§11 first page

Markov chains

§12.1

Learning a random process

Random process

a sequence $X_0, X_1, X_2, \ldots$ of random variables, typically not independent

$$\Pr(x_0, x_1, \ldots, x_n) = \Pr_{X_0}(x_0)\, \Pr_{X_1}(x_1|x_0)\, \Pr_{X_2}(x_2|x_0, x_1) \times \cdots \times \Pr_{X_n}(x_n|x_0 \cdots x_{n-1})$$

$\mathbb{P}(A \text{ and } B \text{ and } C) = \mathbb{P}(A)\,\mathbb{P}(B|A)\,\mathbb{P}(C|B,A)$

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

$= \cdots$

by the chain rule for probability

If we have a dataset of sequences, and we have a probability model (e.g. a RNN or a Transformer neural network) that computes $\Pr_{X_i}(x_i|x_0 \cdots x_{i-1})$, then we can fit it using maximum likelihood estimation.

Markov chain

a random process in which each X_i is generated based **only** on the preceding value X_{i-1}

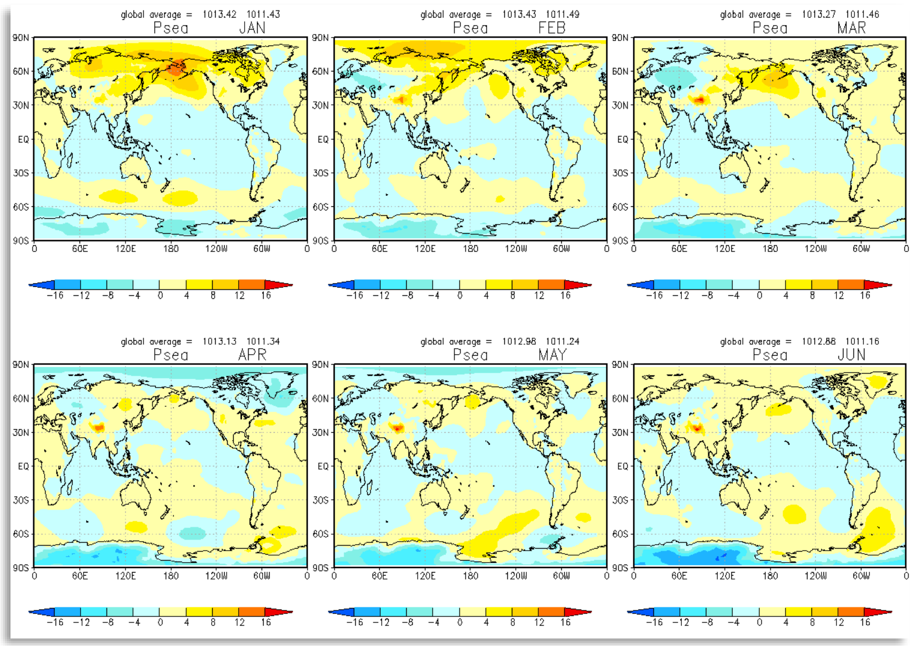
$$X_0 \to X_1 \to X_2 \to \cdots$$

$$\Pr(x_0, x_1, \ldots, x_n) = \Pr_{X_0}(x_0)\, \Pr_{X_1}(x_1|x_0)\, \Pr_{X_2}(x_2|x_1) \times \cdots \times \Pr_{X_n}(x_n|x_{n-1})$$

Because X_2 is generated based only on X_1,

$\Pr_{X_2}(x_2|x_0, x_1) = \Pr_{X_2}(x_2|x_1)$.

Applications of Markov chains: dynamical systems



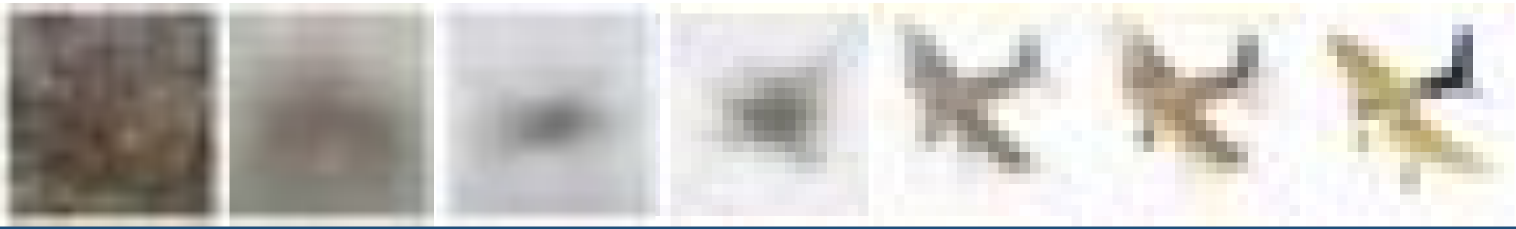
Let X_t be the full state of the system at time t. We'd like to use historical data to learn the dynamics $(X_t | X_{t-1} = x_{t-1})$, so that we can simulate it.

Applications of Markov chains: stable diffusion

Given an image, create a sequence with progressively more and more noise, until we get pure noise.
Do this for many images, to create a training dataset of sequences.

Reverse the sequences. Train a Markov chain to learn the dynamics $(X_t | X_{t-1} = x)$.

If we apply these dynamics to a new pure-noise image, we will generate a novel image.

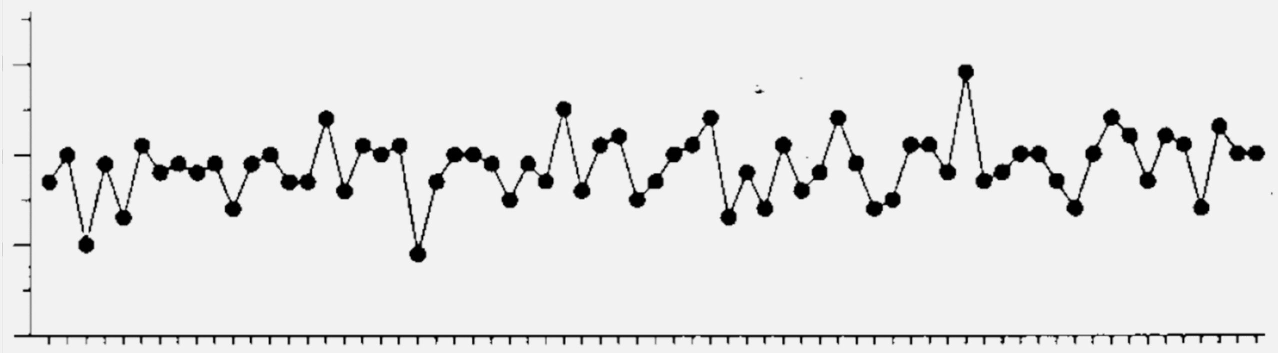


Example 12.1.1: fitting a Markov model

Let $[x_0, x_1, \ldots, x_n]$ be a time series which we believe is generated by

$$X_{i+1} = a + b\,X_i + N(0, \sigma^2).$$

Estimate a, b, and σ using maximum likelihood estimation.



$$\Pr(x_0 x_1 \cdots x_n) = \Pr(x_0) \prod_{i=1}^{n} \Pr(x_i \mid x_{i-1})$$

$$= \Pr(x_0) \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i - (a + b x_{i-1}))^2 / 2\sigma^2}$$

since $X_i \sim N(a + bX_i, \sigma^2)$.

The question tells us nothing at all about the dist. of X_0

$$= ??? \prod_{i=1}^{n} \cdots$$

$$\Pr(x_0, x_1, \cdots, x_n) = ??? \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_i - (a + b x_{i-1}))^2 / 2\sigma^2}$$

To fit our model, we need to maximize this expression over a, b, σ.

predictor	response
x_0	x_1
x_1	x_2
x_2	x_3
$\vdots$	$\vdots$
x_{n-1}	x_n

But this is exactly the same maximization as for the supervised learning task of predicting x_i given x_{i-1} using the model $X_i \sim a + bx_{i-1} + N(0, \sigma^2)$

It's simple to fit using sklearn.

Autoregressive modelling

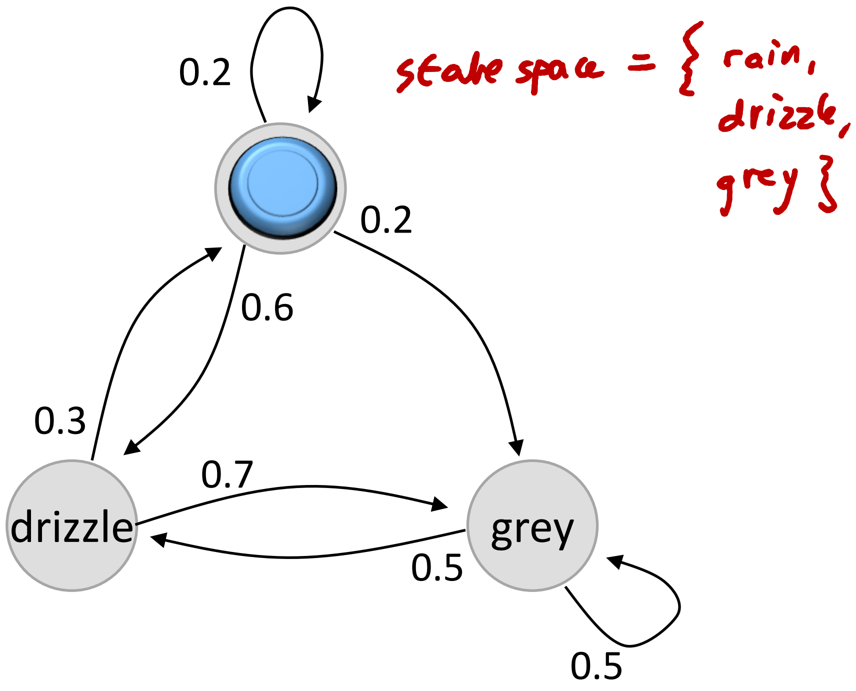
This is a regression (i.e. supervised learning with numerical response).
It's called 'auto' because we're predicting x using x itself as a predictor.

§11.2

Calculations with Markov chains

There are three ways to specify a Markov chain model.

STATE SPACE DIAGRAM



TRANSITION PROBABILITY MATRIX

$$P = \begin{array}{c} \text{rain} \\ \text{drizzle} \\ \text{grey} \end{array} \overset{\text{rain} \quad \text{drizzle} \quad \text{grey}}{\begin{bmatrix} .2 & .6 & .2 \\ .3 & 0 & .7 \\ 0 & .5 & .5 \end{bmatrix}}$$

$$P_{ij} = \mathbb{P}\left(\begin{array}{c}\text{next state} \\ \text{is } j\end{array} \middle| \begin{array}{c}\text{in state} \\ i\end{array} \right)$$

If the state space is $\mathbb{R}$ we can't write out the full matrix so we instead specify $\Pr_{X_t}(x_t | X_{t-1} = x_{t-1})$

CAUSAL DIAGRAM

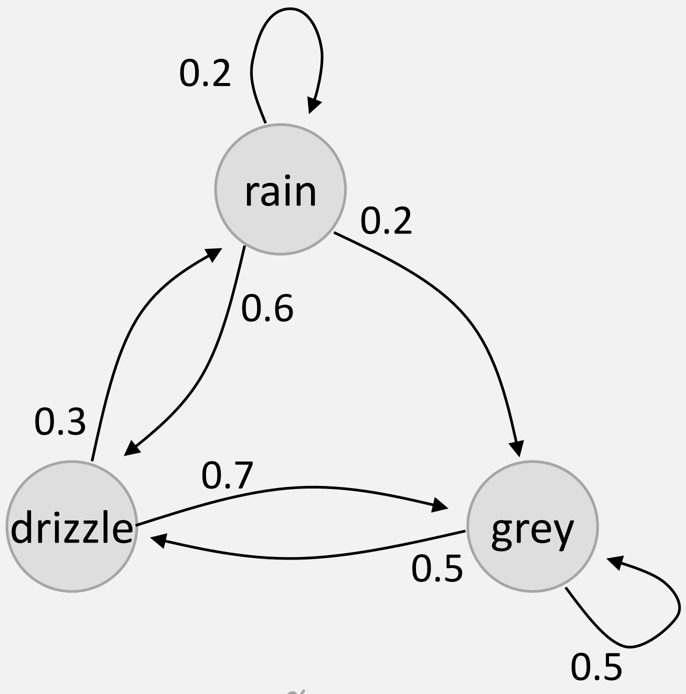
Each X_i is generated based only on the preceding state X_{i-1}:

$$X_1 \to X_2 \to X_3 \to \cdots$$

Example 11.2.1
(Multi-step transition probabilities)

If it's grey today, what's the chance of rain two days from now?

$$X_1 \to X_2 \to X_3 \to \cdots$$



$$P = \begin{array}{c} \\ \text{rain} \\ \text{drizzle} \\ \text{grey} \end{array} \begin{array}{c} \begin{array}{ccc} \text{rain} & \text{drizzle} & \text{grey} \end{array} \\ \begin{bmatrix} .2 & .6 & .2 \\ .3 & 0 & .7 \\ 0 & .5 & .5 \end{bmatrix} \end{array}$$

r = rain
g = grey
d = drizzle.

$$\mathbb{P}(X_2 = r \mid X_0 = g)$$

$$= \sum_x \mathbb{P}(X_2 = r \mid X_1 = x, X_0 = g)\, \mathbb{P}(X_1 = x \mid X_0 = g)$$

Law of Total Probability with baggage $\{X_0 = g\}$

$$= \sum_x \mathbb{P}(X_2 = r \mid X_1 = x)\, \mathbb{P}(X_1 = x \mid X_0 = g)$$

since X_2 is generated based only on X_1, the state at time 0 is irrelevant once we know the state at time 1.

a.k.a. Memorylessness.

$$= \sum_x P_{xr} P_{gx} = \sum_x P_{gx} P_{xr} = [P^2]_{gr}$$

Laws of probability that can help when working with Markov chains

Law of Total Probability

$$\mathbb{P}(A=a) = \sum_b \mathbb{P}(A=a \mid B=b)\,\mathbb{P}(B=b)$$

Law of Total Probability with baggage $\{C=c\}$

$$\mathbb{P}(A=a \mid C=c) = \sum_b \mathbb{P}(A=a \mid B=b, C=c)\,\mathbb{P}(B=b \mid C=c)$$

Bayes's rule

$$\mathbb{P}(A=a \mid B=b) = \frac{\mathbb{P}(A=a)\,\mathbb{P}(B=b|A=a)}{\mathbb{P}(B=b)}$$

Bayes's rule with baggage $\{C=c\}$

$$\mathbb{P}(A=a \mid B=b, C=c) = \frac{\mathbb{P}(A=a|C=c)\,\mathbb{P}(B=b|A=a, C=c)}{\mathbb{P}(B=b|C=c)}$$

Definition of independence

If A and B are independent then

$$\mathbb{P}(A=a \mid B=b) = \mathbb{P}(A=a)$$

Definition of conditional independence

If A and B are conditionally independent given $\{C=c\}$ then

$$\mathbb{P}(A=a \mid B=b, C=c) = \mathbb{P}(A=a \mid C=c)$$

Calculating with Markov Chains

The chain is memoryless

$$X_0 \to X_1 \to \cdots$$

i.e. each item is generated based only on the previous item

Whenever we're doing calculations with Markov chains, we have to wrangle our expression into a form where we can use memorylessness (plus the transition probability matrix).

Often, this will involve conditioning using the Law of Total Probability.

The memorylessness theorem:
conditional on the present,
the future is independent of the past.

future, present, past

$$\mathbb{P}(X_3 = x_3 \mid X_2 = x_2, X_1 = x_1, X_0 = x_0) = \mathbb{P}(X_3 = x_3 \mid X_2 = x_2)$$

future, present, past

$$\mathbb{P}(X_3 = x_3 \mid X_1 = x_1, X_0 = x_0) = \mathbb{P}(X_3 = x_3 \mid X_1 = x_1)$$

future, present, past

$$\mathbb{P}(X_3 = x_3 \mid X_2 = x_2, X_0 = x_0) = \mathbb{P}(X_3 = x_3 \mid X_2 = x_2)$$

Technicalities (*non-examinable)

Formally, a Markov chain is defined by specifying the form of its likelihood function: $\forall x_0, \dots, x_n$

$$\Pr(x_0, x_1, \dots, x_n) = \Pr_{X_0}(x_0) \Pr_{X_1}(x_1 | x_0) \Pr_{X_2}(x_2 | x_1) \times \cdots \times \Pr_{X_n}(x_n | x_{n-1})$$

From this, one can prove memorylessness results such as

$$\Pr_{X_3}(x_3 \mid X_2 = x_2, X_1 = x_1, X_0 = x_0) = \Pr_{X_3}(x_3 \mid X_2 = x_2)$$

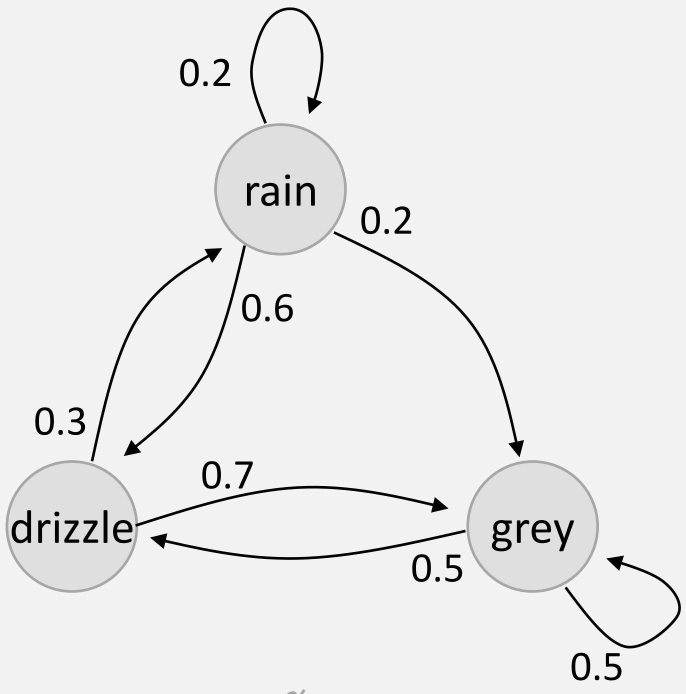
and indeed the full memorylessness theorem.

If you're ever stuck trying to prove a result about Markov chains, and if you can't see a way to use memorylessness, try going back to basics in the form of the likelihood function.

Exercise

Given that yesterday was rain, and tomorrow is rain, what's the chance that today is drizzle?

$$X_1 \to X_2 \to X_3 \to \cdots$$



$$P = \begin{array}{c} \\ \text{rain} \\ \text{drizzle} \\ \text{grey} \end{array} \begin{array}{c} \text{rain} \quad \text{drizzle} \quad \text{grey} \\ \begin{bmatrix} .2 & .6 & .2 \\ .3 & 0 & .7 \\ 0 & .5 & .5 \end{bmatrix} \end{array}$$

today yesterday tomorrow

$\mathbb{P}(X_1 = d \mid X_0 = r, X_2 = r)$

$\mathbb{P}(X_1 = x_1 \mid X_0 = x_0, X_2 = x_2)$

$= \dfrac{\mathbb{P}(X_1 = x_1, X_0 = x_0, X_2 = x_2)}{\mathbb{P}(X_0 = x_0, X_2 = x_2)}$ by definition of conditional probability

numerator $= \mathbb{P}(X_0 = x_0, X_1 = x_1, X_2 = x_2)$ by simple rewriting

$= \mathbb{P}(X_0 = x_0)\, \mathbb{P}(X_1 = x_1 \mid X_0 = x_0)\, \mathbb{P}(X_2 = x_2 \mid X_1 = x_1)$

using the general form of likelihood for a Markov chain (proved using the Chain Rule + Memorylessness)

denominator $= \sum_y \mathbb{P}(X_0 = x_0, X_2 = x_2, X_1 = y)$ by the Sum Rule (a version of the Law of Tot Prob)

$= \sum_y \mathbb{P}(X_0 = x_0)\, \mathbb{P}(X_1 = y \mid X_0 = x_0)\, \mathbb{P}(X_2 = x_2 \mid X_1 = y)$ as above.

$= \dfrac{\mathbb{P}(X_0 = x_0)\, P_{x_0 x_1} P_{x_1 x_2}}{\sum_y \mathbb{P}(X_0 = x_0)\, P_{x_0 y} P_{y x_2}} = \dfrac{P_{x_0 x_1} P_{x_1 x_2}}{\sum_y P_{x_0 y} P_{y x_2}}$

Bayes's rule: $A \to B$, figure out A given B.

What we're doing: $X_0 \to X_1 \to X_2$

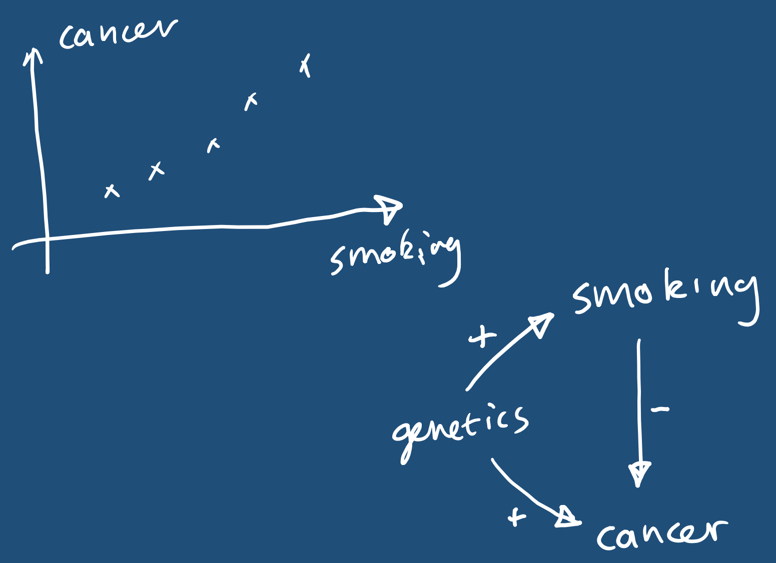
figure out X_1 given X_2 and X_0.

This is a bit like Bayes's, but fancier.

Why I'm excited about this sort of result (* non-examinable)

In science, we don't just want to learn associations, we want to learn causal mechanisms.

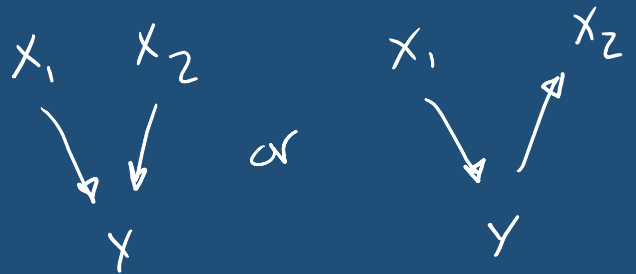
- For example, smoking is associated with getting cancer … but perhaps smoking is protective against cancer, and the association is because of some hidden causal factor (e.g. genetics) that encourages smoking and also predisposes towards cancer.



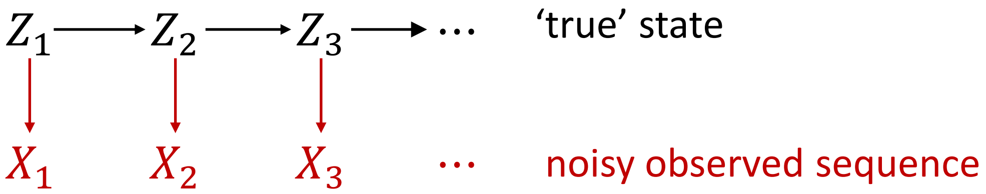
In machine learning, we're often presented with a supervised learning task ("learn to predict y given x_1 and x_2"), and we don't even think about the underlying mechanisms.

- If the causal mechanism is $X_1 \to Y \to X_2$, we can still train a supervised learning model to predict Y (as per the previous exercise)
- Open research question: how can we train ML systems to learn the causal mechanisms, rather than just associations?

x_1	x_2	y
⋮	⋮	⋮



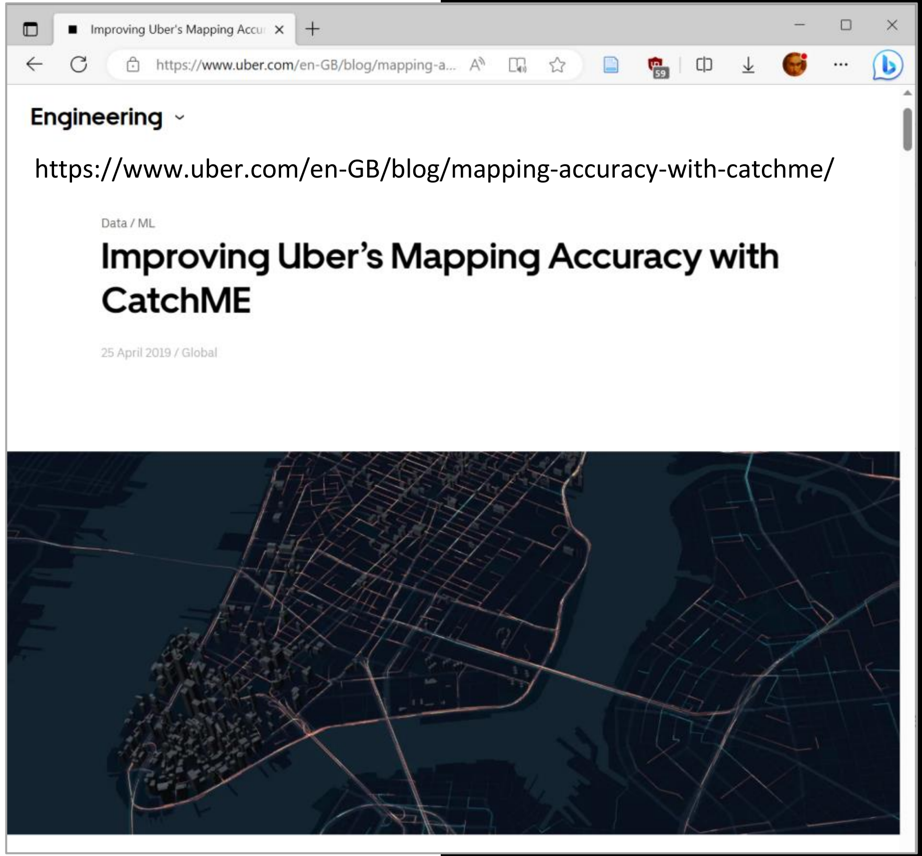
Hidden Markov models



For a hidden Markov model, the likelihood function $\Pr_X(x)$ is nasty, and it's pretty much impossible to learn the model from x data.

So why are hidden Markov models useful?

- Uber collects precise logs (both z and x) from a few drivers, so it can learn the full probability model for how Z and X are generated using straightforward supervised learning
- Then, for regular trips (only x data available), they can infer the posterior $(Z|X=x)$ using Bayes's rule
- (Alternatively, they can simply find the most likely z_T using the Viterbi algorithm)




Data Stoat

Challenge.

Our friend Data Stoat has gone missing!

The GPS sensor that they normally carry has stopped working. But they still have a low-res camera with mobile uplink, so we know what sort of scenery they're in.

Can you help find Data Stoat?

- Use data from animals 1–4 (for which we know both $\underline{z}$ and $\underline{x}$) to learn the probability model.
- Use computational Bayes to find the distribution of $\underline{Z}$ given $\underline{X} = \underline{x}$, and submit your answer as a heatmap.
- Your score will be the probability you assign to Data Stoat's actual location.
- Best answer wins a stylish Data Stoat T-shirt