Computational Analysis of Big Data

Week 5

Machine Learning 2

Decision trees

Decision trees

Lays eggs	Cold blooded	Mammal
0	0	1
0	0	1
0	0	1
0	0	1
0	0	1
1	0	0
1	0	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0

features

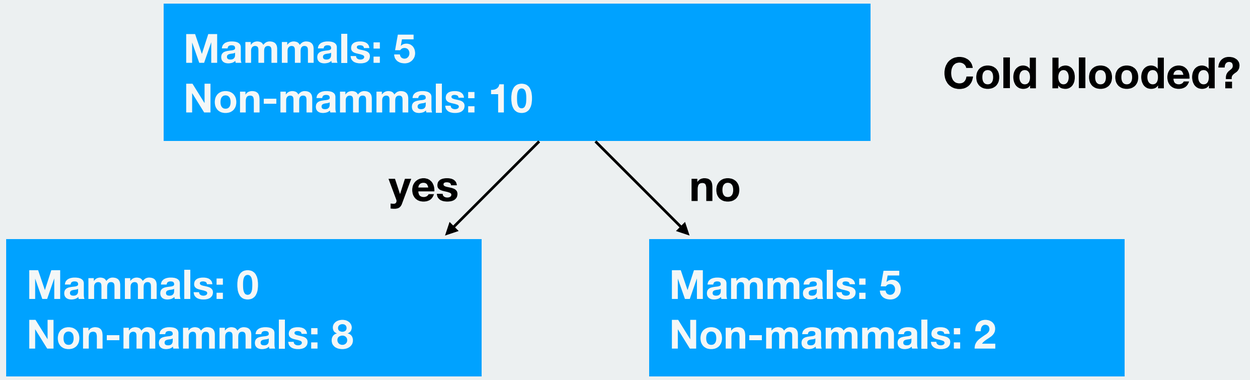
target

Mammals: 5
Non-mammals: 10

Decision trees

Lays eggs	Cold blooded	Mammal
0	0	1
0	0	1
0	0	1
0	0	1
0	0	1
1	0	0
1	0	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0
1	1	0

features **target**

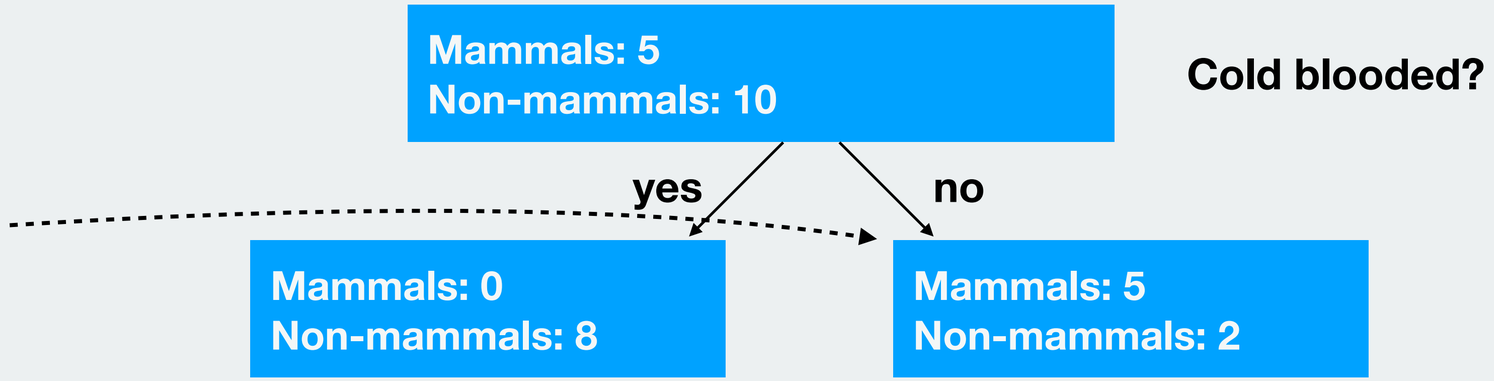


Decision trees

	Lays eggs	Cold blooded	Mammal
no	0	0	1
	0	0	1
	0	0	1
	0	0	1
	0	0	1
	1	0	0
	1	0	0
yes	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0

features

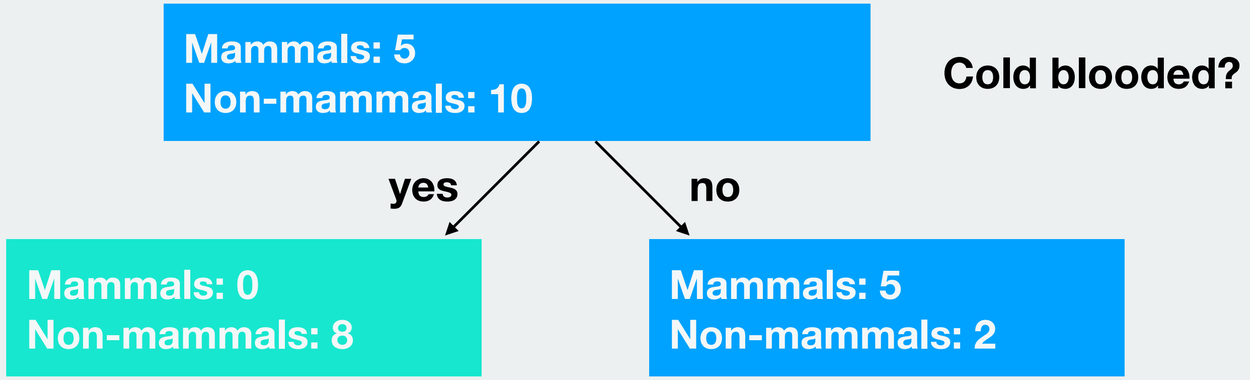
target



Decision trees

	Lays eggs	Cold blooded	Mammal
no	0	0	1
	0	0	1
	0	0	1
	0	0	1
	0	0	1
	1	0	0
	1	0	0
yes	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0

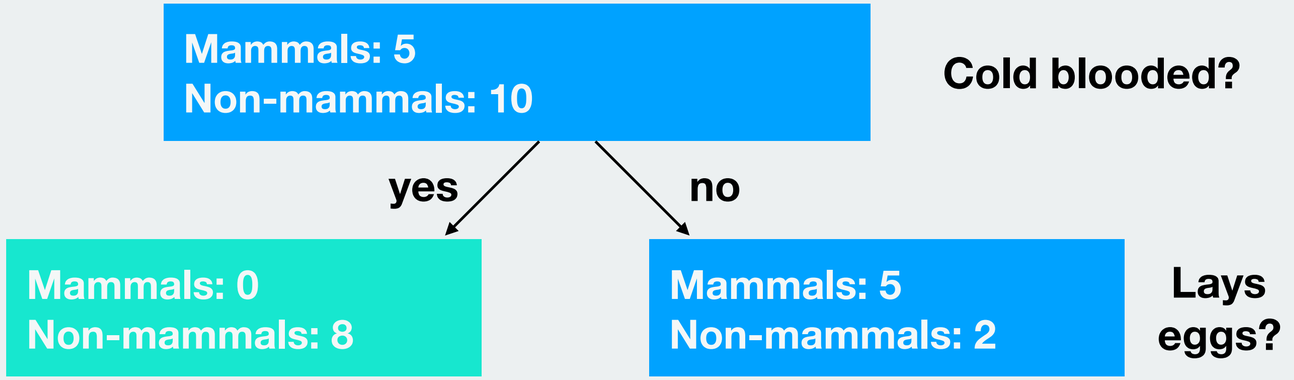
features **target**



Decision trees

	Lays eggs	Cold blooded	Mammal
no	0	0	1
	0	0	1
	0	0	1
	0	0	1
	0	0	1
	1	0	0
	1	0	0
yes	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0

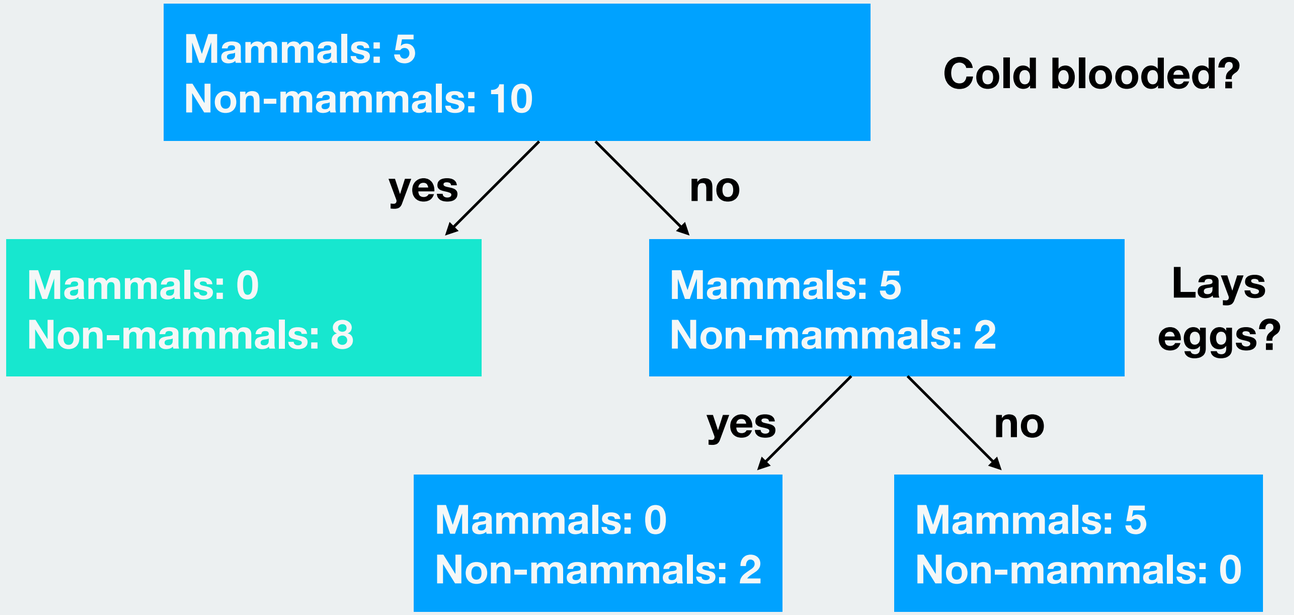
features **target**



Decision trees

	Lays eggs	Cold blooded	Mammal
no	0	0	1
	0	0	1
	0	0	1
	0	0	1
	0	0	1
yes	1	0	0
	1	0	0
yes	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0

features target

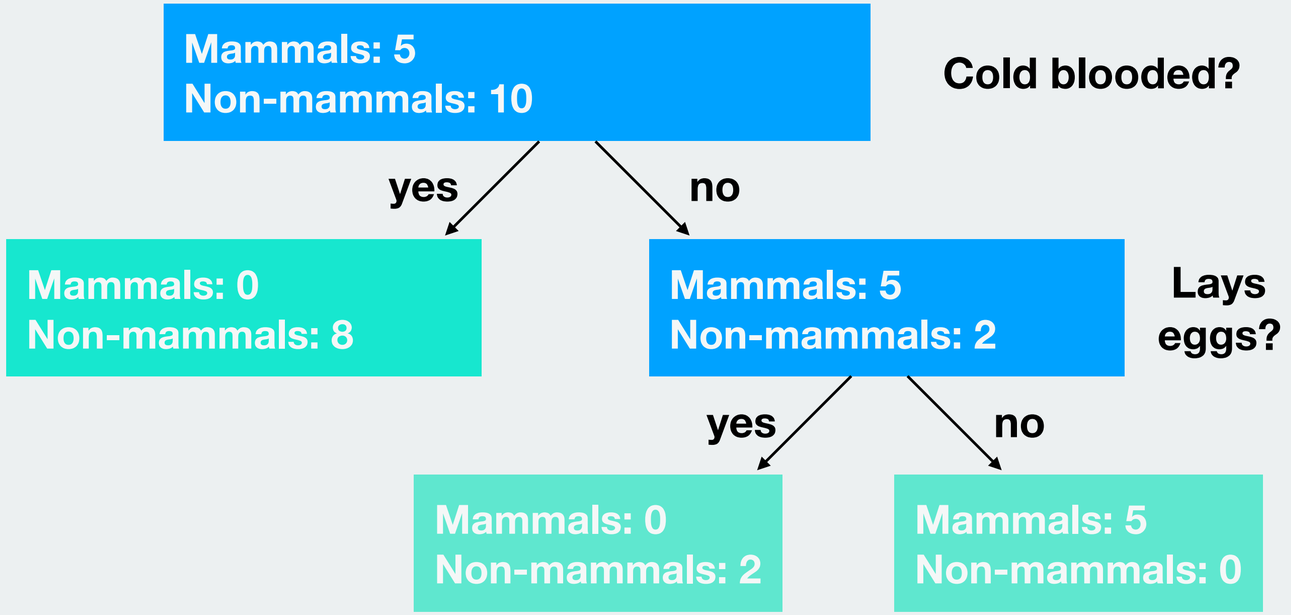


Decision trees

	Lays eggs	Cold blooded	Mammal
no	0	0	1
	0	0	1
	0	0	1
	0	0	1
	0	0	1
yes	1	0	0
	1	0	0
yes	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0

features

target



Decision trees

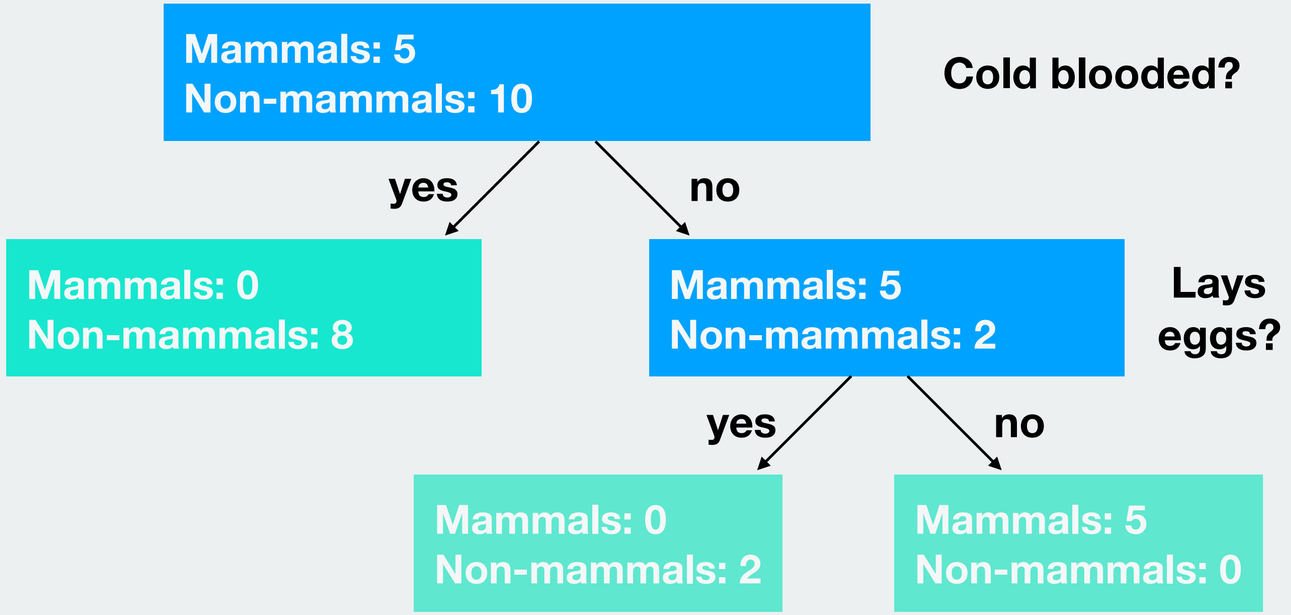
Could we have asked better questions?

Decision trees

	Lays eggs	Cold blooded	Mammal
no	0	0	1
	0	0	1
	0	0	1
	0	0	1
	0	0	1
yes	1	0	0
	1	0	0
yes	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0

features

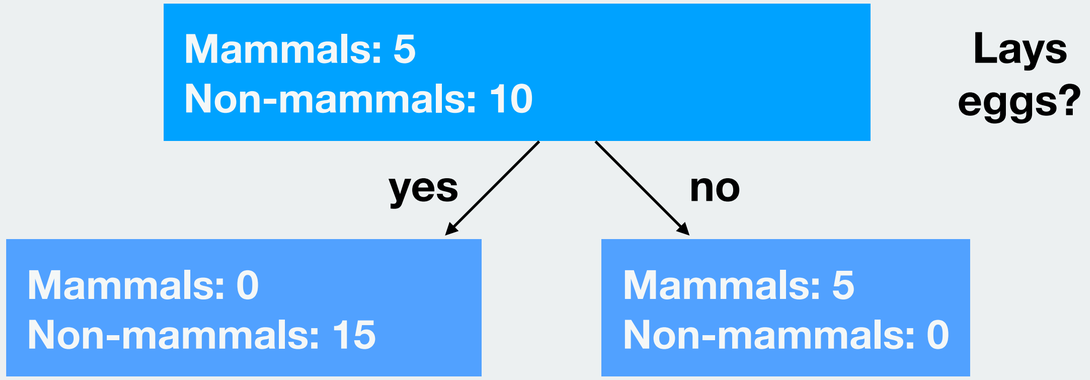
target



Decision trees

	Lays eggs	Cold blooded	Mammal
no	0	0	1
	0	0	1
	0	0	1
	0	0	1
	0	0	1
yes	1	0	0
	1	0	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0

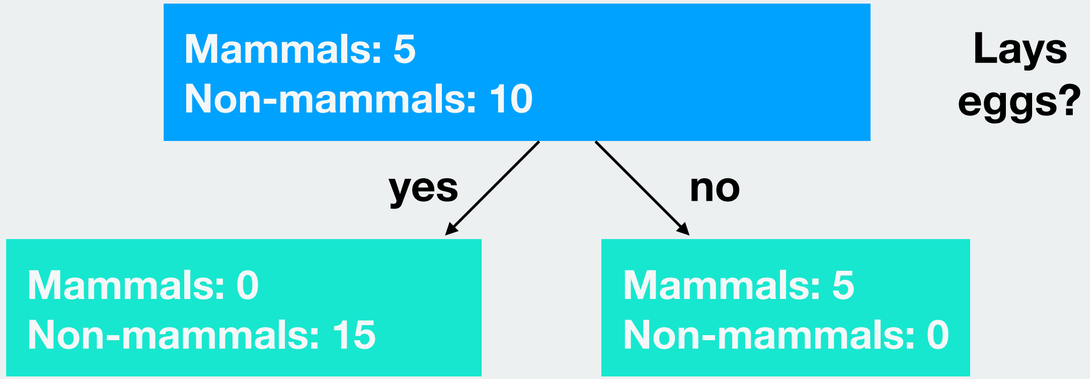
features **target**



Decision trees

	Lays eggs	Cold blooded	Mammal
no	0	0	1
	0	0	1
	0	0	1
	0	0	1
	0	0	1
yes	1	0	0
	1	0	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0
	1	1	0

features **target**



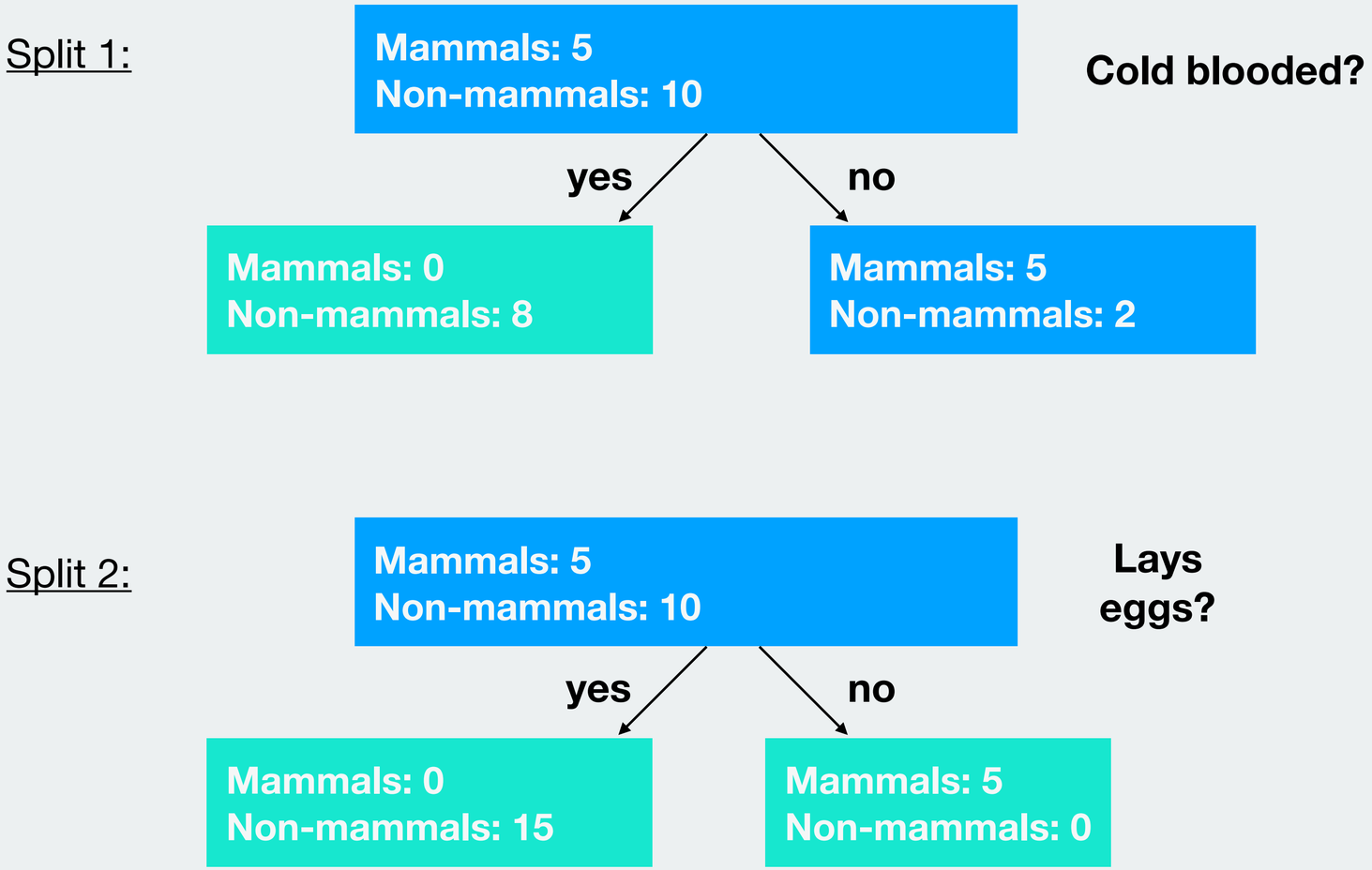
Decision trees

What do we do when we have Big Data?

Decision trees

	Pclass1	Pclass2	Pclass3	Sexfemale	Sexmale	Embarkednan	EmbarkedC	EmbarkedQ	EmbarkedS	CabinFalse	CabinTrue	PassengerId	Age	SibSp	Parch	Fare	Survived
0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	1	22.0	1	0	7.2500	0
1	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	2	38.0	1	0	71.2833	1
2	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	3	26.0	0	0	7.9250	1
3	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	4	35.0	1	0	53.1000	1
4	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	5	35.0	0	0	8.0500	0
5	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	6	NaN	0	0	8.4583	0
6	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	7	54.0	0	0	51.8625	0
7	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	8	2.0	3	1	21.0750	0
8	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	9	27.0	0	2	11.1333	1
9	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	10	14.0	1	0	30.0708	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
881	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	882	33.0	0	0	7.8958	0
882	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	883	22.0	0	0	10.5167	0
883	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	884	28.0	0	0	10.5000	0
884	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	885	25.0	0	0	7.0500	0
885	0.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	886	39.0	0	5	29.1250	0
886	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	887	27.0	0	0	13.0000	0
887	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	888	19.0	0	0	30.0000	1
888	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	889	NaN	1	2	23.4500	0
889	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	890	26.0	0	0	30.0000	1
890	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	891	32.0	0	0	7.7500	0

Automatic split selection



Automatic split selection

(Shannon) $$Entropy = -\sum_i p(i) \log_2 p(i)$$

Input: Probability vector (a list of values between 0 and 1, which sums to 1)

Output: Entropy (a measure of how “spread out” the probability distribution is)

Automatic split selection

$$Entropy = -\sum_i p(i) \log_2 p(i)$$

Mammals: 0
Non-mammals: 8

p = [1, 0]

Automatic split selection

$$Entropy = -\sum_i p(i) \log_2 p(i)$$

Mammals: 0
Non-mammals: 8

p = [1, 0]

Entropy = - (1 · $\log_2(1)$ + 0 · $\log_2(0)$) = 0

Automatic split selection

$$Entropy = -\sum_i p(i) \log_2 p(i)$$

Mammals: 0
Non-mammals: 8

p = [1, 0]

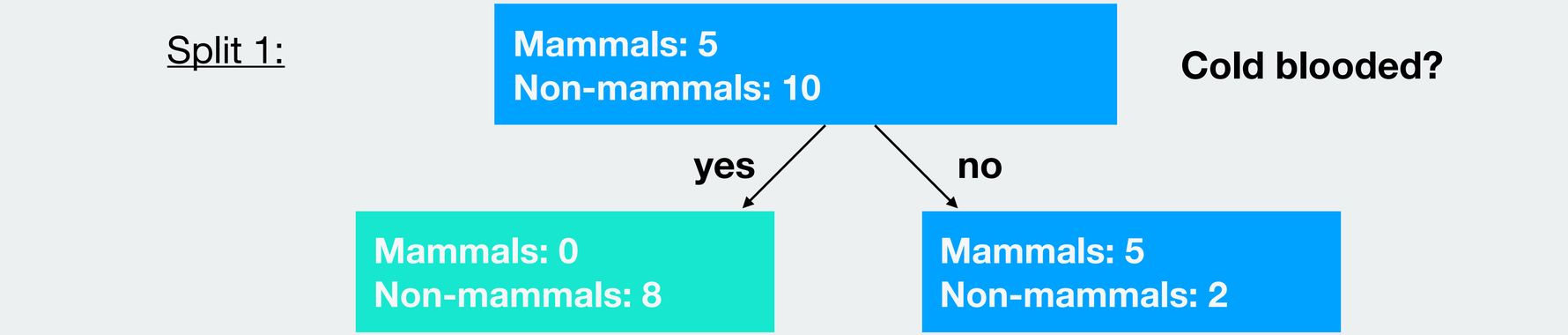
Entropy = - (1 · $\log_2(1)$ + 0 · $\log_2(0)$) = 0

Mammals: 5
Non-mammals: 2

p = [2/7, 5/7]

Entropy = - (2/7 · $\log_2(2/7)$ + 5/7 · $\log_2(5/7)$) = 0.86

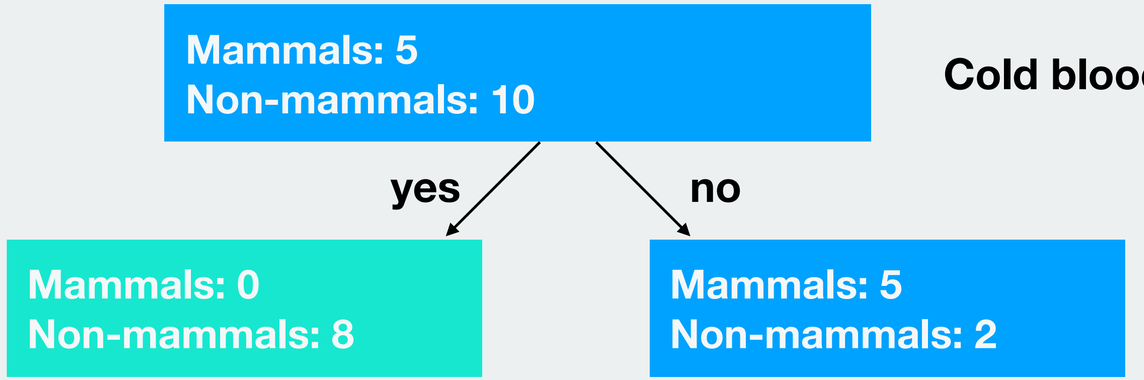
Automatic split selection



split entropy = 8 / 15 · 0 + 7 / 15 · 0.86 = 0.40

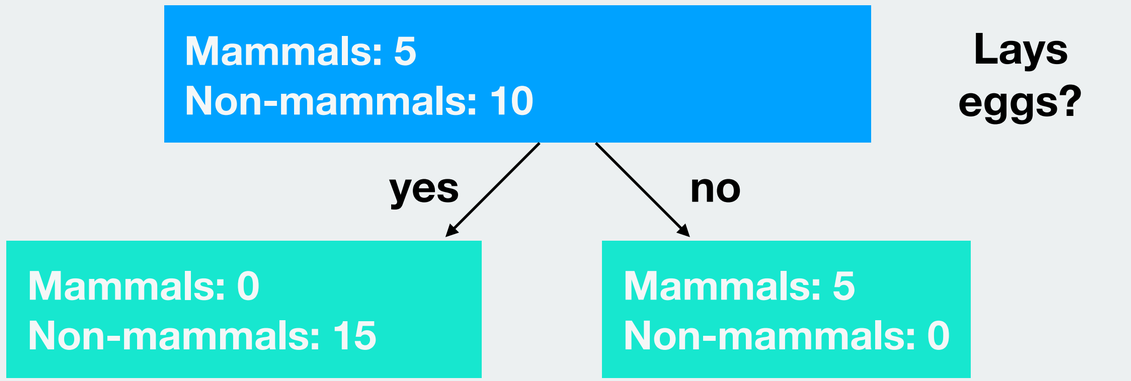
Automatic split selection

Split 1:



split entropy = 8 / 15 · 0 + 7 / 15 · 0.86 = 0.40

Split 2:



split entropy = 8 / 15 · 0 + 7 / 15 · 0 = 0

On steroids:

Ensemble Learning

Ensemble Learning

- Create and train many classification models
- Treat each model as a “voter”
- For each datapoint, classify it according to what models predicts it to be

Random Forest

model1

	Pclass1	Pclass2	Pclass3	Sexfemale	Sexmale	Embarkednan	EmbarkedC	EmbarkedQ	EmbarkedS	CabinFalse	CabinTrue	PassengerId	Age	SibSp	Parch	Fare	Survived
0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	1	22.0	1	0	7.2500	0
1	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	2	38.0	1	0	71.2833	1
2	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	3	26.0	0	0	7.9250	1
3	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	4	35.0	1	0	53.1000	1
4	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	5	35.0	0	0	8.0500	0
5	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	6	NaN	0	0	8.4583	0
6	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	7	54.0	0	0	51.8625	0
7	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	8	2.0	3	1	21.0750	0
8	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	9	27.0	0	2	11.1333	1
9	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	10	14.0	1	0	30.0708	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
881	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	882	33.0	0	0	7.8958	0
882	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	883	22.0	0	0	10.5167	0
883	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	884	28.0	0	0	10.5000	0
884	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	885	25.0	0	0	7.0500	0
885	0.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	886	39.0	0	5	29.1250	0
886	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	887	27.0	0	0	13.0000	0
887	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	888	19.0	0	0	30.0000	1
888	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	889	NaN	1	2	23.4500	0
889	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	890	26.0	0	0	30.0000	1
890	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	891	32.0	0	0	7.7500	0

Random Forest

model2

	Pclass1	Pclass2	Pclass3	Sexfemale	Sexmale	Embarkednan	EmbarkedC	EmbarkedQ	EmbarkedS	CabinFalse	CabinTrue	PassengerId	Age	SibSp	Parch	Fare	Survived
0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	1	22.0	1	0	7.2500	0
1	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	2	38.0	1	0	71.2833	1
2	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	3	26.0	0	0	7.9250	1
3	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	4	35.0	1	0	53.1000	1
4	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	5	35.0	0	0	8.0500	0
5	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	6	NaN	0	0	8.4583	0
6	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	7	54.0	0	0	51.8625	0
7	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	8	2.0	3	1	21.0750	0
8	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	9	27.0	0	2	11.1333	1
9	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	10	14.0	1	0	30.0708	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
881	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	882	33.0	0	0	7.8958	0
882	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	883	22.0	0	0	10.5167	0
883	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	884	28.0	0	0	10.5000	0
884	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	885	25.0	0	0	7.0500	0
885	0.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	886	39.0	0	5	29.1250	0
886	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	887	27.0	0	0	13.0000	0
887	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	888	19.0	0	0	30.0000	1
888	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	889	NaN	1	2	23.4500	0
889	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	890	26.0	0	0	30.0000	1
890	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	891	32.0	0	0	7.7500	0

Random Forest

model3

	Pclass1	Pclass2	Pclass3	Sexfemale	Sexmale	Embarkednan	EmbarkedC	EmbarkedQ	EmbarkedS	CabinFalse	CabinTrue	PassengerId	Age	SibSp	Parch	Fare	Survived
0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	1	22.0	1	0	7.2500	0
1	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	2	38.0	1	0	71.2833	1
2	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	3	26.0	0	0	7.9250	1
3	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	4	35.0	1	0	53.1000	1
4	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	5	35.0	0	0	8.0500	0
5	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	6	NaN	0	0	8.4583	0
6	1.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	7	54.0	0	0	51.8625	0
7	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	8	2.0	3	1	21.0750	0
8	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	9	27.0	0	2	11.1333	1
9	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	0.0	1.0	0.0	10	14.0	1	0	30.0708	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
881	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	882	33.0	0	0	7.8958	0
882	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	883	22.0	0	0	10.5167	0
883	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	884	28.0	0	0	10.5000	0
884	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	885	25.0	0	0	7.0500	0
885	0.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	886	39.0	0	5	29.1250	0
886	0.0	1.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	887	27.0	0	0	13.0000	0
887	1.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	888	19.0	0	0	30.0000	1
888	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	1.0	1.0	0.0	889	NaN	1	2	23.4500	0
889	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	890	26.0	0	0	30.0000	1
890	0.0	0.0	1.0	0.0	1.0	0.0	0.0	1.0	0.0	1.0	0.0	891	32.0	0	0	7.7500	0

Ensemble Learning

- Create and train many classification models
- Treat each model as a “voter”
- For each datapoint, classify it according to what models predicts it to be