

PART II

Handling probability models

PART III

Inference / generalization

What can I say about ...

FITTING

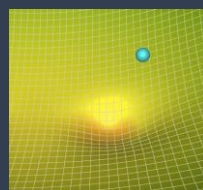
HOLDOUT

... which model to choose?

... parameter values?

MLE / prediction
accuracy

Pick the parameter
values that give the
best-fitting model



LECTURE 2

All of machine learning is based on a single idea:

1. Write out a probability model
2. Fit the model from data

This is behind

- A-level statistics formulae
- our climate model
- ChatGPT training

i.e. estimate its parameters
using Maximum Likelihood
Estimation (mle)

What can I say about ...

FITTING

HOLDOUT

... which model to choose?

MLE / prediction
accuracy
Pick the model that
best fits the data

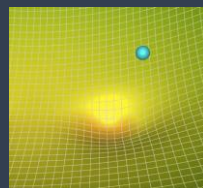
... parameter values?

MLE / prediction
accuracy
Pick the parameter
values that give the
best-fitting model

If we choose between
parameter values by
what value makes the
model fit best

as measured by e.g. the
model's MSE or
 $\Pr(\text{data}; \text{pars})$

then it's natural to
choose between
models in the same
way.



What can I say about ...

FITTING

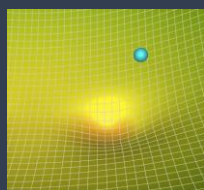
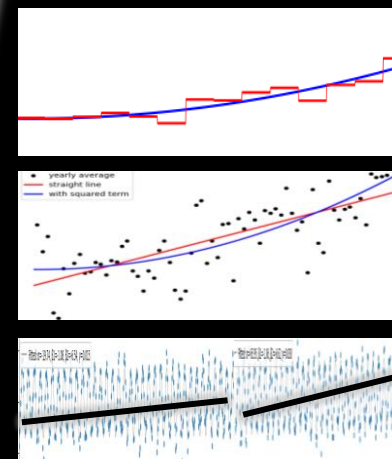
HOLDOUT

... which model to choose?

MLE / prediction
accuracy
Pick the model that
best fits the data

... parameter values?

MLE / prediction
accuracy
Pick the parameter
values that give the
best-fitting model



*linear
step*

linear ... $+ \gamma \epsilon$
quadratic ... $+ \gamma \epsilon + \delta \epsilon^2$

piecewise linear

Julia Wu

Quadratic speed up gives 1689.2 sum of squared residuals, compared to 1710.2 for linear speed up model.

Riki Miyoshi

Thought at this point: the quadratic model is better than the linear model [in residual squared error], which is obvious given it has larger feature space. However, the quadratic term only reduces the residual sum of squares by 1.2%, which may be negligible or not.

This means we need a quantitative way to measure whether this increase in performance is significant.

What can I say about ...

FITTING

HOLDOUT

... which model to choose?

MLE / prediction
accuracy

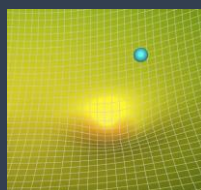
Pick the model that
best fits the data

⚠ **overfitting**

... parameter values?

MLE / prediction
accuracy

Pick the parameter
values that give the
best-fitting model



A “more expressive” model
can always fit better than a
“less expressive” model.

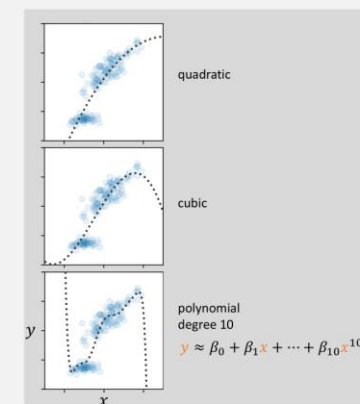
But it may *overfit*

i.e. fit the training data well, but be
useless for new predictions

LECTURE 6

Q. Should we just keep adding more
and more features to our model?

(seeing as the more features we add,
the better we can fit the dataset)



What can I say about ...

FITTING

HOLDOUT

... which model to choose?

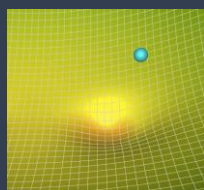
MLE / prediction accuracy
Pick the model that best fits the data

⚠ **overfitting**

Pick the model that best fits
holdout data

... parameter values?

MLE / prediction accuracy
Pick the parameter values that give the best-fitting model



Split the dataset:

- Use part for training (including validation)
- Use a held-out part for evaluation ("testing")

If our model is overfitted to the training data, it'll score badly on holdout data

IA MLRD

Data splits in our experiment



- Training set (1,600)
- Validation (development) set (200) – old "test" set
- Real test set (200) – new today!
- Cross-validation: Combine Train+validation (1800)

What can I say about ...

FITTING

HOLDOUT

... which model to choose?

MLE / ~~prediction~~

~~accuracy~~

Pick the model that
best fits the data

⚠ **overfitting**

Pick the model
that best fits
holdout data

If we choose between
models according to which
fits best

as measured by the model's
 $\Pr(\text{data}; \text{pars})$

... parameter values?

MLE / ~~prediction~~

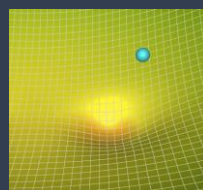
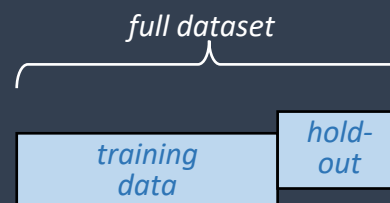
~~accuracy~~

Pick the parameter
values that give the
best-fitting model

(a special case of
model choice)

then it's natural to choose
between *parameter*
values in the same way

since choosing between
parameters θ_1 and θ_2
is really
choosing between
models $\Pr(\cdot; \theta_1)$ and $\Pr(\cdot; \theta_2)$



What can I say about ...

FITTING

HOLDOUT

BEST MODEL:
quadratic

... which model to choose?

MLE / prediction
accuracy
Pick the model that
best fits the data

Pick the model
that best fits
holdout data

FITTED COEFS: with $t_0 = \text{now}$,
Temp ≈ 10.6
 $+ 0.108(t - t_0)$
 $+ 0.0023(t - t_0)^2$
 $+ \dots$

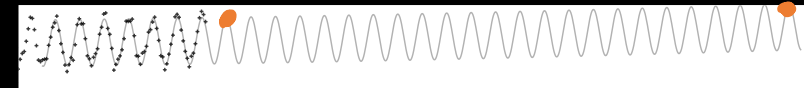
⚠ **overfitting**

... parameter values?

MLE / prediction
accuracy
Pick the parameter
values that give the
best-fitting model

(a special case of
model choice)

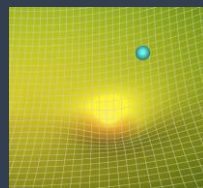
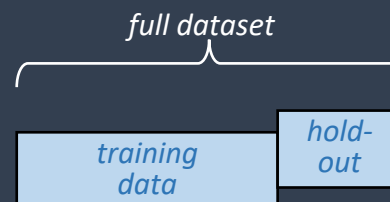
⚠ These methods only rank models & parameters –
how much should we trust the best answer?



PREDICTED TEMP. NEXT JULY:
19.5 °C

PREDICTED TEMP. JULY 2050:
24.0 °C

How confident are you of
these predictions? How
much would you bet on
them?



What can I say about ...

FITTING

HOLDOUT

BAYESIANISM

... which model to choose?

MLE / prediction
accuracy
Pick the model that
best fits the data

⚠ **overfitting**

Pick the model
that best fits
holdout data

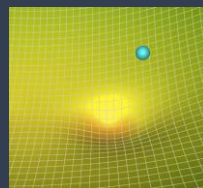
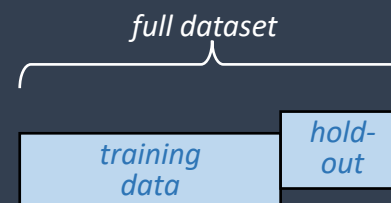
... parameter values?

MLE / prediction
accuracy
Pick the parameter
values that give the
best-fitting model

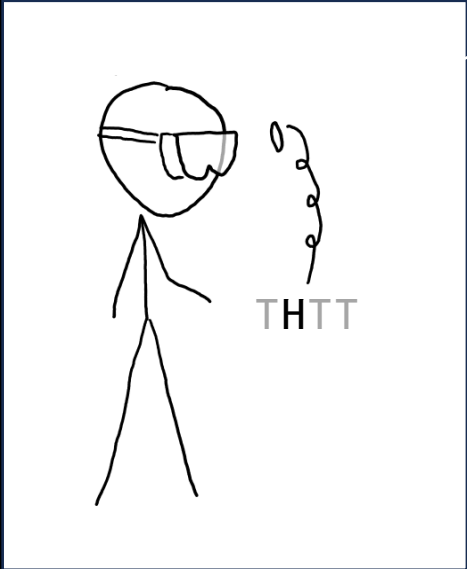
(a special case of
model choice)

Treat unknown
parameters as
random variables
... and this will tell me
how uncertain I should be

⚠ These methods only rank parameter values –
how much should we trust the best answer?



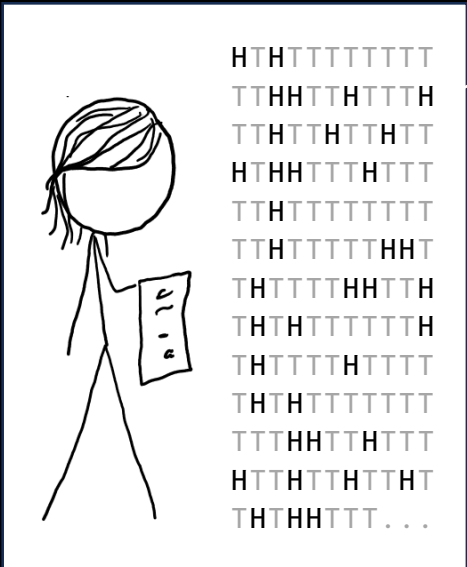
A PARADOX ABOUT THE MEANING OF PROBABILITY



I tossed four coins
and got one head.

Using a $\text{Bin}(n, \theta)$ model, I estimate
the probability of heads is $\hat{\theta} = 25\%$

How much would you bet on this
being borne out by future data?



I tossed four million coins
and got one million heads.

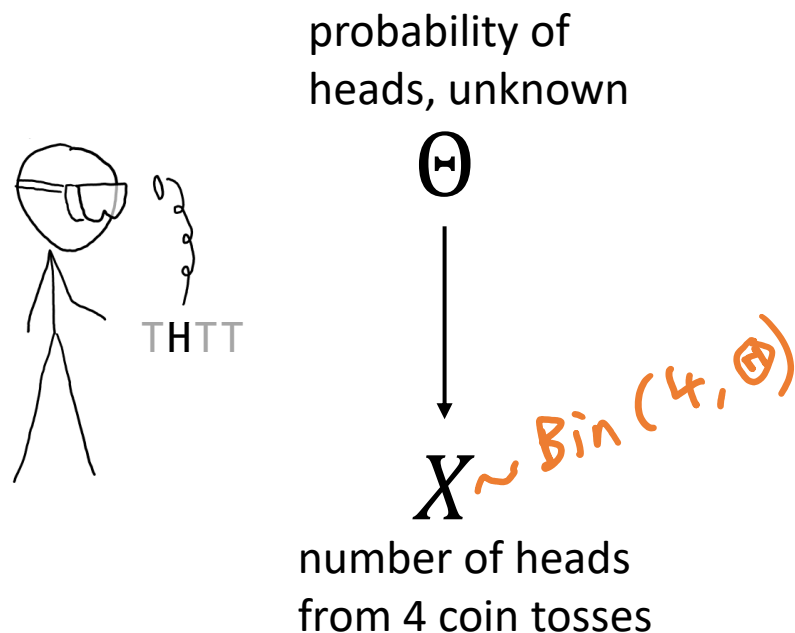
Using a $\text{Bin}(n, \theta)$ model, I estimate
the probability of heads is $\hat{\theta} = 25\%$

and this?

I thought “probability” was how we
measured uncertainty. But it seems
there’s a difference in uncertainty
between these two cases!?

The Bayesianist approach to confidence intervals

By using random variables for unknown quantities,
we can reason about evidence and degree-of-belief.



We don't know the value of Θ .
So let's treat it as a random variable.

We'll start by stating our prior belief.
Say, $\Theta \in \{0.25, 0.5, 0.75\}$, each equally likely.

We observed $X = 1$

We can use Bayes's rule to work out
what this tells us about the unknown
parameter's value ... *prior*

$$P(\Theta = \tfrac{1}{4} \mid X = 1) = \frac{P(\Theta = \tfrac{1}{4}) P(X = 1 \mid \Theta = \tfrac{1}{4})}{P(X = 1)} = \frac{\frac{1}{3} \times \binom{4}{1} (\tfrac{1}{4})^1 (\tfrac{3}{4})^3}{\sum_{\Theta} P(X = 1 \mid \Theta = \Theta) P(\Theta = \Theta)}$$

by Law of Total Probability.

THE BAYESIAN UPDATE

probability of
heads, unknown

 Θ

 $X \sim \text{Bin}(n, \Theta)$

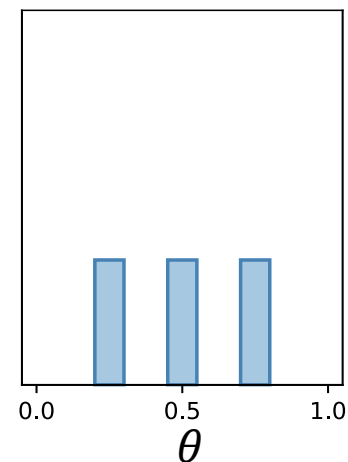
number of heads
from 4 coin tosses



$$\text{prior belief } \mathbb{P}(\Theta = \theta) + \text{data } x \rightarrow \text{posterior belief } \mathbb{P}(\Theta = \theta | X = x)$$

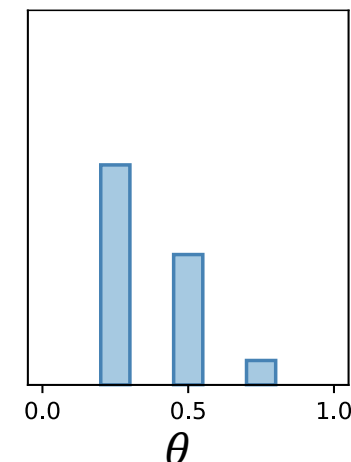
$\mathbb{P}(\Theta = \theta)$ is called
the **prior**.

It expresses our
beliefs *prior* to
having seen the data.



$\mathbb{P}(\Theta = \theta | X = x)$ is
called the **posterior**.

It expresses our
beliefs about Θ
in the light of the data.



THE BAYESIAN UPDATE

$$\begin{array}{ccccc} \text{prior belief} & & + & \text{data} & \rightarrow & \text{posterior belief} \\ \mathbb{P}(\Theta = \theta) & & & x & & \mathbb{P}(\Theta = \theta | X = x) \end{array}$$



$$\Theta \sim U[0,1]$$

to express complete prior ignorance about θ

$$X \sim \text{Bin}(n, \Theta)$$

$$\mathbb{P}(\Theta = \theta | X = x) = \frac{\mathbb{P}(\Theta = \theta) \mathbb{P}(X = x | \Theta = \theta)}{\mathbb{P}(X = x)}.$$

This doesn't help when Θ is continuous:

$$\mathbb{P}(\Theta = \theta) = 0 \text{ for all value of } \theta.$$

It's even worse for models where X is continuous:

$$\text{the denominator is } \mathbb{P}(X = x) = 0. \text{ for all } x.$$

Bayes's rule, done right

For two **discrete** random variables X and Y ,

$$\mathbb{P}(X = x|Y = y) = \frac{\mathbb{P}(X = x)\mathbb{P}(Y = y|X = x)}{\mathbb{P}(Y = y)} \quad \text{when } \mathbb{P}(Y = y) > 0$$

For two **discrete or continuous** random variables X and Y ,

$$\Pr_X(x|Y = y) = \frac{\Pr_X(x) \Pr_Y(y|X = x)}{\Pr_Y(y)} \quad \text{when } \Pr_Y(y) > 0$$



$$\Theta \sim U[0,1]$$

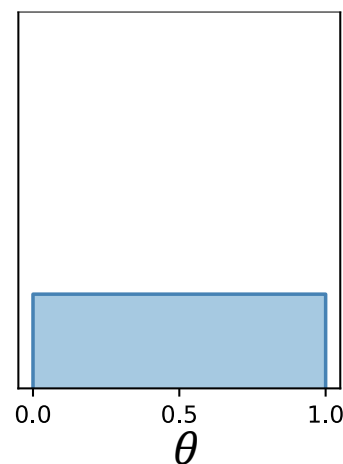
to express complete prior ignorance about θ

$$X \sim \text{Bin}(n, \Theta)$$

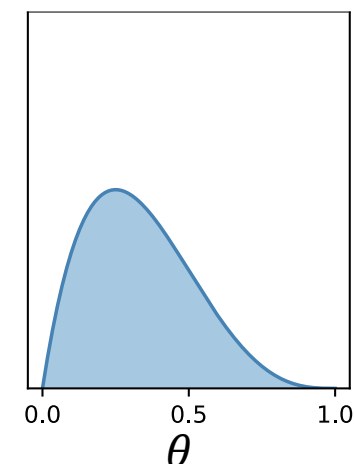
THE BAYESIAN UPDATE

$$\text{prior belief } \mathbb{P}(\Theta = \theta) + \text{data } x \rightarrow \text{posterior belief } \mathbb{P}(\Theta = \theta | X = x)$$

$$\text{prior belief } \Pr_{\Theta}(\theta) + \text{data } x \rightarrow \text{posterior belief } \Pr_{\Theta}(\theta | X = x)$$



THTT





Reverend Thomas
Bayes, 1701–1761

Bayes's rule

$$\Pr_X(x|Y = y) = \frac{\Pr_X(x) \Pr_Y(y|X = x)}{\Pr_Y(y)} \quad \text{when } \Pr_Y(y) > 0$$



Bayesianism

Whenever there's an unknown parameter,
you should express your uncertainty about it
by treating it as a random variable.



Prob of
heads



heads
in coin
toss expt.