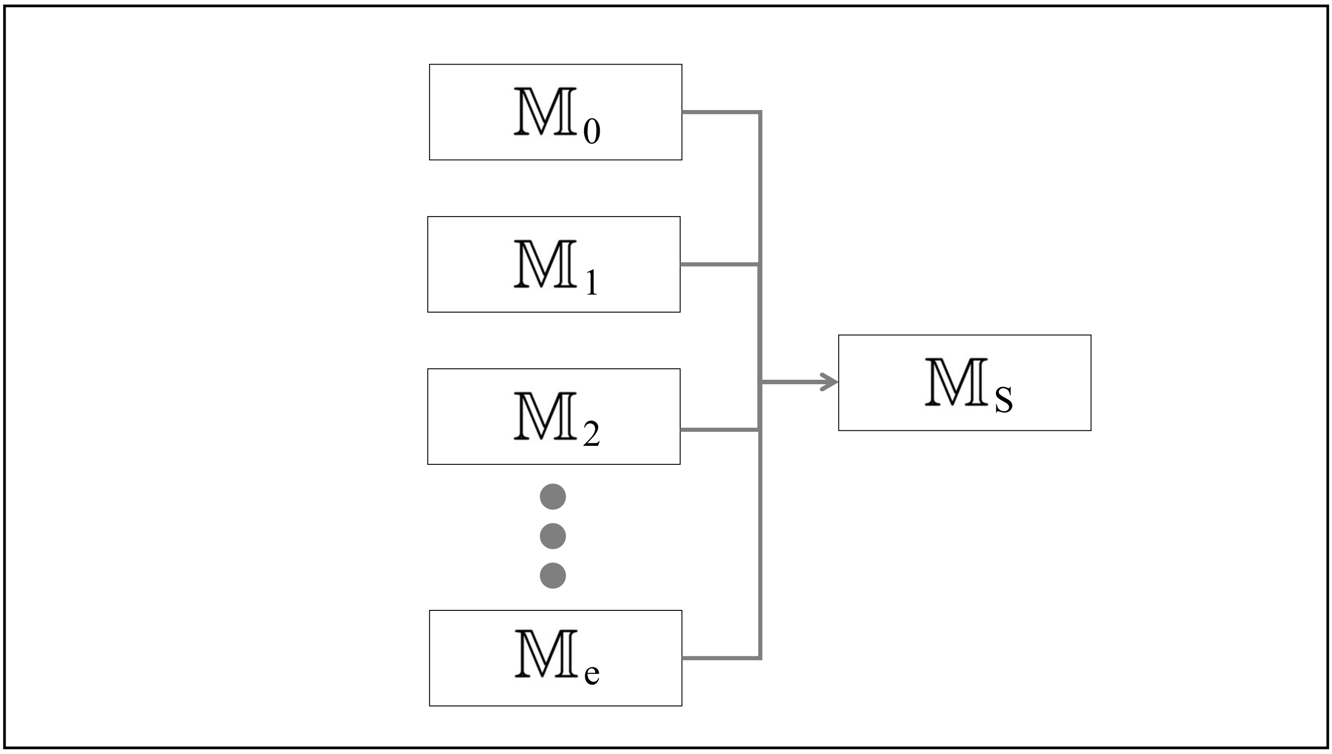
154



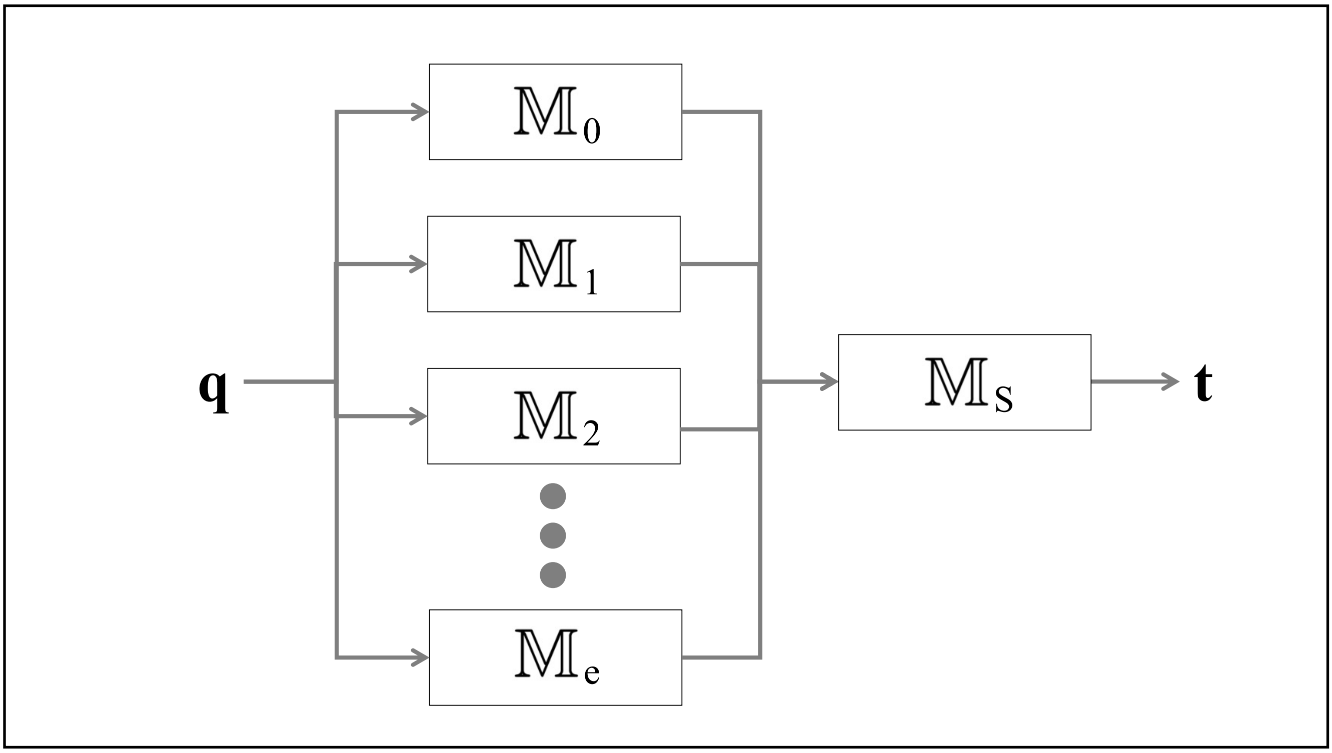
155



156


$\mathbb{M}_0$
$\mathbb{M}_1$
$\mathbb{M}_2$
$\mathbb{M}_e$
$\mathbb{M}_S$

157


q
$\mathbb{M}_0$
$\mathbb{M}_1$
$\mathbb{M}_2$
$\mathbb{M}_e$
$\mathbb{M}_S$
t

158

Stacking

It is very common to use **heterogenous ensembles** with stacking

Stacking takes a bit of work, but can be effective

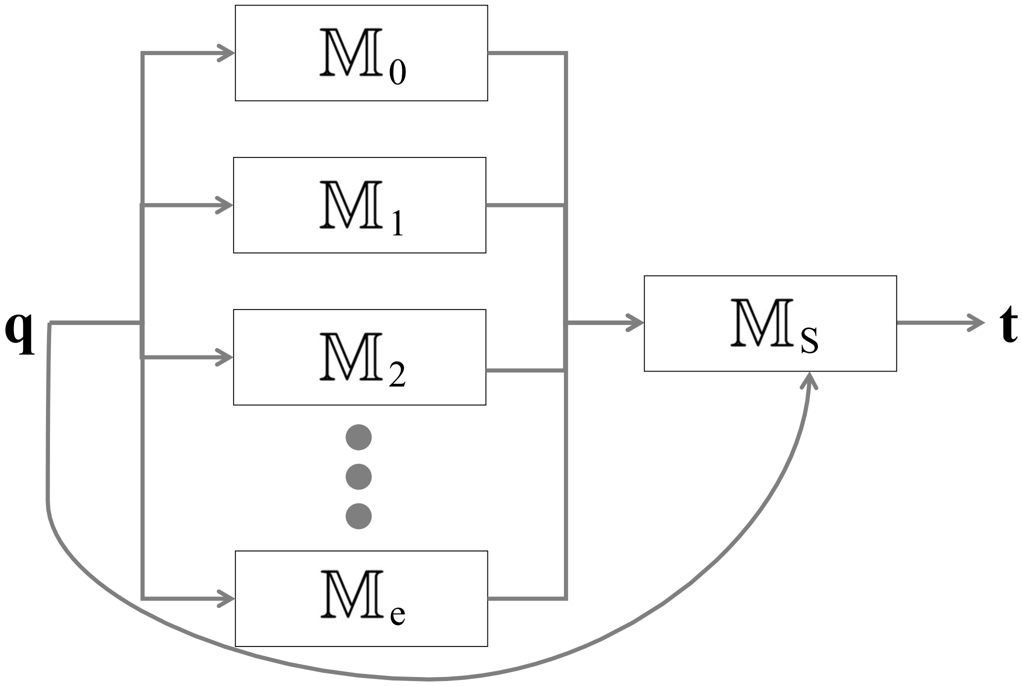
159

Stacking

We can also include the original feature vector as input to the stack layer model

This allows some focus on particular base models for certain areas of the input space

160



161

	$\mathbb{M}_0$	$\mathbb{M}_1$	$\mathbb{M}_2$	$\mathbb{M}_3$	●●●	$\mathbb{M}_e$	d[0]	●●●	d[m]	Target
d_0	0.81	0.22	0.76	0.91		0.11	0.56		-0.41	True
d_1	0.38	0.41	0.29	0.38		0.55	0.78		0.56	False
				⋮						
d_n	0.99	0.76	0.54	0.44		0.38	0.38		0.99	False

162