

§1.5 Better notation for likelihood

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

using maximum
likelihood estimation

Step 1.5. Find an expression
for the likelihood



$\Pr_{\text{RNG}}(\text{val}) = \text{likelihood that RNG produces output val}$

The *likelihood function* for a random variable X is written $\Pr_X(x)$ and defined as

$\Pr_X(x) = \mathbb{P}(X = x)$ in the case where X is discrete

and as

$\Pr_X(x) = \text{pdf}(x)$ in the case where X is continuous with prob. density function $\text{pdf}(x)$

The likelihood for RNG X of value x

For parameterized random variables, write

$\Pr_X(x; \theta)$ or $\Pr_x(x | \theta)$ or $\Pr_x(x)$

$\Pr_{\text{RNG}}(\text{val}) = \text{likelihood that RNG produces output val}$

Example: $\Pr_{X+Y}(0.2)$

I have a RNG function named "X+Y".
(It calls the X RNG and the Y RNG and
adds their outputs together.)

What's the chance that calling "X+Y"
gives output 0.2?

Let's rewrite our rules for likelihood using the Pr notation.

Pairs of random variables:

$\Pr_{X,Y}(x, y)$ is called the *joint likelihood* of X and Y

Independent random variables:

$$\Pr_{X,Y}(x, y) = \Pr_X(x) \Pr_Y(y)$$

For discrete X and Y

$$\begin{aligned}\Pr_{X,Y}(x, y) &= \mathbb{P}(X=x, Y=y) \\ &= \mathbb{P}(X=x) \mathbb{P}(Y=y) \\ &= \Pr_X(x) \Pr_Y(y)\end{aligned}$$

Independent identically-distributed (IID) sample from X :

$$\Pr(x_1, \dots, x_n) = \Pr_X(x_1) \times \dots \times \Pr_X(x_n)$$

Sequential generation of X , then Y based on X :

$$\Pr_{X,Y}(x, y) = \Pr_X(x) \Pr_Y(y; x)$$

TECHNICAL ASIDE

It's the joint likelihood that *defines* the distribution, by specifying the probability of joint events $(X, Y) \in S$.

For discrete random variables,

$$\mathbb{P}(X = x, Y = y) = \Pr_{X,Y}(x, y)$$

$$\mathbb{P}((X, Y) \in S) = \sum_{(x,y) \in S} \Pr_{X,Y}(x, y)$$

For continuous random variables,

$$\mathbb{P}((X, Y) \in S) = \int_{(x,y) \in S} \Pr_{X,Y}(x, y) dx dy$$

See IA Probability lecture 6

Let's rewrite the rules
of probability in
likelihood notation

If X and Y are independent
random variables:

$$\Pr_{X,Y}(x, y) = \Pr_X(x) \Pr_Y(y)$$

Sum rule for random
variables:

$$\Pr_X(x) = \sum_y \Pr_{X,Y}(x, y)$$

$$\Pr_X(x) = \int_y \Pr_{X,Y}(x, y) dy$$

Rules of Probability

Understand what is meant by *sample space*, written Ω , and know that $\mathbb{P}(\Omega) = 1$. Be able to reason about probabilities of events with Venn diagrams. Know the core definitions and laws ...

Conditional probability, or equivalently the chain rule:

$$\mathbb{P}(A | B) = \frac{\mathbb{P}(A, B)}{\mathbb{P}(B)} \quad \text{if } \mathbb{P}(B) > 0$$

$$\mathbb{P}(B, A) = \mathbb{P}(B) \mathbb{P}(A | B) \quad (\text{chain rule})$$

If A and B are independent:

If A and B are independent events:

$$\mathbb{P}(A, B) = \mathbb{P}(A) \mathbb{P}(B)$$

$$\mathbb{P}(A | B) = \mathbb{P}(A)$$

Sum rule, and the law of total probability, for events $\{B_1, B_2, \dots\}$ that partition the sample space (i.e. for events that are mutually exclusive and where $\bigcup_i B_i = \Omega$):

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A, B_i)$$

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A | B_i) \mathbb{P}(B_i) \quad (\text{law of total probability})$$

Bayes's rule:

$$\mathbb{P}(A | B) = \frac{\mathbb{P}(B | A) \mathbb{P}(A)}{\mathbb{P}(B)} \quad \text{if } \mathbb{P}(B) > 0.$$

Let's rewrite MLE in Pr notation

Maximum Likelihood Estimation, again

If we've seen an outcome x , and we've proposed a probability model X , and if its distribution involves some unknown parameters θ ,

the *maximum likelihood estimator* for θ is

$$\hat{\theta} = \arg \max_{\theta} \Pr_X(x ; \theta)$$

- x could be discrete or continuous
- x could be a single observation or a dataset with many observations

The point of the likelihood notation is so that we can write down a single equation and have it cover all these cases.

$\Pr_{\text{RNG}}(\text{val}) = \text{likelihood that RNG produces output val}$

Brain teaser

Consider this random variable X :

$$X = \begin{cases} \text{cat} & \text{with prob. } 1/6 \\ \text{stoat} & \text{with prob. } 3/6 \\ \text{dog} & \text{with prob. } 2/6 \end{cases}$$

What is $\Pr_X(X)$?

Define the function f via a lookup table,

$f(\text{cat}) = \text{meh}$
 $f(\text{stoat}) = \text{cute}$
 $f(\text{dog}) = \text{meh}$

Q1. What is $\Pr_{f(X)}(\text{meh})$?

Q2. What is $\Pr_X$? What is its type?

Q3. What is $f(X)$? What is its type?

Easy WRONG answer: $\Pr_X(X) = \mathbb{P}(X=X) = 1$.

correct answer: $\Pr_X$ is a function (see Q2), call it g .
we want $g(X)$, which is a random variable (see Q3).

$$\left. \begin{aligned} g(\text{cat}) &= 1/6 \\ g(\text{stoat}) &= 3/6 \\ g(\text{dog}) &= 2/6 \end{aligned} \right\} \text{ so } g(X) = \begin{cases} 1/6 & \text{w. prob. } 1/6 \\ 3/6 & \text{w. prob. } 3/6 \\ 2/6 & \text{w. prob. } 2/6 \end{cases}$$

Q1: $\Pr_{f(X)}(\text{meh}) = \frac{1}{2}$, because to get $f(X) = \text{meh}$ we need $X = \text{cat}$ or $X = \text{dog}$, $\text{prob.} = \frac{1}{6} + \frac{2}{6} = \frac{1}{2}$.

Q2: $\Pr_X$ is a function

x	$\Pr_X(x)$
cat	1/6
stoat	3/6
dog	2/6

Q3: $f(X)$ is random: $f(X) = \begin{cases} \text{meh} & \text{w. prob. } 1/2 \\ \text{cute} & \text{w. prob. } 1/2 \end{cases}$

There are standard numerical random variables that you should know:

§1.2

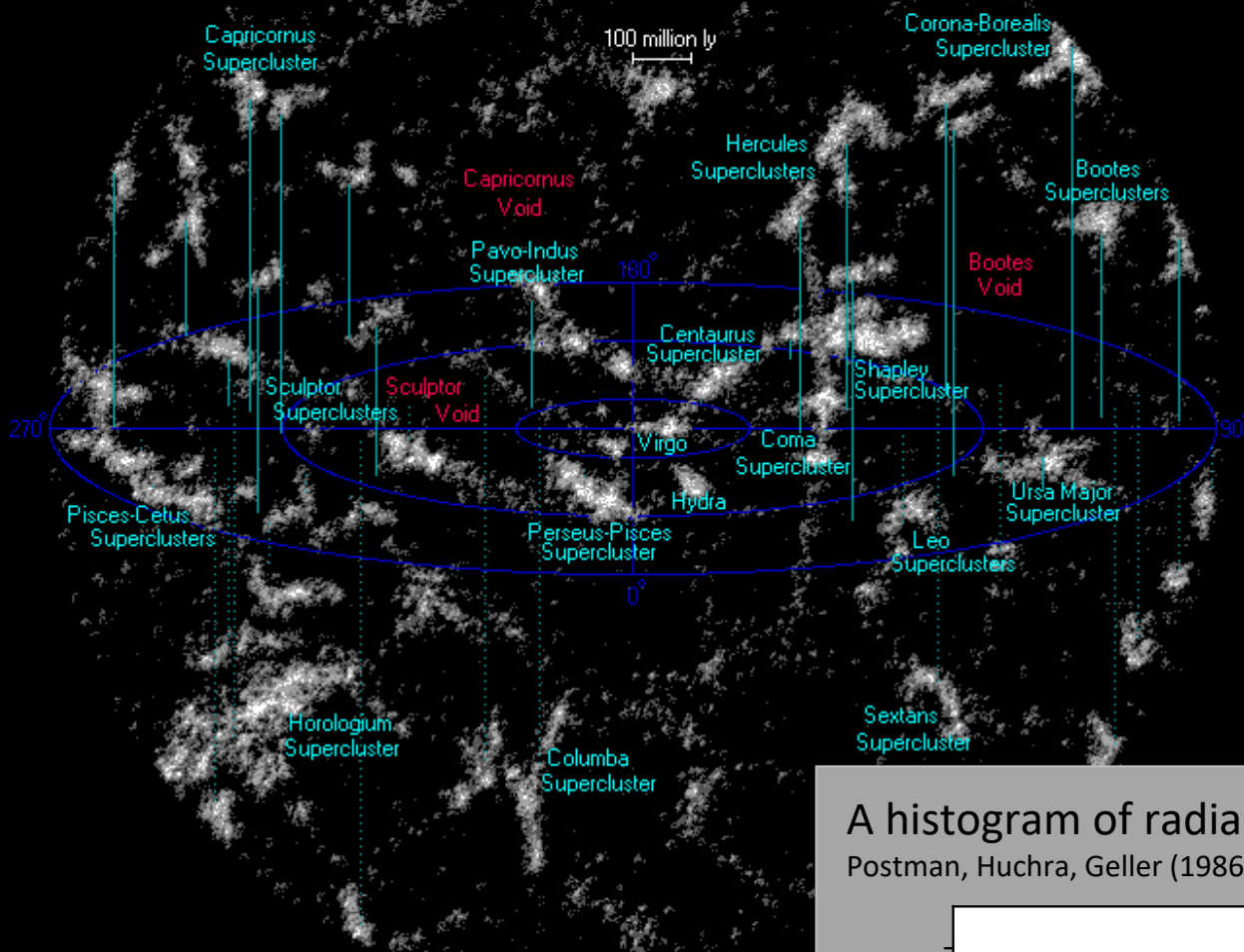
DISCRETE RANDOM VARIABLES

Binomial $X \sim \text{Bin}(n, p)$	$\mathbb{P}(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$ $x \in \{0, 1, \dots, n\}$	For count data, e.g. number of heads in n coin tosses
Poisson $X \sim \text{Pois}(\lambda)$	$\mathbb{P}(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$ $x \in \{0, 1, \dots\}$	For count data, e.g. number of buses passing a spot
Categorical $X \sim \text{Cat}([p_1, \dots, p_k])$	$\mathbb{P}(X = x) = p_x$ $x \in \{1, \dots, k\}$	For picking one of a fixed number of choices

CONTINUOUS RANDOM VARIABLES

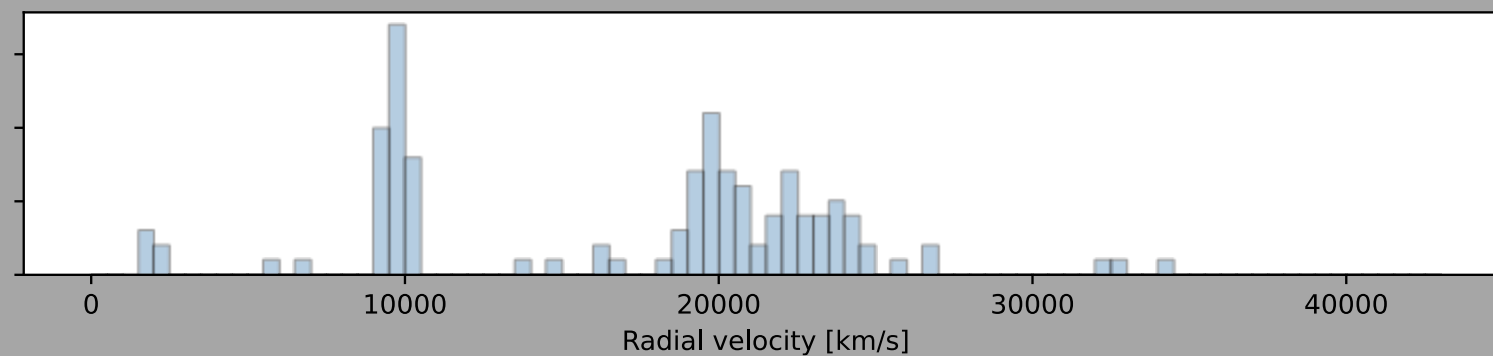
Uniform $X \sim U[a, b]$	$\text{pdf}(x) = \frac{1}{b - a}$ $x \in [a, b]$	A uniformly-distributed floating point value
Normal / Gaussian $X \sim N(\mu, \sigma^2)$	$\text{pdf}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$ $x \in \mathbb{R}$	For data about magnitudes, e.g. temperature or height
Pareto $X \sim \text{Pareto}(\alpha)$	$\text{pdf}(x) = \alpha x^{-(\alpha+1)}$ $x \geq 1$	For data about “cascade” magnitudes, e.g. forest fires
Exponential $X \sim \text{Exp}(\lambda)$	$\text{pdf}(x) = \lambda e^{-\lambda x}$ $x > 0$	For waiting times, e.g. time until next bus
Beta $X \sim \text{Beta}(a, b)$	$\text{pdf}(x) \propto x^{a-1} (1 - x)^{b-1}$ $x \in (0, 1)$	Arises in Bayesian inference

Image: Richard Powell, Creative Commons Attribution-Share Alike 2.5 Generic license.

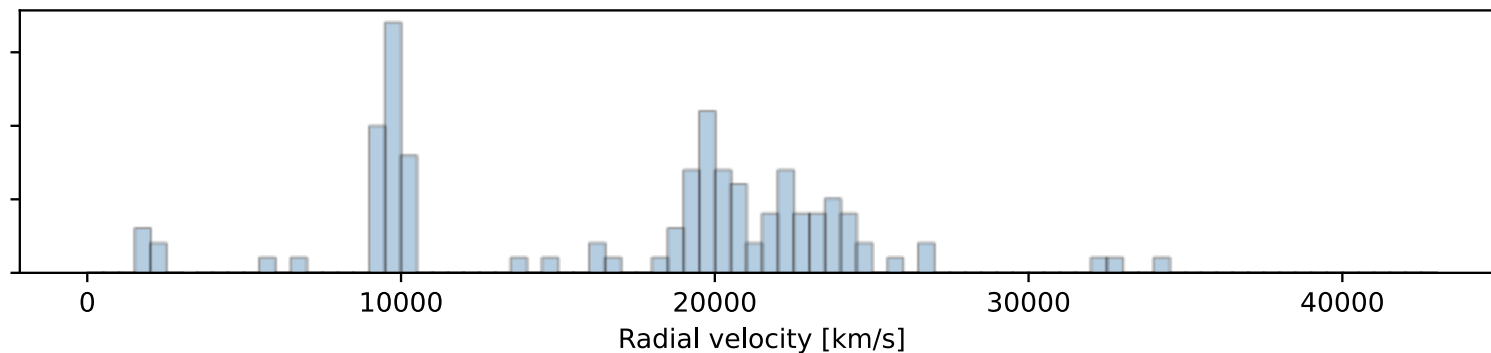


A histogram of radial velocities of 120 galaxies in the Corona Borealis region

Postman, Huchra, Geller (1986)

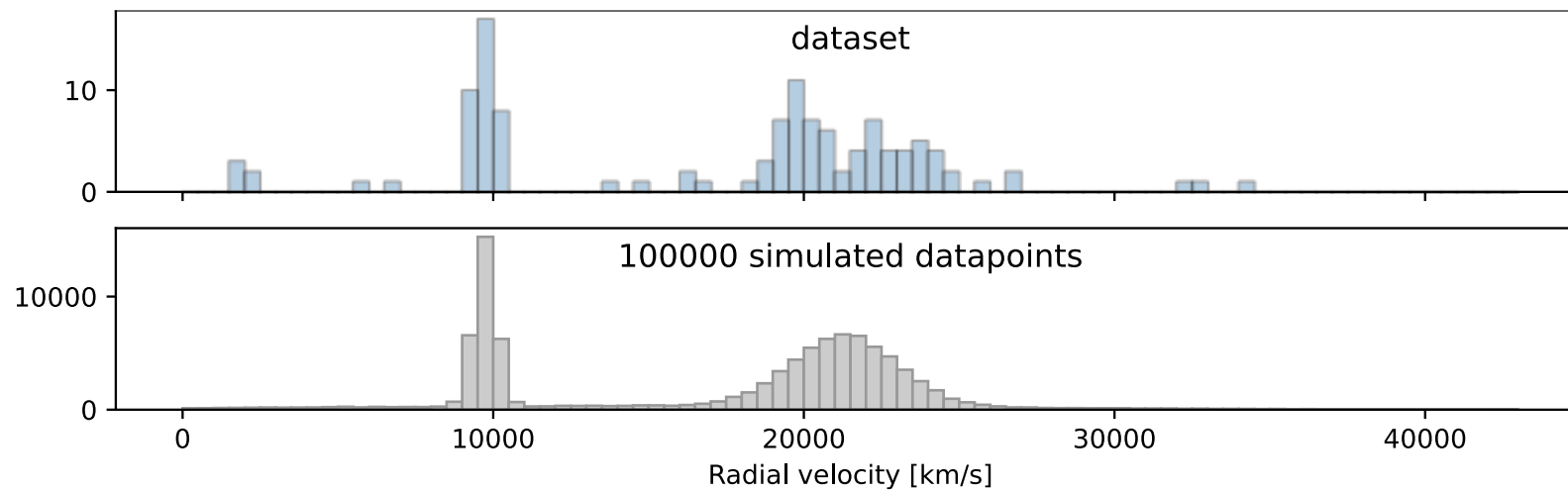


A histogram of radial velocities of 120 galaxies



How might you complete this code?

```
def rgalaxy(...):  
    # TODO: return a single random galaxy speed  
  
def rgalaxies(size):  
    return [rgalaxy(...) for _ in range(size)]
```



```
def rgalaxy(p,μ,σ):
    k = np.random.choice([0,1,2], p=p)
    x = np.random.normal(loc=μ[k], scale=σ[k])
    return x

def rgalaxies(size, p,μ,σ):
    return [rgalaxy(p,μ,σ) for _ in range(size)]
```

```
p = [0.28, 0.54, 0.18]
μ = [9740, 21300, 15000]
σ = [340, 1700, 10600]
```

How do I know if I've got the right model?



*"All models are wrong
but some are useful"*

George Box

... so don't be shy when building your own.
There *is* no "correct model".

EXERCISE 1.1.3. Write this model in random variable notation.

$$K \sim \text{cat}(p)$$

$$x \sim N(\mu_K, \sigma_K^2)$$

parameters are $p \in (0,1)^3$ s.t. $p_1 + p_2 + p_3 = 1$
 $\mu \in \mathbb{R}^3$ $\sigma \in \mathbb{R}_{>0}^3$

EXERCISE 1.7.4. What's the likelihood, $\Pr(\text{data}; \text{pars})$?

$$\text{data} = x_1, \dots, x_n \quad \text{pars} = p, \mu, \sigma$$

$$\Pr_{K,X}(k, x; p, \mu, \sigma) = \Pr_K(k) \Pr_X(x; k)$$

$$= p_k \frac{1}{\sqrt{2\pi}\sigma_k} e^{-(x - \mu_k)^2 / (2\sigma_k^2)}$$

$$\Pr_X(x) = \sum_k p_k \frac{1}{\sqrt{2\pi}\sigma_k} \dots$$

$$\Pr(\text{data}; p, \mu, \sigma) = \prod_{i=1}^n \sum_k p_k \dots$$

↙ expr with x_i

EXERCISE 1.7.4. How would you fit this model? **MLE**

Stick the likelihood function into a numerical optimizer, and maximize it over p, μ, σ !

Watch out for constraints: $p \in (0,1)^3$ with $p_1 + p_2 + p_3 = 1$, $\sigma > 0$.

(We can handle these using a parameter transformation — e.g. use softmax for p .)

There are standard numerical random variables that you should know:

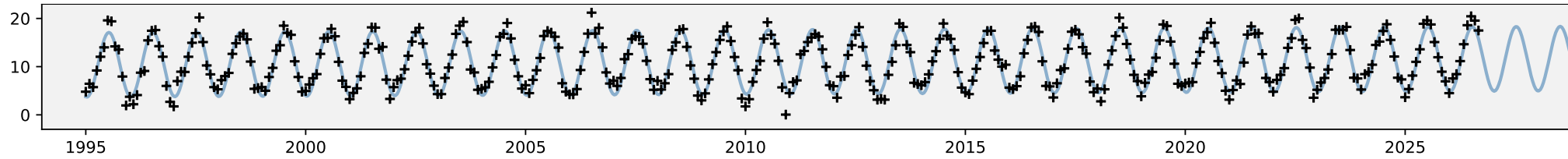
DISCRETE RANDOM VARIABLES

Binomial $X \sim \text{Bin}(n, p)$	$\mathbb{P}(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$ $x \in \{0, 1, \dots, n\}$	For count data, e.g. number of heads in n coin tosses
Poisson $X \sim \text{Pois}(\lambda)$	$\mathbb{P}(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$ $x \in \{0, 1, \dots\}$	For count data, e.g. number of buses passing a spot
Categorical $X \sim \text{Cat}([p_1, \dots, p_k])$	$\mathbb{P}(X = x) = p_x$ $x \in \{1, \dots, k\}$	For picking one of a fixed number of choices

CONTINUOUS RANDOM VARIABLES

Uniform $X \sim U[a, b]$	$\text{pdf}(x) = \frac{1}{b - a}$ $x \in [a, b]$	A uniformly-distributed floating point value
Normal / Gaussian $X \sim N(\mu, \sigma^2)$	$\text{pdf}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$ $x \in \mathbb{R}$	For data about magnitudes, e.g. temperature or height
Pareto $X \sim \text{Pareto}(\alpha)$	$\text{pdf}(x) = \alpha x^{-(\alpha+1)}$ $x \geq 1$	For data about “cascade” magnitudes, e.g. forest fires
Exponential $X \sim \text{Exp}(\lambda)$	$\text{pdf}(x) = \lambda e^{-\lambda x}$ $x > 0$	For waiting times, e.g. time until next bus
Beta $X \sim \text{Beta}(a, b)$	$\text{pdf}(x) \propto x^{a-1} (1 - x)^{b-1}$ $x \in (0, 1)$	Arises in Bayesian inference

EXERCISE



Take the Cambridge monthly weather data from 1995 to the present.
[Code to load the data is in the slides for lecture 1.]

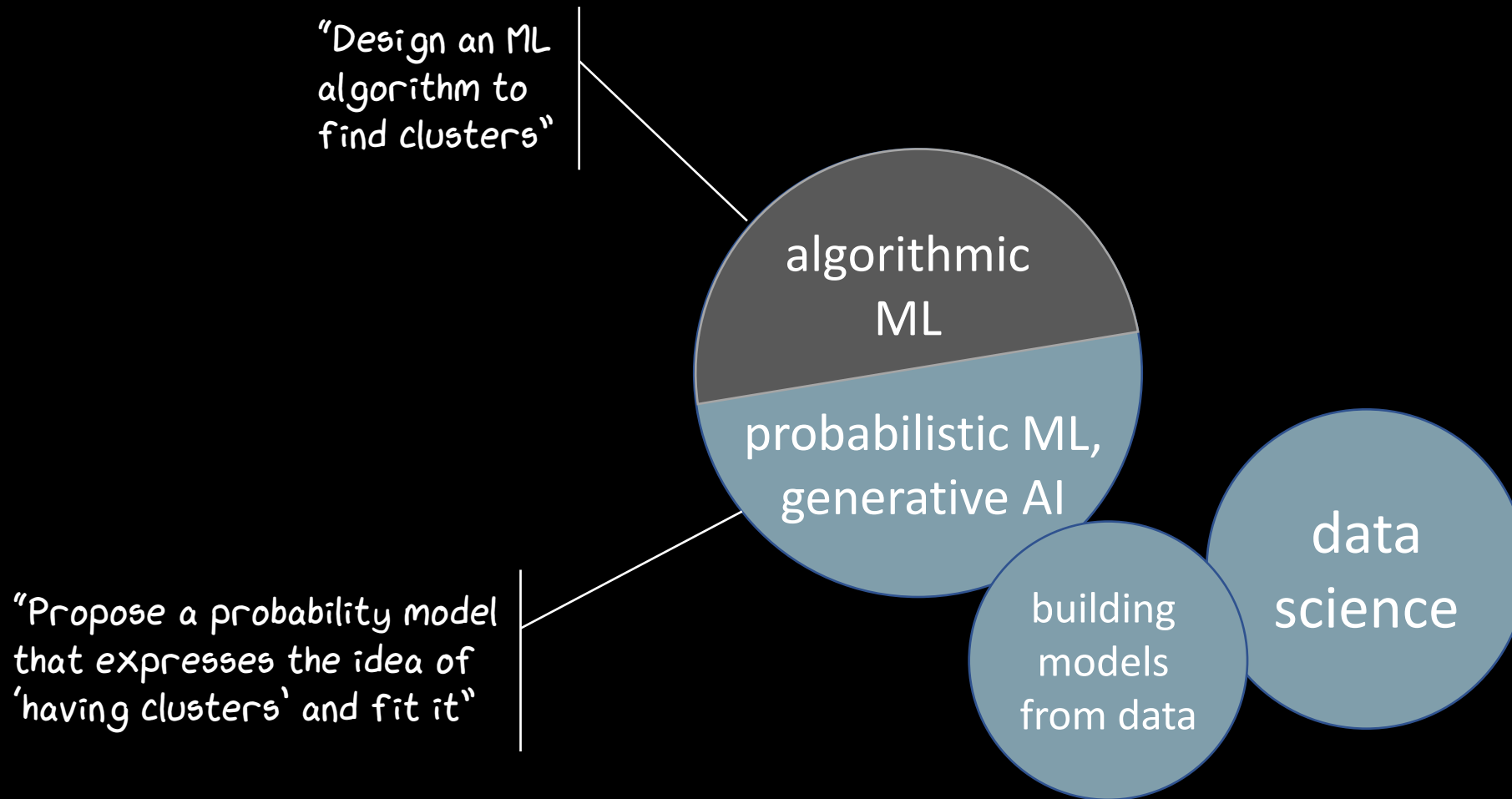
Estimate γ in the following model:

```
def rtemp(t,  $\alpha$ ,  $\phi$ , c,  $\gamma$ ,  $\sigma$ ):  
    pred = c +  $\alpha$  * sin(2* $\pi$ *(t+ $\phi$ )) +  $\gamma$ *t  
    return random.normal(loc=pred, scale= $\sigma$ )
```

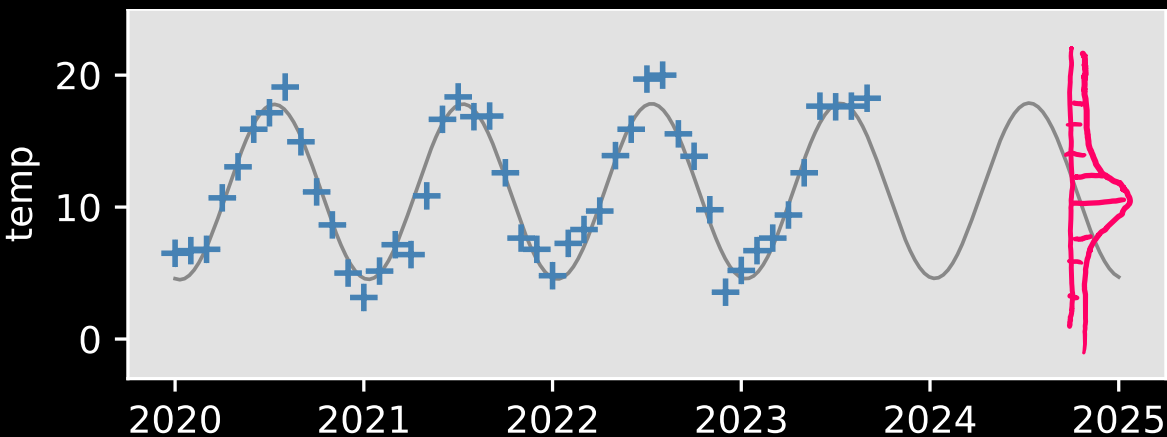
1. Write the model in random variable notation
2. Find the likelihood, $\Pr(\text{data}; \text{pars})$
3. Fit it numerically with `scipy.optimize.fmin`

Submit your answer (in $^{\circ}\text{C}/\text{century}$) on Moodle before Friday's lecture.

We just invented a cluster-finding algorithm!
Our fitted parameters (p, μ, σ) are the cluster sizes and shapes.



supervised learning

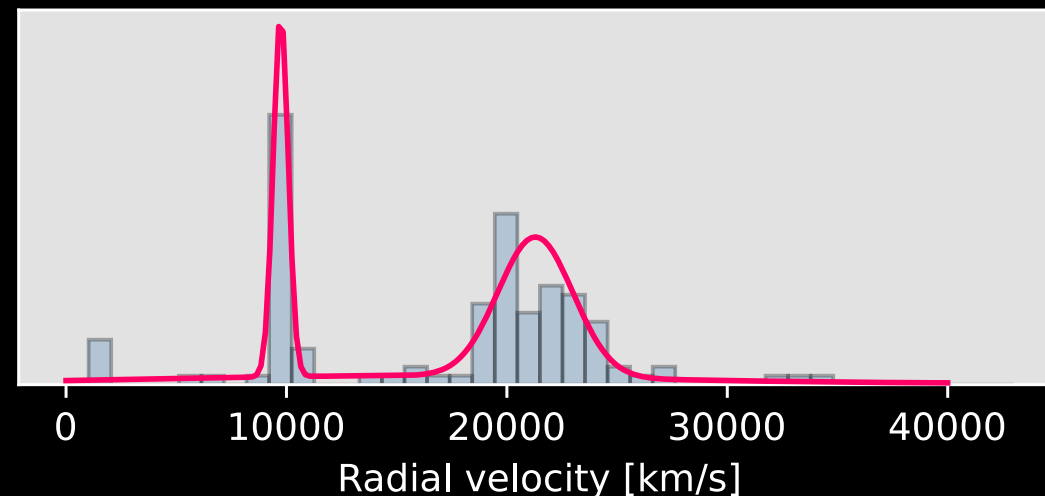


Given a dataset of (t_i, temp_i) pairs, $i \in \{1, \dots, n\}$,

I'd like to predict temp as a function of t .

I'd like to fit a probability model for Temp, where the parameters of the distribution depend on t .

unsupervised learning

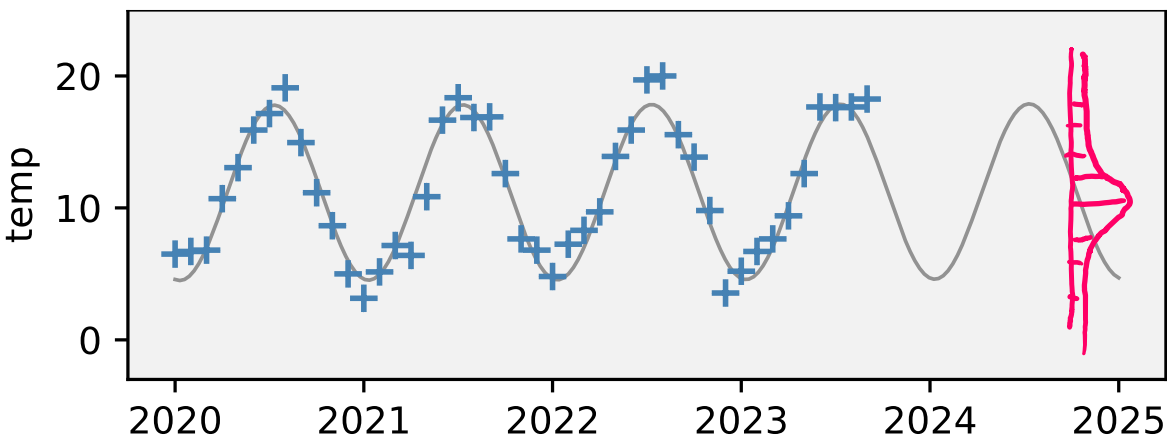


Given a dataset $[x_1, \dots, x_n]$ of galaxy speeds,

I'd like to identify the clusters.

I'd like to fit a probability model for speed X , using a model that has clusters.

supervised learning



Given a dataset $(x_1, y_1), \dots, (x_n, y_n)$ where y_i is the label in record i and x_i is the predictor variable or variables:

0. Propose a probability model

$Y \sim \dots$ or $X \sim \dots$

1. Write out the likelihood for a single datapoint:

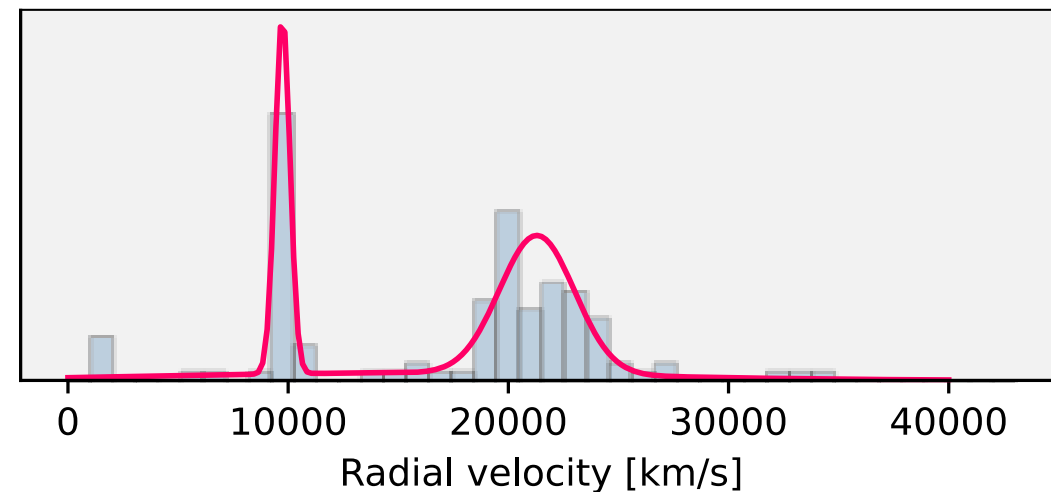
$\Pr_Y(y; x, \theta)$ or $\Pr_X(x; \theta)$

2. Model the dataset as independent observations and write out the likelihood for the full dataset:

$\Pr(\text{dataset}) = \prod_{i=1}^n \Pr_Y(y_i; x_i, \theta)$ or $\prod_{i=1}^n \Pr_X(x_i; \theta)$

3. Learn θ using maximum likelihood estimation

unsupervised learning



Given a dataset $x_1, \dots, x_n$:

Terminology for supervised learning

station	yyyy	mm	t	af	rain	sun	tmin	tmax	temp
Cambridge	1985	1	1985.00	23	37.3	40.7	-2.2	3.4	0.6
Cambridge	1985	2	1985.08	13	14.6	79	-1.9	4.9	1.5
Cambridge	1985	3	1985.16	10	45.8	97.8	1.1	8.7	4.9
⋮									

called the PREDICTOR variable,
or the FEATURE,
or the COVARIATE

called the RESPONSE,
or the LABEL variable,
or the GROUND TRUTH.

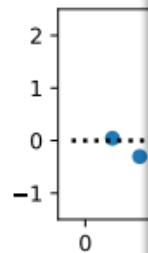
- Here the response is real-valued, so we call it REGRESSION.
- If the response were categorical, we'd call it CLASSIFICATION.

For Friday: have a look at the mock exam, try fitting the climate data.

COMPUTER SCIENCE TRIPOS Part IB – mock – Paper 6

1 Foundations of Data Science (DJW)

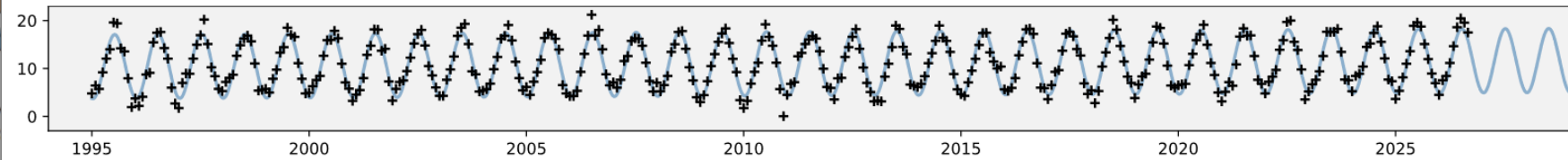
- (a) A 0/1 signal is being transmitted. The transmitted signal at timeslot $i \in \{1, \dots, n\}$ is $x_i \in \{0, 1\}$, and we have been told that this signal starts at 0 and then flips to 1, i.e. the $x_i = 1_{i > \theta}$. The value of this θ is unknown. The received signal in timeslot i is y_i , where ε is known.



where ε is known.

- (i) Given received signals $(y_1, \dots, y_n)$, estimate θ . Explain your working.

EXERCISE



Take the Cambridge monthly weather data from 1995 to the present.
[Code to load the data is in the slides for lecture 1.]

Estimate γ in the following model:

```
def rtemp(t,  $\alpha$ ,  $\phi$ ,  $c$ ,  $\gamma$ ,  $\sigma$ ):  
    pred =  $c + \alpha * \sin(2*\pi*(t+\phi)) + \gamma*t$   
    return random.normal(loc=pred, scale= $\sigma$ )
```

[5 marks]