Probabilistic Time Series Analysis: Lab 1

Tim Kunisky

September 5, 2018

Some Organizational Details

- **Your TA:** Dmitriy (Tim) Kunisky, Courant PhD Student
  - Email: `kunisky@cims.nyu.edu`
  - If you want to discuss something outside of office hours, email me and I'll try to find a time.
- **Office Hours:** Right after lab in the same room today, TBA in the future.
- **Lab Work:** 10% of your grade is participation in lab and class—please participate!
  - We will do exercises in lab. Sometimes we will just go over them together, but sometimes I will ask you to hand in a small amount of written work or code before the next lab.

Basic Notions of Probability

- *State space:* the set of all possible outcomes of a random experiment or process, often written Ω.
  - E.g. rolling a die once: $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- *Events:* sets of outcomes.
  - "The event that I roll an even number" $= \{2, 4, 6\}$.
- *Random variables:* measurements of an outcome, usually with scalar, vector, or categorical values.
  - X = "The number that I rolled, mod 2" $\in \{0, 1\}$
  - If X is a random variable, then $\{\omega \in \Omega : X(\omega)$ has some property$\}$ is...**an event**.
- *Probabilities:* our certainty an event will happen, measured in $[0, 1]$. **For this class, we are Bayesian!**
  - Random variables have *probability distributions* or *densities* over the set of values they take, $p(x) = \mathbb{P}(X = x)$, such that $\sum_x p(x) = 1$ or $\int p(x)dx = 1$.

Conditioning

- Recalculate all probabilities, assuming some event has happened:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \text{ and } B)}{\mathbb{P}(B)}, \text{ usually written } \frac{\mathbb{P}(A, B)}{\mathbb{P}(B)}.$$

- **Question:** Can conditioning increase a probability? Decrease a probability? Leave it the same? Can you give examples for one dice roll? Which of these situations has a special name?
- An important consequence (stay tuned!): **Bayes' rule:**

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

and therefore...

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

Independence

There are two equivalent descriptions of events A, B being independent:

1. $\mathbb{P}(A, B) = \mathbb{P}(A)\mathbb{P}(B)$ (maybe more familiar)
2. $\mathbb{P}(A \mid B) = \mathbb{P}(A)$ (maybe more intuitive: B gives no information about whether A happens)

The same ideas apply for *conditional independence*: either one of the equivalent conditions

1. $\mathbb{P}(A, B \mid C) = \mathbb{P}(A \mid C)\mathbb{P}(B \mid C)$
2. $\mathbb{P}(A \mid B, C) = \mathbb{P}(A \mid C)$ (if you know already whether C happens, B gives no further information about whether A happens)

Marginalization

- The process of “forgetting irrelevant information” about a probability distribution.
- Usually with random variables: given the joint distribution of X, Y, sum over all possible values of Y to get the *marginal distribution* of X:

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

- Same applies for subsets of variables, with conditioning, etc.

Getting Information from Joint Distributions

- Combining **conditioning** and **marginalization**, we can pass from a joint distribution to the distribution of any subset of variables conditional on any any other subset of variables.
- **Question:** suppose we have a joint distribution $p(A, B, C, D, E)$. How to compute $\mathbb{P}(A = a \mid B = b)$?

$$\mathbb{P}(A = a, B = b) = \sum_{c,d,e} \mathbb{P}(A = a, B = b, C = c, D = d, E = e)$$

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

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

- **All of Bayesian inference is about trying to do this faster!**

Using Bayes' Rule for Statistics

- *Assume a data-generating process:* data $\boldsymbol{y}$ is generated by some random process from a model parametrized by $\boldsymbol{\theta}$.
  - E.g. we make observations at some fixed points $\boldsymbol{x}$, and observe $y_i = ax_i + b +$ noise, a linear model.
  - **Question:** What is $\boldsymbol{\theta}$ here?
- We see $\boldsymbol{y}$, and want to *infer* $\boldsymbol{\theta}$.
- Bayes' rule tells us how:

$$\text{posterior} = p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

- The evidence does not depend on $\boldsymbol{\theta}$, so for inferring $\boldsymbol{\theta}$ we often drop it and say "posterior $\propto$ likelihood $\times$ prior."
- **Important note:** often a confusing shorthand is used, where

$$p(x) = \mathbb{P}(\text{the variable called } x = \text{ the value } x).$$

Exercise 1

Suppose $x_1, \ldots, x_n$ are fixed, $\boldsymbol{\theta} = (a, b)$ is a parameter vector with independent priors $a \sim \mathcal{N}(0, 1)$ and $b \sim \mathcal{N}(0, 1)$, and the data $y_1, \ldots, y_n$ is generated as $y_i = ax_i + b + \xi_i$, where $\xi_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d. gaussian noise.

Write down a formula for

$$\begin{aligned} p(\boldsymbol{\theta} \mid \boldsymbol{y}) &\propto p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) \\ &= ??? \end{aligned}$$

Graphical Models

- A way of keeping track of dependencies among variables in complicated data-generating processes.
- Always possible to factorize a joint distribution naively into conditional distributions (*chain rule*):

$$
\begin{aligned}
p(x_1, x_2, x_3) &= p(x_3, x_2 \mid x_1) p(x_1) \\
&= p(x_3 \mid x_2, x_1) p(x_2 \mid x_1) p(x_1)
\end{aligned}
$$

- But if some variables are *conditionally independent*, then this can be simplified: e.g., if

$$p(x_3, x_2 \mid x_1) = p(x_3 \mid x_1) p(x_2 \mid x_1),$$

then

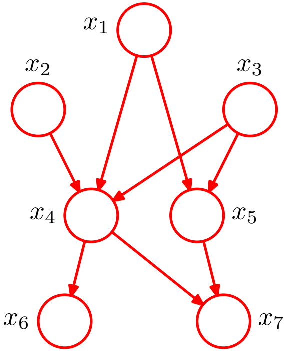
$$p(x_1, x_2, x_3) = p(x_3 \mid x_1) p(x_2 \mid x_1) p(x_1).$$

Graphical Models

- To capture this, describe a *formula* for $p(x_1, \ldots, x_n)$ as a product of $p(x_i \mid x_a, x_b, \ldots)$ by a *directed graph* where every dependence is described by an arrow.
- Examples:
  - Chain rule $\rightsquigarrow$ complete graphs
  - Conditional independences $\rightsquigarrow$ less connectivity (but be careful! examples to come)
- **Question:** What must be true about a directed graph representing a valid graphical model?
- **More advanced question:** Do you know another name for a graphical model that is a directed path?

Exercise 2.1

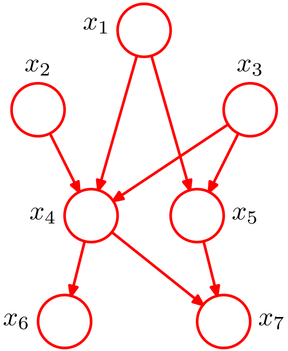
Write down a formula for the joint distribution $p(x_1, \ldots, x_7)$ described by the following graphical model.



(From Bishop, Chapter 8)

Exercise 2.2

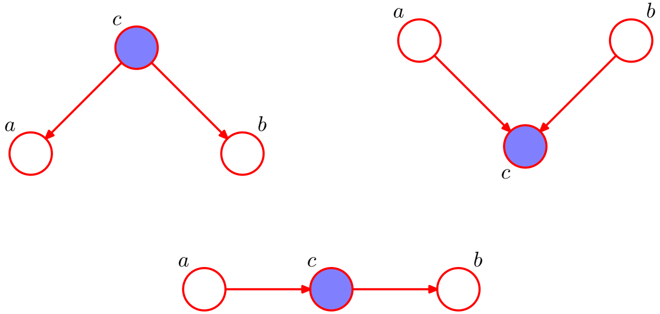
If this describes a statistical model, which variables will have priors assigned to them?



(From Bishop, Chapter 8)

Exercise 3

In each of the following graphical models, are a and b independent? Are they independent conditional on c?



(From Bishop, Chapter 8)