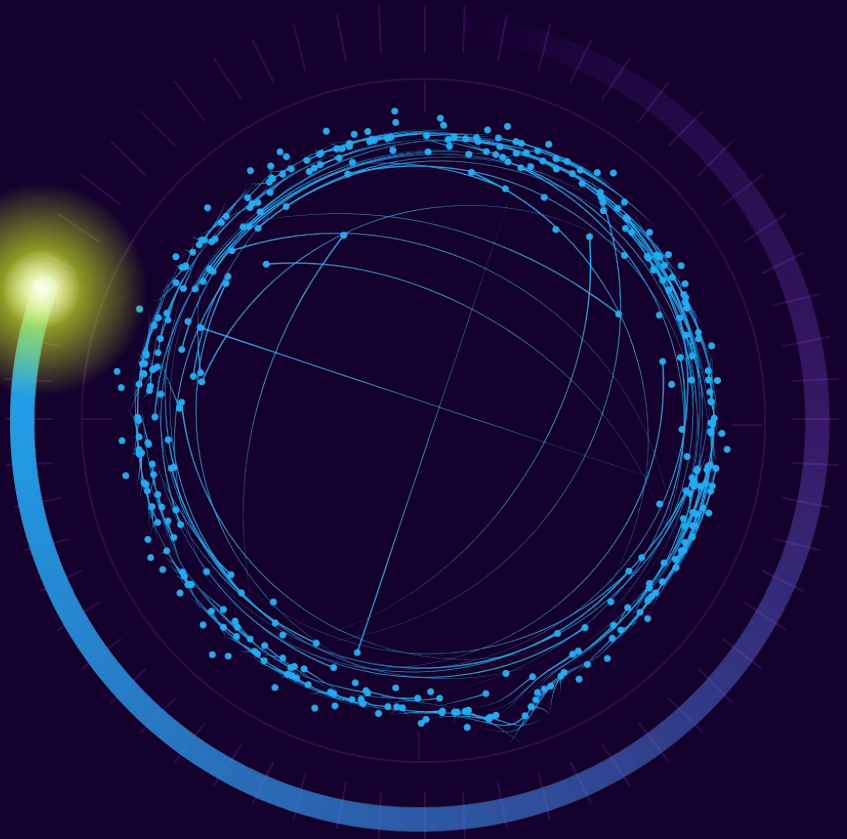
influxdata
Act in Time

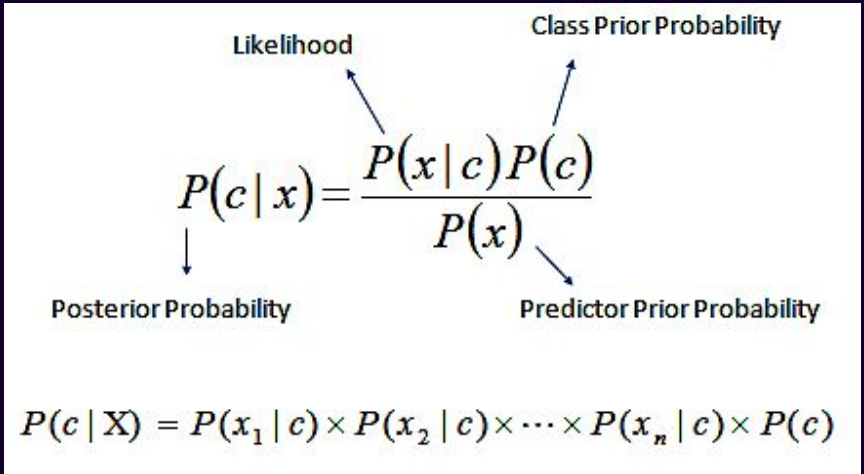
Building a Naive Bayes classifier using Flux

by Team Magic



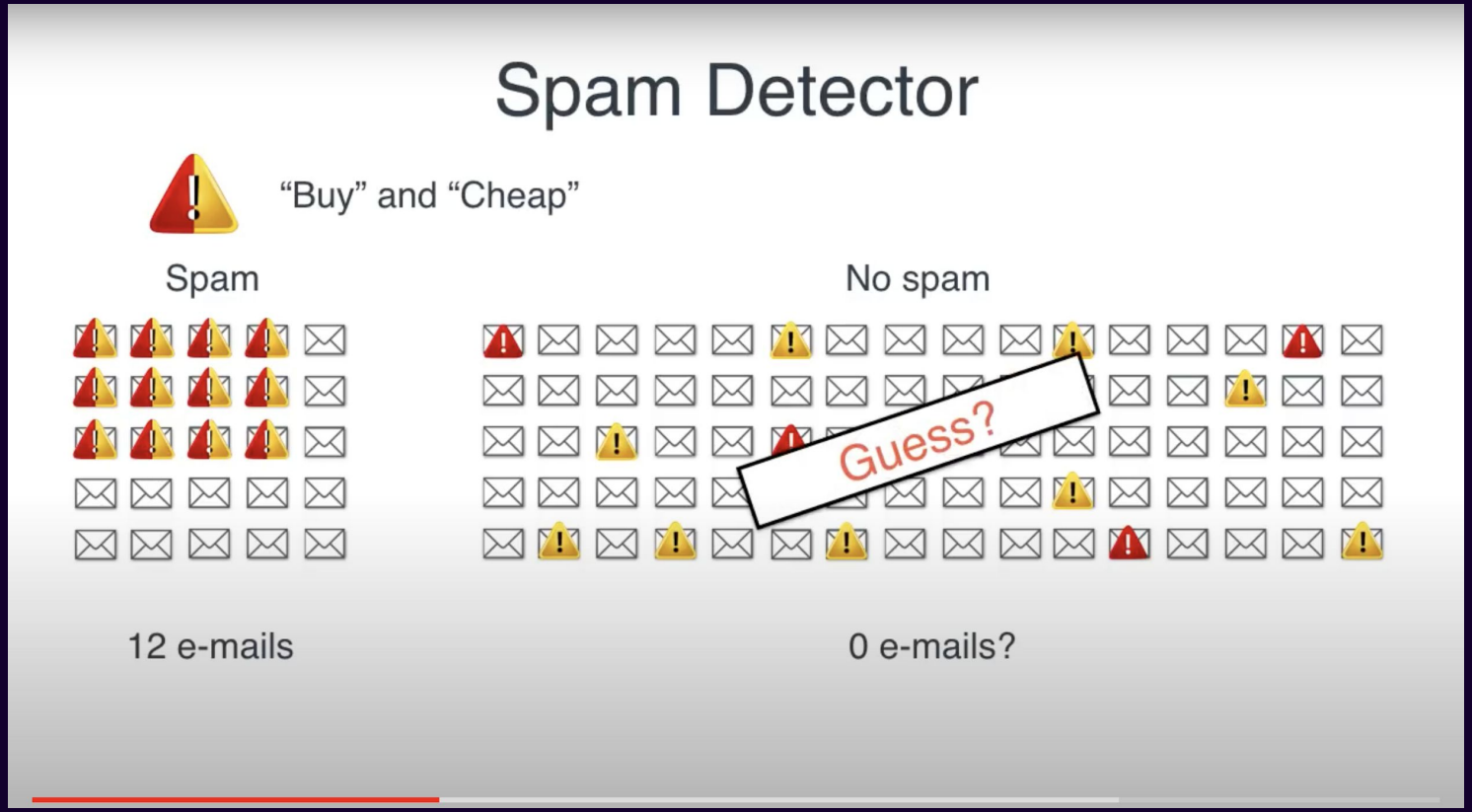
The Naive Bayes Algorithm

Naive Bayes classifiers are a family of simple "probabilistic classifiers" based on applying Bayes' theorem with strong (naïve) independence assumptions between the features.

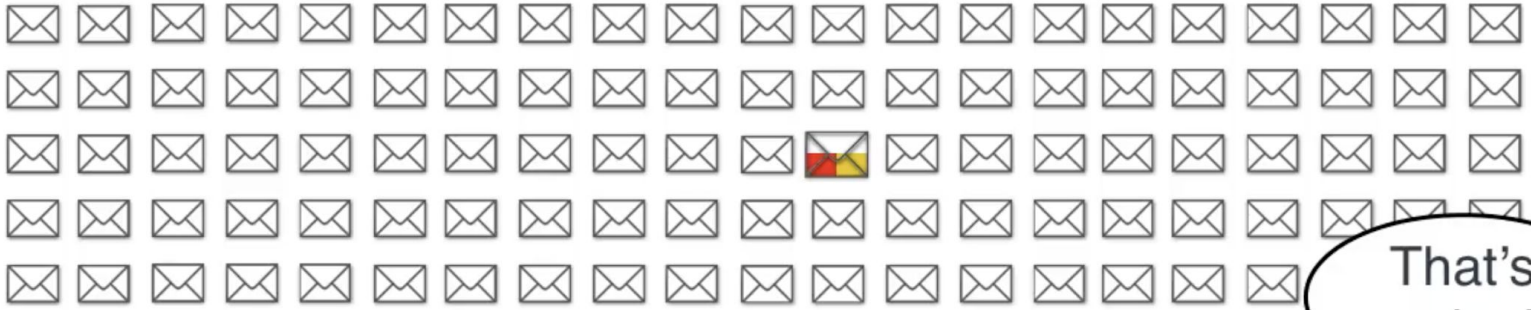


Example: a fruit can be considered to be an apple if it is red, round, and about 10 cm in diameter. A naive Bayes classifier assumes each of these features contribute independently to the probability that this fruit is an apple, regardless of any possible correlations between the color, roundness, and diameter features.

A Great Example: Spam Filter



Spam Detector



That's naive!

100 e-mails

5 "Buy"

10 "Cheap"

5% "Buy"

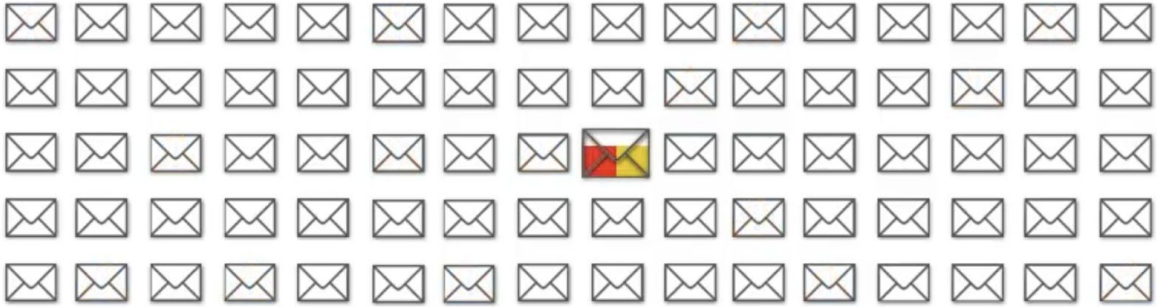
10% "Cheap"

Independent

0.5% "Buy" and "Cheap"

Spam Detector

No spam



75 e-mails

5 “Buy” 1/15

10 “Cheap” 2/15

2/225 × 75 = 2/3 “Buy” and “Cheap”

Naive Bayes Classifier

"Buy" and "Cheap" → 94.737%

Spam

No spam

12

94.737%

2/3

5.263%

Quiz: If an e-mail contains the words "buy" and "cheap", what is the probability that it is spam?

$$\frac{12}{12 + 2/3} = \frac{36}{38}$$

$$= 94.737\%$$

Why Flux? 🤔

We chose **flux** to implement our classifier to illustrate it's **unique, data-intensive** capabilities

Building a simple classifier with simple data

Simple Data:

```
TrainingData = "
#datatype,string,long,string,string,string,dateTime:RFC3339,string
#group,false,false,true,true,true,false,false
#default,_result,,,,,,
,result,table,_measurement,_field,Class,_time,_value
,,0,m1,f1,Yes,2018-12-19T22:13:30Z,A
,,0,m1,f1,Yes,2018-12-19T22:13:40Z,A
,,0,m1,f1,Yes,2018-12-19T22:13:50Z,A
,,0,m1,f1,Yes,2018-12-19T22:14:00Z,B
,,1,m1,f1,No,2018-12-19T22:14:10Z,A
,,1,m1,f1,No,2018-12-19T22:14:20Z,B
,,1,m1,f1,No,2018-12-19T22:13:30Z,B
,,1,m1,f1,No,2018-12-19T22:13:40Z,B
"
```

Feature	Class	Value
f1	Yes	A
f1	Yes	A
f1	Yes	A
f1	Yes	B
f1	No	A
f1	No	B
f1	No	B
f1	No	B

Building a simple classifier with simple data

Simple Classifier:

_value	Class	p_k	p_x	P_x_k	Probability
A	No	0.5	0.5	0.25	0.25
A	Yes	0.5	0.5	0.75	0.75
B	No	0.5	0.5	0.75	0.75
B	Yes	0.5	0.5	0.25	0.25

Question: The result occurs if A, what is the probability this statement is true?

P(Yes | A) = P(A | Yes) * P(Yes) / P(A) = P_x_k * p_k / p_x = 0.75 * 0.5 / 0.5 = 0.75

_value	Class	Probability
A	Yes	0.75

Probability an animal is airborne given its aquatic

Animal_name	P_x_k	Probability	_field_Probabilit…	_field_r	_value
buffalo	0.5636363636363636	0.6888888888888888	aquatic	aquatic	0
bear	0.5636363636363636	0.6888888888888888	aquatic	aquatic	0
boar	0.5636363636363636	0.6888888888888888	aquatic	aquatic	0
calf	0.5636363636363636	0.6888888888888888	aquatic	aquatic	0
cavy	0.5636363636363636	0.6888888888888888	aquatic	aquatic	0
cheetah	0.5636363636363636	0.6888888888888888	aquatic	aquatic	0
aardvark	0.5636363636363636	0.6888888888888888	aquatic	aquatic	0
chicken	0.5636363636363636	0.6888888888888888	aquatic	aquatic	0
antelope	0.5636363636363636	0.6888888888888888	aquatic	aquatic	0
clam	0.5636363636363636	0.6888888888888888	aquatic	aquatic	0
clam	0.7368421052631579	0.31111111111111106	aquatic	aquatic	0
bear	0.7368421052631579	0.31111111111111106	aquatic	aquatic	0
boar	0.7368421052631579	0.31111111111111106	aquatic	aquatic	0
calf	0.7368421052631579	0.31111111111111106	aquatic	aquatic	0
cavy	0.7368421052631579	0.31111111111111106	aquatic	aquatic	0
cheetah	0.7368421052631579	0.31111111111111106	aquatic	aquatic	0
aardvark	0.7368421052631579	0.31111111111111106	aquatic	aquatic	0
chicken	0.7368421052631579	0.31111111111111106	aquatic	aquatic	0
antelope	0.7368421052631579	0.31111111111111106	aquatic	aquatic	0
buffalo	0.7368421052631579	0.31111111111111106	aquatic	aquatic	0
carp	0.43636363636363634	0.8275862068965516	aquatic	aquatic	1
crab	0.43636363636363634	0.8275862068965516	aquatic	aquatic	1
chub	0.43636363636363634	0.8275862068965516	aquatic	aquatic	1
catfish	0.43636363636363634	0.8275862068965516	aquatic	aquatic	1
bass	0.43636363636363634	0.8275862068965516	aquatic	aquatic	1
carp	0.2631578947368421	0.17241379310344826	aquatic	aquatic	1

Demo time!

Looking forwards, we'd like to...

- add more fields/features to our classifier to improve the accuracy of our results
- consider using non-binary data
- implement potential density functions (i.e. Gauss)
- use our algorithm to classify more relevant datasets (for example, slack incidents)
- create a graphic user interface that allows users to feed in training/test data

A huge thank you to Anais and Adam!
We couldn’t have done it without you!!!