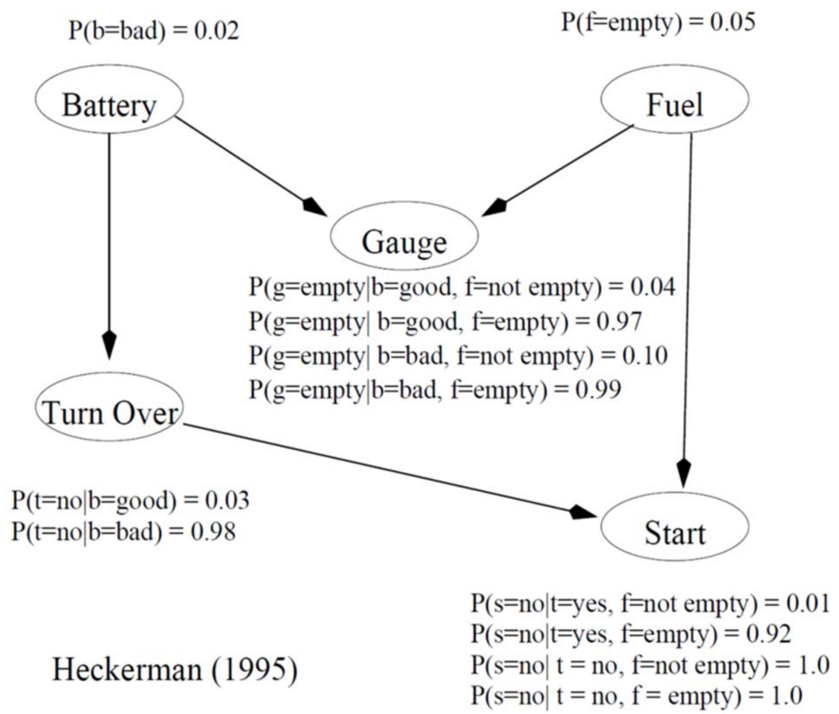
iPMR

Learning

- Being able to learn is of critical importance
  - Need to refine our uncertain knowledge using available data
- Almost all methods of learning require some form of inference over variables.
- Sometimes very approximate inference can still provide sufficient “signal” for learning.
- This week: learning is just another form of inference. It is inference over parameters.
  - Inference: inference using information from data items.
  - Learning: inference using information from data sets.

Parameters

- Previous lecture:



Uncertainty

- But where do the numbers come from?
- If we don't know the numbers what should we do?

Parameters

- If we don’t know the numbers we can represent them using parameters.
- But parameters are just unknown items with uncertainty.
- How are they different from random variables?

Parameters

- How are they different from random variables?

- They are not different
  - In fact we can just include parameters in our model and infer them in the same way.
- They are different
  - Parameters are *extrinsic* rather than *intrinsic* quantities. (I'll Explain)
  - To infer them we need to condition on data sets not data items.

Parameterising a Distribution.

- Previously

$$P(\mathbf{x}) = \frac{1}{Z} \exp\left(-E(\mathbf{x})\right)$$

- Now

$$P(D|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})^{N_D}} \prod_{n=1}^{N_D} \exp\left(-E(\mathbf{x}^n|\boldsymbol{\theta})\right)$$

where D denotes the data set $D = \{x_1, x_2, \ldots, x^{N_D}\}$.
$\boldsymbol{\theta}$ denotes the collection of all the parameters.

- Note same parameter for each *n*.

Using Parameters

For $P(\text{Toothache}, \text{Cavity})$ we can write

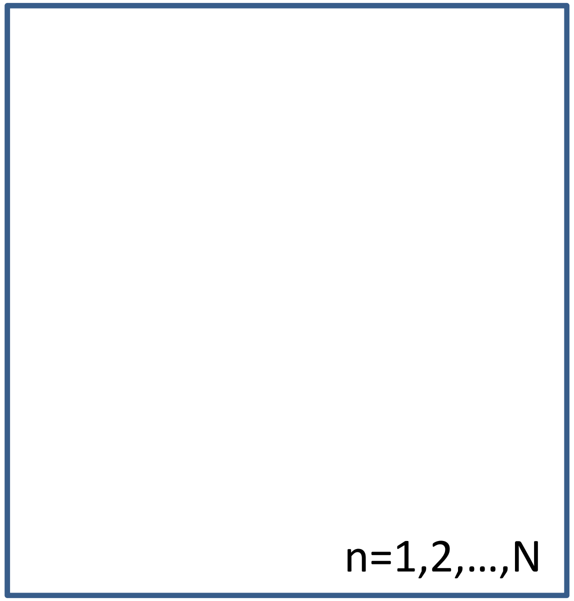
	Toothache = true	Toothache = false
Cavity = true	0.04	0.06
Cavity = false	0.01	0.89

For $P(\text{Toothache}, \text{Cavity})$ we can write

	Toothache = true	Toothache = false
Cavity = true	θ_1	θ_3
Cavity = false	θ_2	$1 - \theta_1 - \theta_2 - \theta_3$

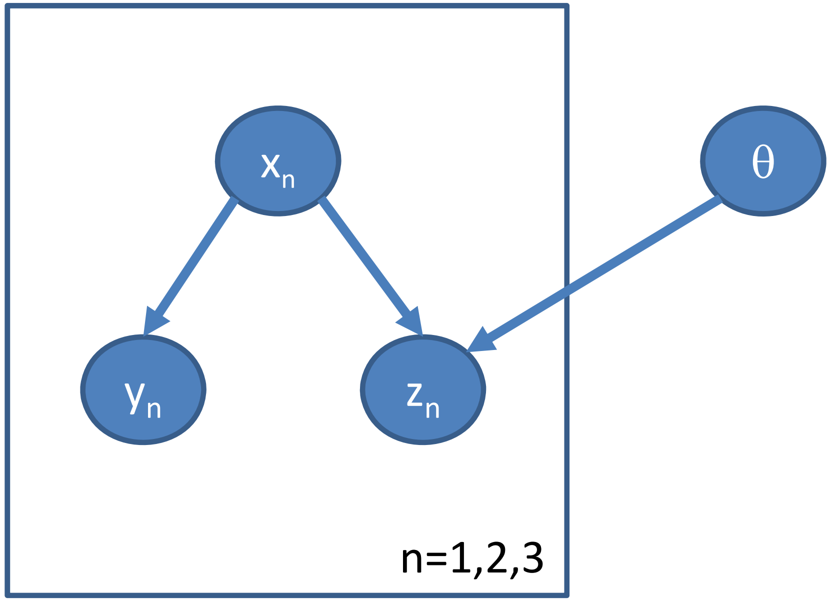
Plate Notation

- Represent repetition in graphical models as a ‘plate’:

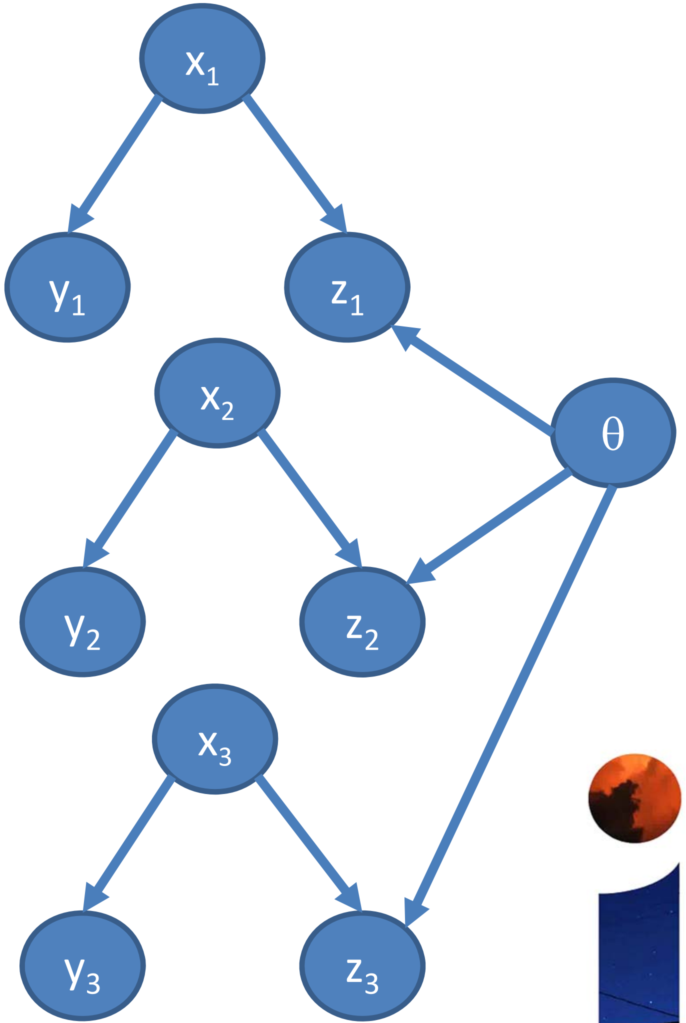


- Items inside the plate, or crossing the plate are repeated each time.
- Items outside are not repeated.

Example



=



$$P(\theta)\prod_{n=1}^{3} P(x_n)P(y_n|x_n)P(z_n|x_n,\theta)$$

Break

Learning as Inference

- Given data $D = \{\mathbf{v}^n | n = 1,2,\ldots, N_D\} = \mathbf{v}^{1:n}$. Have full model

$$\prod_{n=1}^{N_D} P(\mathbf{v}^n, \mathbf{x}^n, \mathbf{h}^n | \boldsymbol{\theta})$$

- Define a prior distribution $P(\boldsymbol{\theta})$ over parameters to give joint model

$$P(\mathbf{v}^{1:n}, \mathbf{x}^{1:n}, \mathbf{h}^{1:n}, \boldsymbol{\theta}) = \left[\prod_{n=1}^{N_D} P(\mathbf{v}^n, \mathbf{x}^n, \mathbf{h}^n | \boldsymbol{\theta})\right] P(\boldsymbol{\theta})$$

- Now consider a new test case (not in the training set). Index by $n = 0$

$$P(\mathbf{v}^{0:n}, \mathbf{x}^{0:n}, \mathbf{h}^{0:n}, \boldsymbol{\theta}) = \left[\prod_{n=0}^{N_D} P(\mathbf{v}^n, \mathbf{x}^n, \mathbf{h}^n | \boldsymbol{\theta})\right] P(\boldsymbol{\theta})$$

- Compute what we want via inference:

$$P(\mathbf{x}^0 | \mathbf{v}^0, D) \propto P(\mathbf{x}^0, \mathbf{v}^0 | D) = \sum_{\boldsymbol{\theta}, \mathbf{h}^{0:n}, \mathbf{x}^{1:n}} \left[\prod_{n=0}^{N_D} P(\mathbf{v}^n, \mathbf{x}^n, \mathbf{h}^n | \boldsymbol{\theta})\right] P(\boldsymbol{\theta})$$

Prior

- Note we needed prior $P(\boldsymbol{\theta})$ to do this.
- Prior comes from assumptions about parameters.
- Sometimes people choose reasonable assumptions.
- Sometimes people choose assumptions that make the maths easy.
- Sometimes people choose a vague prior (but it is still an assumption)
- Note also that the "sum" over $\boldsymbol{\theta}$ is really likely to be an integral over $\boldsymbol{\theta}$ as parameters will likely be real valued.
- We will be doing some multivariate integrals…

Independent Data Items

- If, conditioned on the parameters, each data item is independent.
  - Can use plate notation.
  - The test data is only connected to the training data via the parameters.

$$P(\mathbf{x}^0, \mathbf{v}^0|D) = \int d\boldsymbol{\theta} P(\mathbf{x}^0, \mathbf{v}^0|\boldsymbol{\theta}) P(\boldsymbol{\theta}|D)$$

  - $P(\boldsymbol{\theta}|D)$ is the *posterior* distribution.
  - It is what we now know about the parameters having seen the data.

Example

Example

1 0 0 1 0 1 0 1 0 0 0 0 0 1 0 1 1 1 0 1. $P(p) \propto p(1-p)$

- *p* denotes the probability of a 1 turning up. What is the probability that the next item x^* is 1?
- *p* takes one value for all data items. But we don't know what that is! Use a distribution to represent and compute with the uncertainty.

$$\int dp\; P(x_* = 1|p)P(p|D) \text{ where } P(p|D) = \frac{P(D|p)P(p)}{P(D)}.$$

$$\text{Now } P(D|p) = \prod_n P(x^n|p) = p^{N_1}(1-p)^{N_0} = p^9(1-p)^{11}$$

$$\text{so } P(p|D) = \frac{23!}{10!12!}p^{10}(1-p)^{12}$$

meaning $P(x_* = 1) = \int dp\; P(x_* = 1|p)P(p|D)$. Computing this gives $11/24$. (Hint: look up Beta distribution).

Note $N_1 = 9$ is the number of ones, and $N_0 = 11$ is the number of zeros.

Learning can be hard.

- Why can learning be hard?

$$P(\mathbf{v},\mathbf{h}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})}\exp(-E(\mathbf{v},\mathbf{h};\boldsymbol{\theta}))$$

where

$$Z(\boldsymbol{\theta}) = \int d\mathbf{v}d\mathbf{h}\exp(-E(\mathbf{v},\mathbf{h};\boldsymbol{\theta}))$$

- *Z* depends on the parameters.
- It is not always easy to compute.

Our Journey

- Introduction to Learning
- Next: Q&A
- Next: Learning exponential family models, with examples.