Lecture 23

Statistical

Machine

Learning

Ingredients — data driven

↓ algorithm

decisions

CAPTCHA: users to label data with correct answers

Social media (algorithms)

medicine

GPT3/ LLM → neural networks

↳ predictive text, also input

AI

rule-based

Statistical

1. build a statistical model from some data
2. use the model to make predictions

Supervised

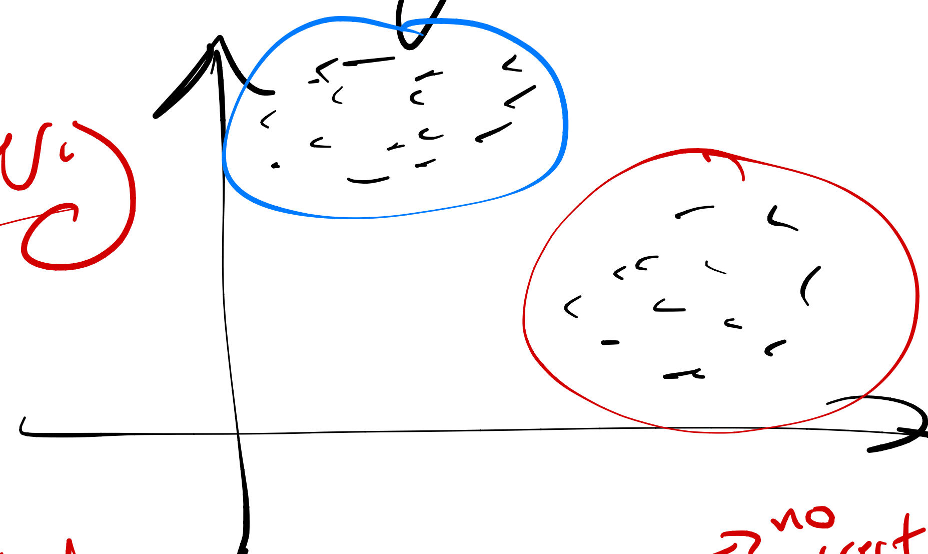
"training data"

inputs with the correct outputs labeled

Unsupervised

learn patterns with no training data

clustering



"test data" / actual examples

→ no correct answer

Classification problem

input $\longrightarrow$ classification

email $\longrightarrow$
- spam
- not spam

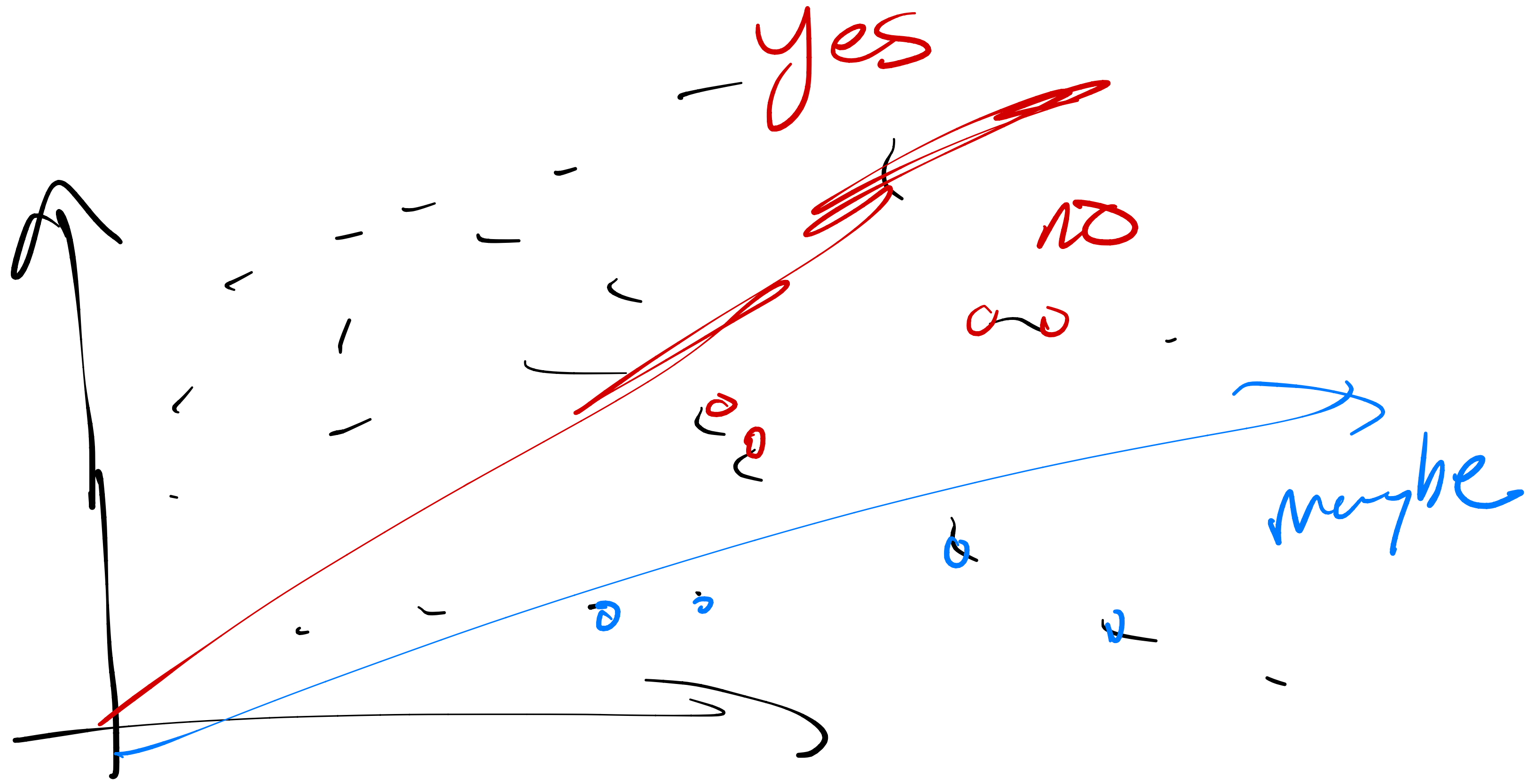
email $\longrightarrow$
- inbox
- promos
- social

⋮

document $\longrightarrow$ category:

news articles

- economics
- govt
- sports


yes
no
maybe

email 1 yes
email 2 no

Idea:

1. build a statistical model from training data
   examples labeled w/ answer

2. use that model to classify a new document (no answer yet)

email 187 → ?

test

live

Naive Bayes Classification

Probability of an "event"

$P(A)$ is a real number between 0.0 and 1.0

A ↳ event

0 not going to happen

1 certainly going to happen

P(it will rain today) is 0.75

Conditional probability:

outcome, input

$$P(A \mid B)$$

"given"

Probability that event A happens

If event B happens

email is spam

input

$P($spam | words are "credit card number...."$)$

$$P(\text{spam} \mid \text{input words})$$ if bigger ↪ goes in Junk

$$> ?$$

$$P(\text{not spam} \mid \text{input words})$$ if bigger ↪ goes in Inbox

$P(\text{sports} \mid \text{article})$

$P(\text{local} \mid \text{article})$

$P(\text{world} \mid \text{article})$

$P(\text{you} \mid \text{articles})$

pick highest probability

Idea: predict this from training data

article1 world

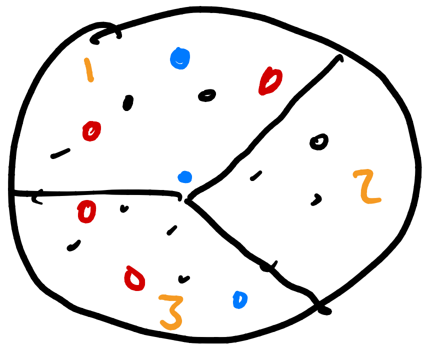
article sports

⋮

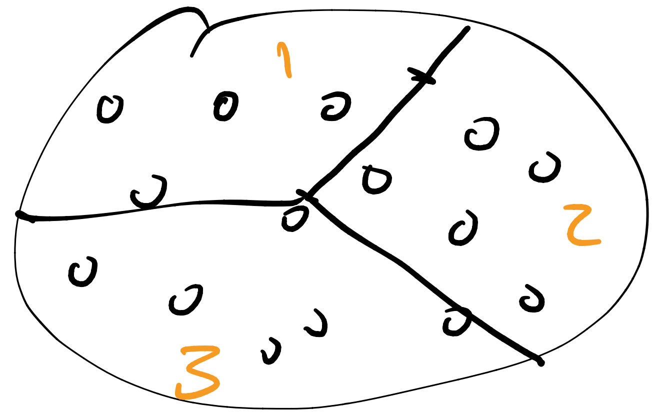
Bayes' rule

$$\underbrace{P(B \mid A)}_{\text{want to know}} = \underbrace{\frac{P(A \mid B)\, P(B)}{P(A)}}_{\text{something you can observe}}$$

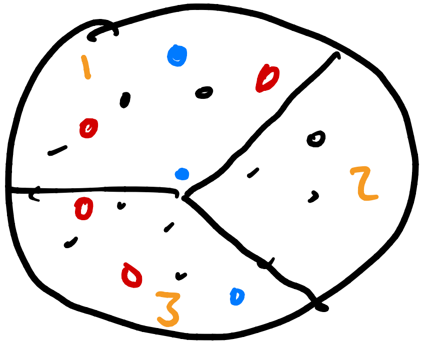
M&M



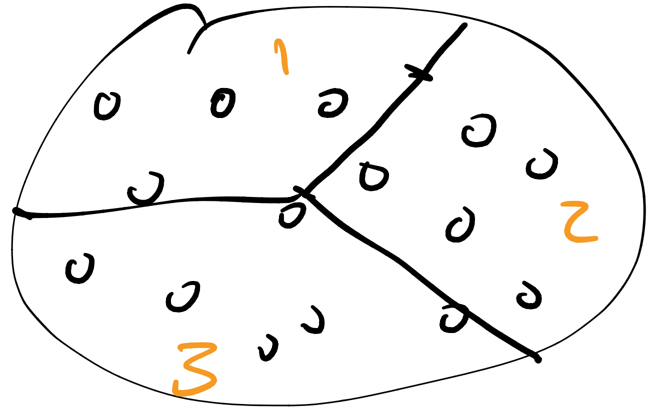
Chocolate chip
Cookie



try to predict whether
the cookie is M&M
or not based on 1 bite



candy cookie



no candy cookie

making use of an observation

category
spam

text
text

Want to know $P(\text{candy cookie} \mid \text{no candy bite})$

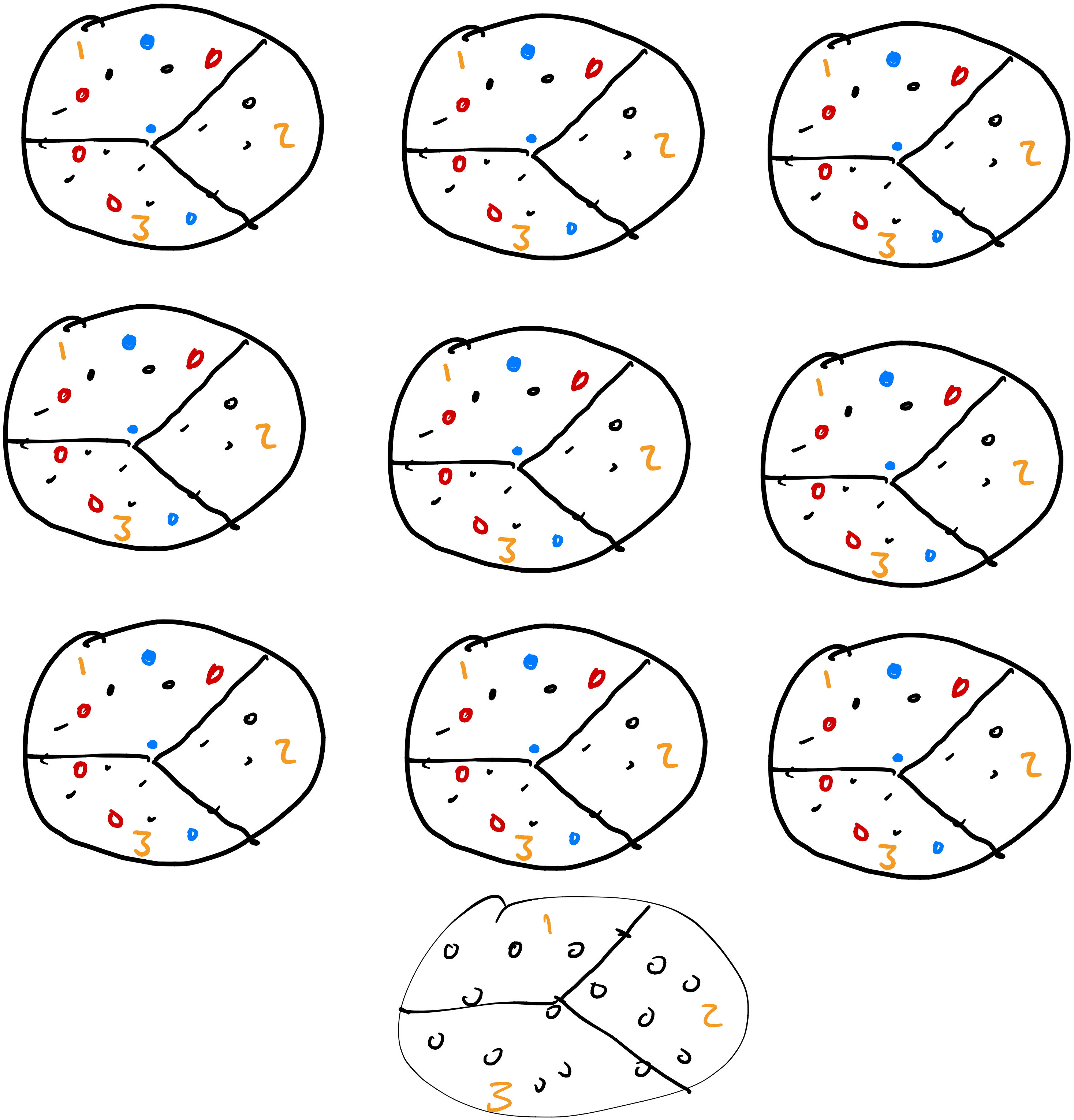
want to know

observation

flip

$P(\text{no candy bite} \mid \text{candy cookie}) = \frac{1}{3}$

$P(\text{no candy bite} \mid \text{no candy cookie}) = 1$



base rate

prior

30 bites

12 have no candy

9 from candy cookies

3 from no candy cookies

$$P(\overset{B}{\text{candy cookie}} \mid \overset{A}{\text{no candy bite}}) \qquad P(B \mid A) = \frac{P(A \mid B)\,P(B)}{P(A)}$$

$$= \frac{P(\overset{A}{\text{no candy bite}} \mid \overset{B}{\text{candy cookie}})\; P(\overset{B}{\text{candy cookie}})}{P(\text{no candy bite})}$$

$$= \frac{\frac{1}{3} \times \overset{\text{base rate}}{\frac{9}{10}}}{\frac{12}{30}} = \frac{\frac{9}{30}}{\frac{12}{30}} = \frac{9}{12} = \frac{3}{4}$$

Text classification

Have training data:

emails labeled with spam or not

Classify a new email with words

$\langle w_1, w_2, w_3 \ldots \rangle$

$$P\left(\begin{matrix}\text{email is}\\ \text{spam}\end{matrix} \,\middle|\, \begin{matrix}\text{words}\\ w_1, w_2, \ldots\end{matrix}\right) = ?$$

$\geq$?

$$P\left(\begin{matrix}\text{email is}\\ \text{not}\\ \text{spam}\end{matrix} \,\middle|\, \begin{matrix}\text{words}\\ w_1, w_2, \ldots\end{matrix}\right) = ?$$

$$P\left(\begin{matrix}\text{email is}\\ \text{spam}\end{matrix} \,\middle|\, \begin{matrix}\text{words}\\ w_1, w_2, \ldots\end{matrix}\right) \approx = \frac{P\left(\begin{matrix}\text{words}\\ w_1 \ldots w_n\end{matrix} \,\middle|\, \text{spam}\right) P(\text{spam})}{\cancel{P(\text{words})}_{w_1, w_2 \ldots w_n}}$$

$\sum$?

$$P\left(\begin{matrix}\text{email is}\\ \text{not}\\ \text{spam}\end{matrix} \,\middle|\, \begin{matrix}\text{words}\\ w_1, w_2, \ldots\end{matrix}\right) \approx = \frac{P\left(\begin{matrix}\text{words}\\ w_1 \ldots w_n\end{matrix} \,\middle|\, \begin{matrix}\text{not}\\ \text{spam}\end{matrix}\right) P\left(\begin{matrix}\text{not}\\ \text{spam}\end{matrix}\right)}{\cancel{P(\text{words})}_{w_1 w_2 \ldots w_n}}$$

not needed for relative

$$P\left(\substack{\text{email is} \\ \text{spam}} \;\middle|\; \substack{\text{words} \\ w_1, w_2, \ldots}\right) \approx \underbrace{P\left(\substack{\text{words} \\ w_1 \ldots w_n} \;\middle|\; \text{spam}\right) P(\text{spam})}$$

can be observed from the training data

$\boxed{\text{Training data}}$

email 1 $w_1 \ldots w_n$	spam
email 2 $v_1 \ldots v_n$	not spam
⋮	spam

just look at spam ones

$$= \underbrace{\text{Naive Bayes}} : \text{bag of words} \quad \prod_{i=1}^{n} \underline{P\left(\substack{\text{word} \\ w_i} \;\middle|\; \text{spam}\right)} \left(\frac{\#\text{ spam emails}}{\text{total }\#\text{ emails}}\right) \text{ base rate}$$

$$= \left(\prod_{i=1}^{n} \frac{\#\text{times } w_i \text{ occurs in spam}}{\#\text{ words in spam}}\right) \left(\frac{\#\text{ spam emails}}{\#\text{ emails}}\right)$$