Prof. Alex Rogers
Department of Computer Science
University of Oxford
31st October 2020

Building Models in PyMC3
Session 4

Contents

Exercises

1. Use the knowledge that the Dirichlet distribution is identical to the beta distribution in the special case that $k = 2$ to derive a relationship between a Gamma function and a factorial.

$$\rho(n_1, \ldots, n_k | p_1, \ldots, p_k) = \frac{N!}{\prod_{i=1}^{k} n_i!} \prod_{i=1}^{k} p_i^{n_i}$$

$$\rho(x_1, \ldots, x_k | a_1, \ldots, a_k) = \frac{\Gamma\left(\sum_{i=1}^{k} a_i\right)}{\prod_{i=1}^{k} \Gamma\left(a_i\right)} \prod_{i=1}^{k} x_i^{a_i - 1}$$

$$\Gamma(x) = (x-1)!$$

2. For the example of measurement noise in the notes, confirm that a Gamma distribution can be used as a prior for both the value of the physical property being measured and also the standard deviation of measurement noise. Propose suitable values for the parameters of these Gamma distributions.

3. A supermarket is trying to estimate the number of customers who will visit a store each day and uses a Poisson process to model the arrival of customers. If 40 people visit the store on the first day, how many are expected to visit the store on any day? What is the uncertainty in this estimate? If on subsequent days 30, 60, and 50 customers actually do visit the store, how does the estimate of the expected number of customers change?

4. A common probabilistic model is a Gaussian mixture model where there are two possible Gaussian distributions from which a measurement is made and a binary latent variable is used to indicate which one any measurement actually comes from. This is similar to the example from Session 1 where rather than two independent Gaussians, one Gaussian and an offset was used to explain the height measurements observed.

   (a) Modify the example from Session 1 to be a true mixture of two independent Gaussians, each with their own mean and variance. Generate some suitable data and show that your model is able to infer the correct properties of the two Gaussian distributions.

   (b) The mixture model in Session 1 used the following Bernoulli distribution to describe whether a measurement fell into each category:

   ```
   male_or_female = pm.Bernoulli('male_or_female', 0.5, shape=N)
   ```

   What does `shape=N` do here? What happens if it is removed? Why might this be an appropriate thing to do in some models?

Model Building

Model Building Methodology

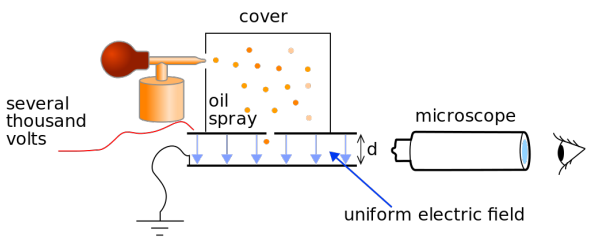
1. Understand the setting and the scientific question being asked of the data.
2. What are the sources of noise and imprecision in the process being considered.
   - What likelihood function is appropriate?
     - Are we making error-prone continuous measurements or are our observations discrete?
     - Do we know anything about the expected accuracy of the measurements?
3. Think about what prior information is available.
   - How should that be represented within the model?
     - What prior probability distributions should we use?
     - Are random variable discrete or continuous?

Millikan Oil Drop Experiment

Millikan Oil Drop Experiment

https://en.wikipedia.org/wiki/Oil_drop_experiment

Millikan Oil Drop Experiment



https://en.wikipedia.org/wiki/Oil_drop_experiment

Millikan Oil Drop Experiment

1. Make noisy measurements of the charge on each oil drop.
   - The range of the measurement that we can make is around 0 to 10 x 1e-19 coulombs.
   - Our measurements should be accurate to +/- 0.1 x 1e-19 coulombs if we are careful.
2. We think each oil drop will pick up excess charge equivalent to just a few electrons.
   - The charge is due to an integer excess of electrons.
3. Previous experiments suggest that the charge on an electron is between 1 and 2 x 1e-19 coulombs.
   - What do our experimental results suggest for this value?
   - How accurate is our experimental setup?

oil_drop.ipynb

Describing the Model

1. To fully describe the model:
   - Introduce the notation used to describe things.
   - Describe the likelihood function.
   - Describe the priors.
2. We should then describe the way in which the model was solved (e.g. run PyMC3 with 2000 tuning steps and 5000 draws).
3. We should separate describing the model and running inference on the model.
4. It is convenient to use Latex since the mathematics are nicely presented for us.

report.tex

https://www.overleaf.com/

Your Project

An AI to Help Reduce Heating Bills

You are asked by an energy company to develop an AI to help their customers reduce their heating bills.

They want to be able to identify customers who are using too much energy for their home heating, either:

- Because they have their **heating** turned up **too high**.
- Or because their home is **poorly insulated** and uses a lot of energy to maintain a comfortable temperature.

An AI to Help Reduce Heating Bills

Each home has a **smart meter** that measures the daily total electricity used by the home.

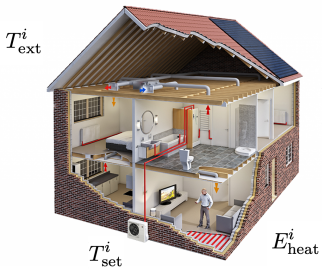
Electricity is used for running household **appliances** and for **heating** the home.

The heating is only on during the winter months when the external temperature drops below the **thermostat set point** temperature.

Heat leaks from the home at a rate that is proportional to the difference between the **internal temperature** and the **external temperature**.

The amount of electrical energy used by the heating system each day matches the amount of heat energy that leaks out of the house.

An AI to Help Reduce Heating Bills



$$24 \times \text{HPLC} \times \left(T^i_{\text{set}} - T^i_{\text{ext}}\right)$$

$$E^i_{\text{heat}}$$

$$E^i_{\text{heat}} = 24 \times \text{HPLC} \times \left(T^i_{\text{set}} - T^i_{\text{ext}}\right)$$

Heating off	$E^i_{\text{meter}} = E^i_{\text{app}}$	$T^i_{\text{ext}} > T^i_{\text{set}}$
Heating on	$E^i_{\text{meter}} = E^i_{\text{heat}} + E^i_{\text{app}}$	$T^i_{\text{ext}} \leq T^i_{\text{set}}$

An AI to Help Reduce Heating Bills

You will be provided with **external temperature** data (°C) and the **daily electricity consumption** measured by the smart meter (kWh) from five houses.

Design a robust probabilistic model that captures the features of this data.

Use the model to estimate:

1. The thermostat set point temperature of each house [°C].
2. The heating power loss coefficient of each house [kW/°C].

Houses 4 and 5 present additional challenges and will require a more sophisticated model.

home_heating.ipynb

Next Time

Explore and plot the dataset.

Propose a suitable model for houses 1-3.

What might be explaining houses 4 and 5?