Introduction to Bayesian Models and Inference

Jeremy R. Manning
Contextual Dynamics Lab

January 30, 2019

WARNING

- You’re going to hear a lot of terms
- If this is your first exposure to Bayesian models, most of those terms will be unfamiliar
- The best way to learn is to let the words “wash over” you and just keep practicing and trying
- Eventually the concepts will stick and you’ll be fluent
- Ask questions, repeatedly, until you start to understand— it will help everyone!

What is a model?

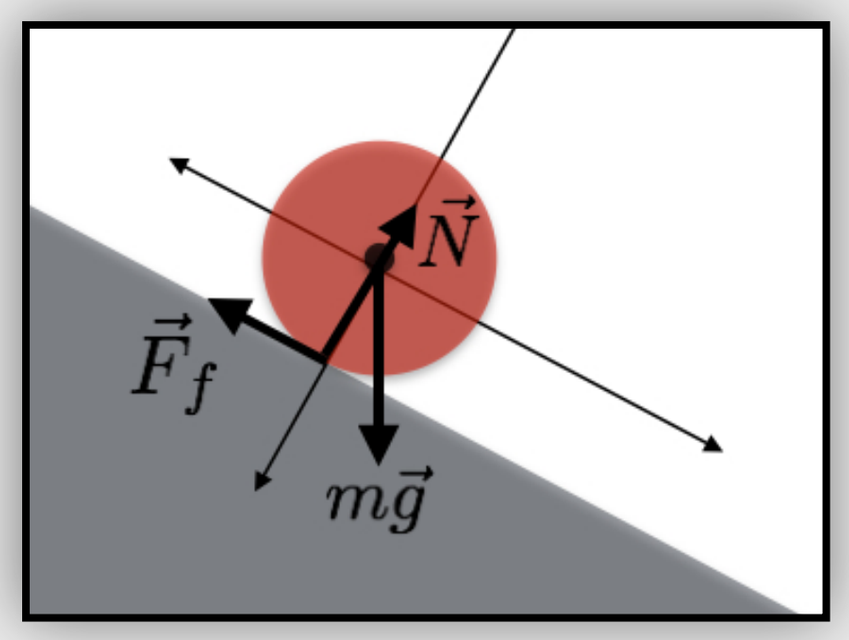
- A **model** is a machine that takes **inputs** and produces **outputs**
- **Inputs** are numbers that correspond to **observable stuff in the world** (e.g. pixel intensities in an image, EEG voltages, counts, sensor measurements, task instructions, etc.)
- **Outputs** are numbers that correspond to stuff we are trying to **explain, predict, or do** (e.g. probability that we are looking at a cat, amount to turn the steering wheel, estimated price of a stock 24 hours from now, etc.)

What is a model?

- Creating a model often entails defining a series of steps that explains “how the world works” within the limited microcosm you’re trying to account for
- Why have a model— aren’t our observations the best predictors of what happened (or will happen)?
  - Observations can be noisy; we can never really know the “truth”
  - Models force us to be explicit about our assumptions, providing a context for our observations

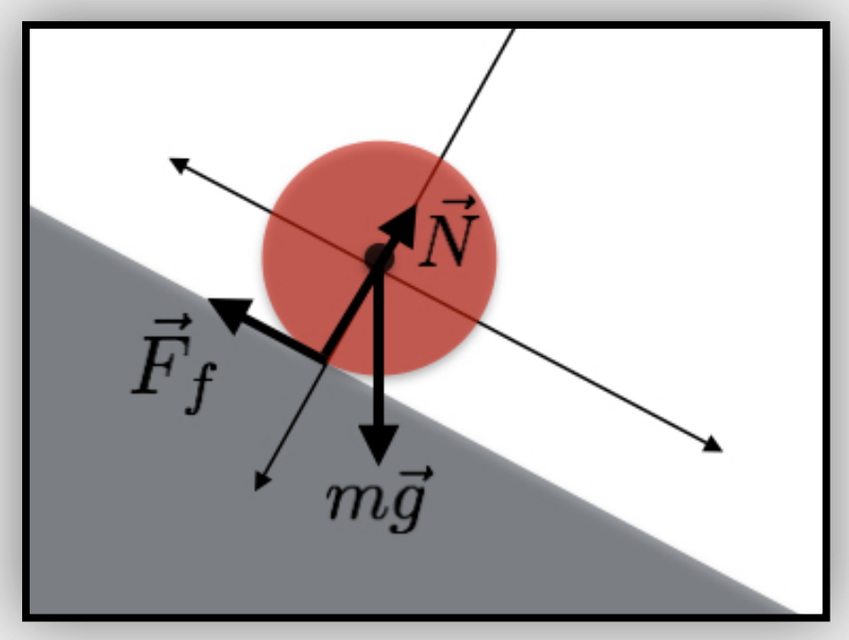
Example: ball rolling down a hill

- **Inputs:**
  - Current status of the ball (position, velocity, mass, size, surface texture, etc.)
  - Properties of the hill (surface texture, slope)
  - Properties of the world (gravity, air resistance, wind)
- **Outputs:**
  - Status of the ball in the next moment



Example: ball rolling down a hill

- If we wanted to infer the rate of acceleration due to gravity, what other stuff would we need to know?
- Suppose we didn't know any of the specific numbers (e.g. amount of friction, magnitudes of different forces, etc.), but we **did** know how the forces interacted and where the ball was located at each moment. What could we learn about the system?
- Suppose some of our assumptions about the physical forces acting on the ball were wrong. How would that change our answers?

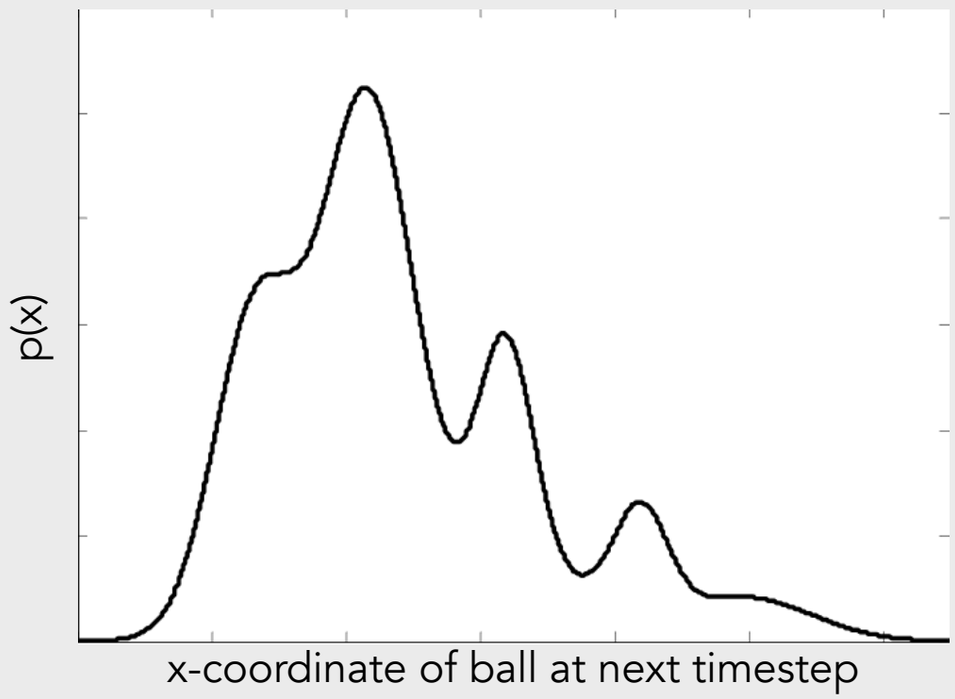


Uncertainty

- When any of the following aren't 100% reliable, we can't perfectly predict what an outcome will be:
  - What we **measure** as the current state of the system
  - What we **assume** about the underlying mechanisms of the system
  - How **reliable** the system itself is, even when everything else is held fixed ("noise")

Uncertainty

- We can use **probability distributions** to represent uncertainty
- They tell us: considering each **possible** state of the system, how likely is it that the system is in that state?

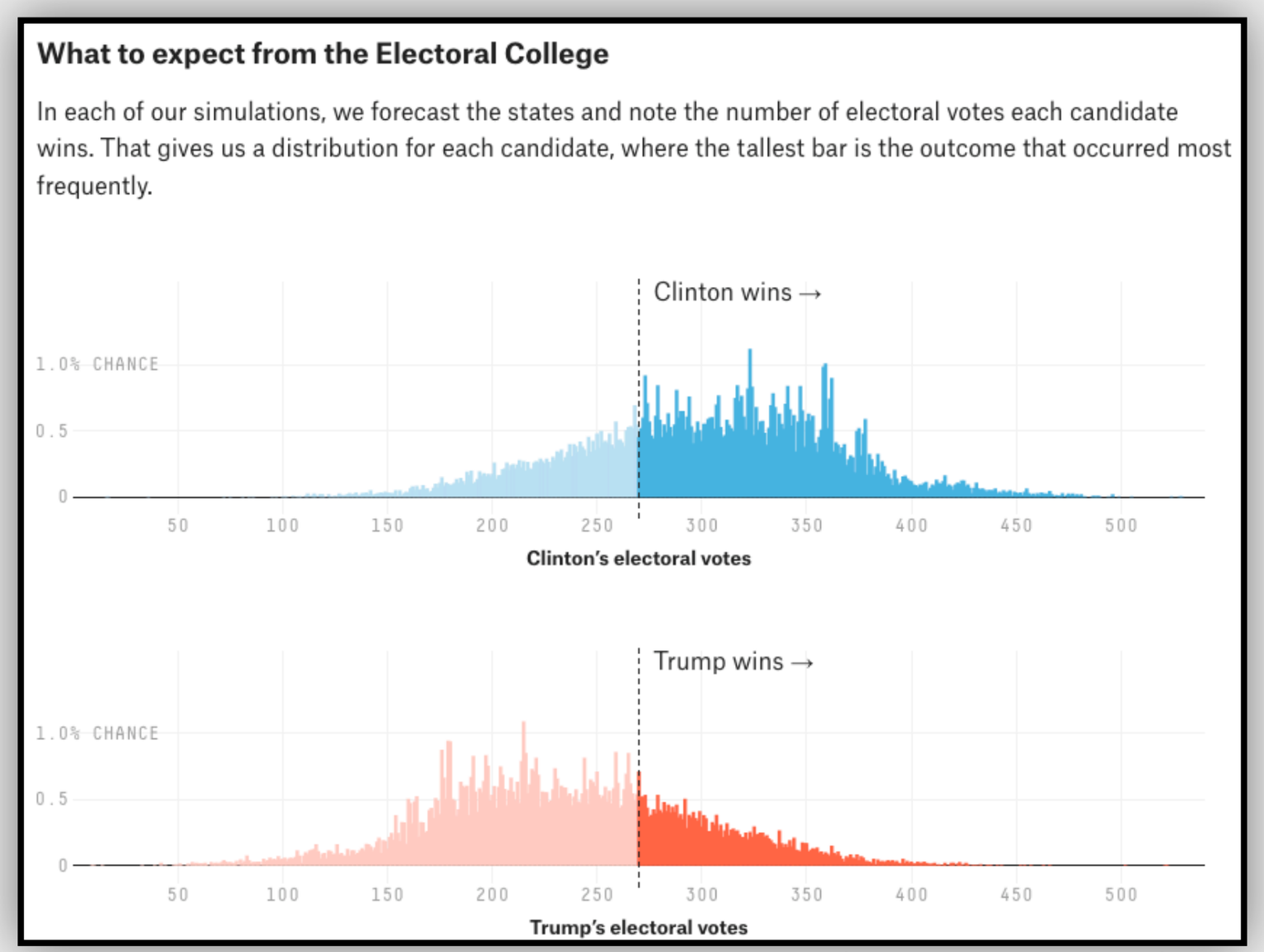


More about probability distributions…

- In order to come up with an estimate of how likely each state is, we need to account for how those states arise
- E.g. are all states equally likely? Does the state depend on the previous state? What are the constraints on the system (e.g. Newton's laws, connectivity diagrams, etc.)?
- Building in assumptions about how likely each state is entails building a model (i.e. accounting for constraints)
- **Probability distributions are models**

More about probability distributions…

- Example: simulating election outcomes
- Where do the local peaks and troughs come from—why aren't these distributions smooth?



https://projects.fivethirtyeight.com/2016-election-forecast/

More about probability distributions…

- "All models are wrong, but some are useful" (Box, 1976)
- There is no such thing as a perfect model—the question is how detailed (and accurate) our model **needs to be** in order to provide useful insights into the question(s) we care about

Accounting for prior beliefs

- Which number(s) don’t belong?

1, 2, 3, 923847523495 87, 5, 6, 7, 8, 9, 10

Accounting for prior beliefs

- Which number(s) don't belong?

1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1

Accounting for prior beliefs

- Which number(s) don't belong?

1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1

Accounting for prior beliefs

- When we interpret results, we often bring intuitions about which sorts of answers are within the realm of "reasonable"
- We can account for our assumptions by up-weighting the assigned probability of states we think are more reasonable, even in the absence of evidence
- We could define a new distribution that tells us how probable each state is, before we measure or observe anything

Names for three key ideas so far

- **Likelihood:** a probability distribution that tells us how likely it is that our system is in each possible state, **given our observations**
- **Prior:** a probability distribution that tell us how likely it is that our system is in each possible state, **before we observe anything**
- **Posterior:** a probability distribution that tells us how likely it is that our system is in each possible state, accounting for the likelihood and prior

Working with probability distributions

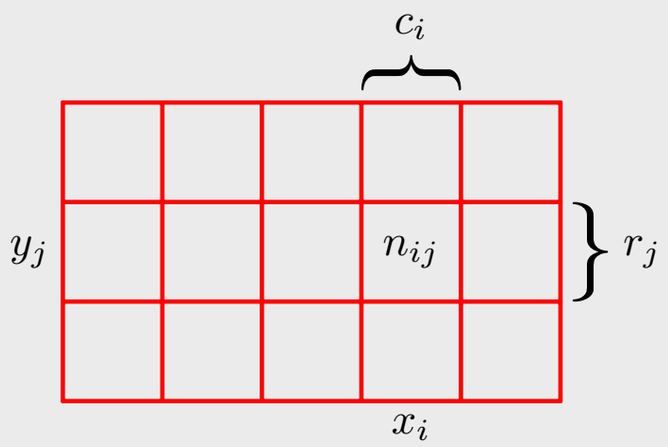
- p(X) is shorthand for p(X = x)
  - X: a random variable (e.g. something that could take on many possible values)
  - x: a specific value of the random variable
- **Joint probability:** p(X, Y)
- "The probability of X **and** Y" — in other words, the probability that X has some specific value (x) AND ALSO Y has some specific value (y).
- Think about: what does p(X, Y) look like?

Working with probability distributions

- **Conditional probability:** p(X | Y)
  - "The probability of X **given** Y"
  - If we assume that Y = y, then what's the probability that X = x?
  - When X and Y are independent, p(X | Y) = p(X)

Working with probability distributions (Bishop, 2006)

Figure 1.10 We can derive the sum and product rules of probability by considering two random variables, X, which takes the values $\{x_i\}$ where $i = 1, \ldots, M$, and Y, which takes the values $\{y_j\}$ where $j = 1, \ldots, L$. In this illustration we have $M = 5$ and $L = 3$. If we consider a total number N of instances of these variables, then we denote the number of instances where $X = x_i$ and $Y = y_j$ by n_{ij}, which is the number of points in the corresponding cell of the array. The number of points in column i, corresponding to $X = x_i$, is denoted by c_i, and the number of points in row j, corresponding to $Y = y_j$, is denoted by r_j.



The Rules of Probability

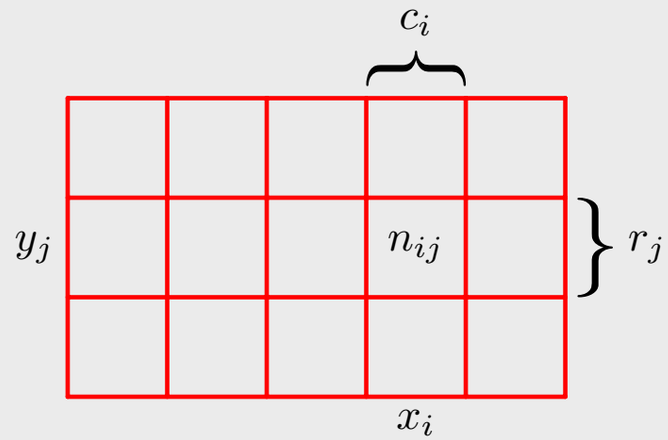
sum rule
$$p(X) = \sum_{Y} p(X, Y) \tag{1.10}$$

product rule
$$p(X, Y) = p(Y|X)p(X). \tag{1.11}$$

Working with probability distributions (Bishop, 2006)

sum rule $p(X) = \sum_Y p(X, Y)$

product rule $p(X, Y) = p(Y|X)p(X)$

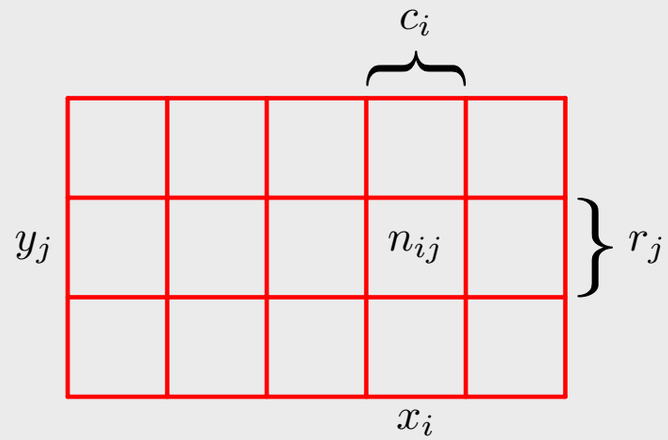


- The **sum rule** entails summing down the rows of the table
- If you know p(X, Y), then you can compute p(X) by summing p(X, Y=y1) + p(X, Y=y2), …, y(X, y=yN)

Working with probability distributions (Bishop, 2006)

sum rule $p(X) = \sum_Y p(X, Y)$

product rule $p(X, Y) = p(Y|X)p(X)$



- The **product rule** entails two steps:
  - Consider the **conditional probability** p(Y|X). In other words, what is p(Y) given each possible value of X?
  - Now multiply each of those conditional probabilities by the probability of each value of X
  - This tells you about the probabilities of each possible combination of the values of X and Y

Deriving Bayes' Theorem

- **Product rule**: p(X, Y) = p(Y | X) p(X)
- This rule is symmetric: we could have just as correctly written p(X, Y) = p(X | Y) p(Y)
- Therefore p(Y | X) p(X) = p(X | Y) p(Y)
- This gives us a way of flipping conditional probabilities (divide both sides by p(X)):

$$p(Y \mid X) = p(X \mid Y)\, p(Y) \,/\, p(X)$$

Bayes' Theorem

$$p(Y \mid X) = \frac{p(X \mid Y)\, p(Y)}{p(X)}$$

Bayes' Theorem

$$p(Y \mid X) = \frac{\overbrace{p(X \mid Y)}^{\text{Likelihood}}\, p(Y)}{p(X)}$$

Bayes' Theorem

$$p(Y \mid X) = \frac{\overbrace{p(X \mid Y)}^{\text{Likelihood}}\ \overbrace{p(Y)}^{\text{Prior}}}{p(X)}$$

Bayes' Theorem

$$p(Y \mid X) = \frac{\overbrace{p(X \mid Y)}^{\text{Likelihood}}\ \overbrace{p(Y)}^{\text{Prior}}}{\underbrace{p(X)}_{\substack{\text{Normalizing} \\ \text{term}}}}$$

Bayes' Theorem

$$\underbrace{p(Y \mid X)}_{\text{Posterior}} = \frac{\overbrace{p(X \mid Y)}^{\text{Likelihood}}\overbrace{p(Y)}^{\text{Prior}}}{p(X)}$$

Normalizing term

Bayes' Theorem

$$\underbrace{p(Y \mid X)}_{\text{Posterior}} = \frac{\overbrace{p(X \mid Y)}^{\text{Likelihood}} \overbrace{p(Y)}^{\text{Prior}}}{\underbrace{p(X)}_{\text{Normalizing term}}}$$

- Suppose **X is the data**— e.g. stuff you observe, and **Y is the set of parameters you want to estimate**— e.g. the slope of a line, the state of a system, the locations of things you're tracking using a noisy sensor, etc.

- The **likelihood p(X | Y)** says: if we knew the parameters, how likely would we be to observe different values in our data?

- The **prior p(Y)** says: how likely are different parameters, not accounting for anything we know about the data

- The **normalizing term p(X)** says: how likely are different values in our data? Often we don't know or care about this term, so we estimate everything else up to a (missing) constant.

- The **posterior says p(Y | X)** says: given the observed data, how probable are different parameter values?

Bayes' Theorem

$$\underbrace{p(Y \mid X)}_{\text{Posterior}} = \frac{\overbrace{p(X \mid Y)}^{\text{Likelihood}}\overbrace{p(Y)}^{\text{Prior}}}{\underbrace{p(X)}_{\substack{\text{Normalizing} \\ \text{term}}}}$$

- Most of the "modeling" in Bayesian models happens in the likelihood term. This entails building a set of dependencies between the model parameters and the data.

- The prior is where we can impose different constraints on the parameters (e.g. force them to be Real numbers, non-negative Integers, positive semi-definite matrices, etc.). This generally entails selecting from a set of basic "building blocks" (pre-existing distributions) that have the desired properties.

Likelihood

- Think of word problems from math or science courses— try to solve for the "one missing thing" if you know all of the other pieces and the relevant rules

- The likelihood function describes how all of the different pieces relate to each other. It's kind of like a recipe.

- The first step is to start with whatever your assumptions are

- Then use the rules of the model to build on those assumptions; this repeats as additional rules are applied to the results of previously applied rules

- Eventually you produce the data (observations). That's the "result" of your recipe.

- If you fix the assumptions and rules in your system, you may get slightly different results each time you follow the recipe. The likelihood distribution tells you how likely it is that you'll see each of those outcomes.

Likelihood

- Another way of thinking about the likelihood is like setting up a complex machine with many moving parts
- The likelihood tells you what the parts are and how they interact
- The particular settings of the parameters (the prior, which we'll get to soon!) are going to also affect how the machine behaves (e.g. angles of each lever, tension on the springs, etc.)

Likelihood: baking a cake

- Assumptions: the ingredients and amounts
- Recipe: combine the ingredients in a particular way. This might entail building "interim cake parts" (e.g. suspension of eggs + oil; mixture of dry ingredients; batter; batter + pan raw cake; etc.)
- Output: given the starting ingredients and particulars of the recipe, how likely is it that you'll get any given cake instantiation? E.g. what are the chances that you'll get the precise amount and distribution of frosting, the precise sprinkles configurations, precise air pocket placements, etc.

Likelihood

- Remember that the likelihood is a probability distribution. We need to define some space that describes all possible outputs, where each coordinate is some possible thing we “could have” observed.

- Then we can ask, for each coordinate in that space, what were the chances of generating that observation?

- Pro tip 1: ideally we’d want the likelihood to have some nice (mathematical) form to it, so that we don’t need to manually enumerate every possible outcome

- Pro tip 2: in defining the likelihood function, you want to simultaneously optimize several things:

    - Simplicity (easy to write down equations and compute with)

    - Flexibility (you want the things you care about to be representable)

    - Specificity (you want things you **don’t** care about to be excluded— e.g. if you want to compute likelihoods of cakes, you probably don’t need the flexibility to represent different species of spiders)

Prior

- The prior tells you how to "start" your recipe. E.g. before you start measuring things out, blending, or heating them up, you first need to select which ingredients to use
- The prior constrains the solution space (e.g. it'd be hard to end up with spaghetti sauce if you were starting with flour, eggs, butter, sugar, water, baking soda, etc.). But it **doesn't** tell you anything particular about your specific data— e.g. given a particular set of ingredients, you might end up with a pancake, a layered cake, a bunch of cupcakes, etc.

Prior

- Given however the likelihood is set up (i.e. what the parts of the machine are), the prior tells you about the specific settings of its components
- You'll need to consider what form those specifications might take (e.g. scalars, vectors, matrices, etc.)
- You also need to consider which settings of those parameters are more "reasonable," and which are allowed in the first place (e.g. binary, complex, Real valued, sum to 1, symmetric, between -30 and +26 except on Tuesdays, etc.)

Prior

- Pro tip: in practice, defining the prior (over each parameter of your model) entails choosing some existing distribution that has the desired properties (next slides…)
- Then you can make a guess about the parameters of those distributions (called “hyperparameters”— parameters on parameters!) to help define the general range of answers that might be possible to get out. This is somewhat of an art, and often requires some exploration in order to find hyperparameters that are sufficiently flexible to allow you to model a broad range of phenomena within the domain you’re studying.
- The best way to get a feel for what works well is to build models and play around with them (i.e. more tutorials!)

Bernoulli distribution

- Outputs a single binary variable (e.g. a coin flip)

Bernoulli

This is the distribution for a single binary variable $x \in \{0, 1\}$ representing, for example, the result of flipping a coin. It is governed by a single continuous parameter $\mu \in [0, 1]$ that represents the probability of $x = 1$.

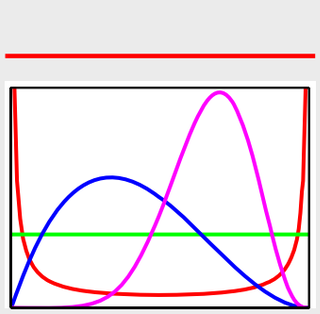
$$\begin{aligned}
\text{Bern}(x|\mu) &= \mu^x(1-\mu)^{1-x} && \text{(B.1)}\\
\mathbb{E}[x] &= \mu && \text{(B.2)}\\
\text{var}[x] &= \mu(1-\mu) && \text{(B.3)}\\
\text{mode}[x] &= \begin{cases} 1 & \text{if } \mu \geqslant 0.5, \\ 0 & \text{otherwise} \end{cases} && \text{(B.4)}\\
\text{H}[x] &= -\mu \ln \mu - (1-\mu)\ln(1-\mu). && \text{(B.5)}
\end{aligned}$$

The Bernoulli is a special case of the binomial distribution for the case of a single observation. Its conjugate prior for μ is the beta distribution.

Bishop (2006)

Beta distribution

- Outputs a single scalar value between 0 and 1 (inclusive); e.g. a probability



Beta

This is a distribution over a continuous variable $\mu \in [0, 1]$, which is often used to represent the probability for some binary event. It is governed by two parameters a and b that are constrained by $a > 0$ and $b > 0$ to ensure that the distribution can be normalized.

$$\text{Beta}(\mu|a, b) = \frac{\Gamma(a + b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1 - \mu)^{b-1} \tag{B.6}$$

$$\mathbb{E}[\mu] = \frac{a}{a + b} \tag{B.7}$$

$$\text{var}[\mu] = \frac{ab}{(a + b)^2(a + b + 1)} \tag{B.8}$$

$$\text{mode}[\mu] = \frac{a - 1}{a + b - 2}. \tag{B.9}$$

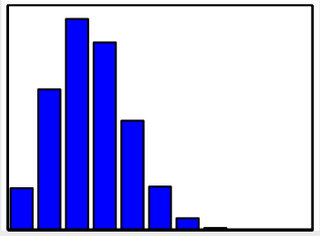
The beta is the conjugate prior for the Bernoulli distribution, for which a and b can be interpreted as the effective prior number of observations of $x = 1$ and $x = 0$, respectively. Its density is finite if $a \geqslant 1$ and $b \geqslant 1$, otherwise there is a singularity at $\mu = 0$ and/or $\mu = 1$. For $a = b = 1$, it reduces to a uniform distribution. The beta distribution is a special case of the K-state Dirichlet distribution for $K = 2$.

Bishop (2006)

Binomial distribution

- The probability of observing m occurrences with a particular outcome, from a set of N samples
- Outputs a count (leq N)

Binomial



The binomial distribution gives the probability of observing m occurrences of $x = 1$ in a set of N samples from a Bernoulli distribution, where the probability of observing $x = 1$ is $\mu \in [0, 1]$.

$$\mathrm{Bin}(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m} \tag{B.10}$$

$$\mathbb{E}[m] = N\mu \tag{B.11}$$

$$\mathrm{var}[m] = N\mu(1-\mu) \tag{B.12}$$

$$\mathrm{mode}[m] = \lfloor (N+1)\mu \rfloor \tag{B.13}$$

where $\lfloor (N+1)\mu \rfloor$ denotes the largest integer that is less than or equal to $(N+1)\mu$, and the quantity

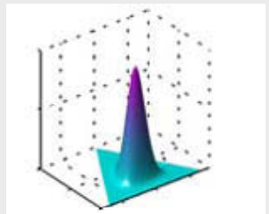
$$\binom{N}{m} = \frac{N!}{m!(N-m)!} \tag{B.14}$$

denotes the number of ways of choosing m objects out of a total of N identical objects. Here $m!$, pronounced 'factorial m', denotes the product $m \times (m-1) \times \ldots, \times 2 \times 1$. The particular case of the binomial distribution for $N = 1$ is known as the Bernoulli distribution, and for large N the binomial distribution is approximately Gaussian. The conjugate prior for μ is the beta distribution.

Bishop (2006)

Dirichlet distribution

- Outputs a vector of mixing proportions that sums to 1



Dirichlet

The Dirichlet is a multivariate distribution over K random variables $0 \leqslant \mu_k \leqslant 1$, where $k = 1, \ldots, K$, subject to the constraints

$$0 \leqslant \mu_k \leqslant 1, \qquad \sum_{k=1}^{K} \mu_k = 1. \tag{B.15}$$

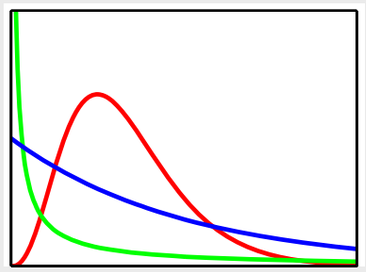
Denoting $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^{\mathrm{T}}$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^{\mathrm{T}}$, we have

$$\begin{aligned}
\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) &= C(\boldsymbol{\alpha}) \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} && \text{(B.16)}\\
\mathbb{E}[\mu_k] &= \frac{\alpha_k}{\widehat{\alpha}} && \text{(B.17)}\\
\mathrm{var}[\mu_k] &= \frac{\alpha_k(\widehat{\alpha} - \alpha_k)}{\widehat{\alpha}^2(\widehat{\alpha} + 1)} && \text{(B.18)}\\
\mathrm{cov}[\mu_j \mu_k] &= -\frac{\alpha_j \alpha_k}{\widehat{\alpha}^2(\widehat{\alpha} + 1)} && \text{(B.19)}\\
\mathrm{mode}[\mu_k] &= \frac{\alpha_k - 1}{\widehat{\alpha} - K} && \text{(B.20)}\\
\mathbb{E}[\ln \mu_k] &= \psi(\alpha_k) - \psi(\widehat{\alpha}) && \text{(B.21)}\\
\mathrm{H}[\boldsymbol{\mu}] &= -\sum_{k=1}^{K} (\alpha_k - 1)\{\psi(\alpha_k) - \psi(\widehat{\alpha})\} - \ln C(\boldsymbol{\alpha}) && \text{(B.22)}
\end{aligned}$$

Bishop (2006)

Gamma distribution

- Outputs a positive number



Gamma

The Gamma is a probability distribution over a positive random variable $\tau > 0$ governed by parameters a and b that are subject to the constraints $a > 0$ and $b > 0$ to ensure that the distribution can be normalized.

$$\mathrm{Gam}(\tau|a,b) = \frac{1}{\Gamma(a)} b^a \tau^{a-1} e^{-b\tau} \tag{B.26}$$

$$\mathbb{E}[\tau] = \frac{a}{b} \tag{B.27}$$

$$\mathrm{var}[\tau] = \frac{a}{b^2} \tag{B.28}$$

$$\mathrm{mode}[\tau] = \frac{a-1}{b} \quad \text{for } \alpha \geqslant 1 \tag{B.29}$$

$$\mathbb{E}[\ln \tau] = \psi(a) - \ln b \tag{B.30}$$

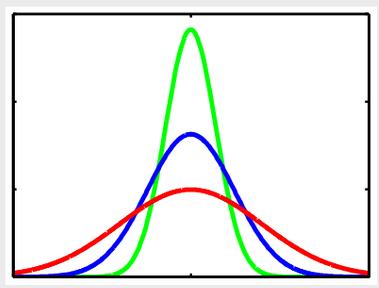
$$\mathrm{H}[\tau] = \ln \Gamma(a) - (a-1)\psi(a) - \ln b + a \tag{B.31}$$

where $\psi(\cdot)$ is the digamma function defined by (B.25). The gamma distribution is the conjugate prior for the precision (inverse variance) of a univariate Gaussian. For $a \geqslant 1$ the density is everywhere finite, and the special case of $a = 1$ is known as the *exponential* distribution.

Bishop (2006)

Gaussian distribution

- Outputs a Real number (or a vector of Real numbers)



Gaussian

The Gaussian is the most widely used distribution for continuous variables. It is also known as the *normal* distribution. In the case of a single variable $x \in (-\infty, \infty)$ it is governed by two parameters, the mean $\mu \in (-\infty, \infty)$ and the variance $\sigma^2 > 0$.

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{B.32}$$

$$\mathbb{E}[x] = \mu \tag{B.33}$$

$$\mathrm{var}[x] = \sigma^2 \tag{B.34}$$

$$\mathrm{mode}[x] = \mu \tag{B.35}$$

$$\mathrm{H}[x] = \frac{1}{2}\ln\sigma^2 + \frac{1}{2}\left(1 + \ln(2\pi)\right). \tag{B.36}$$

The inverse of the variance $\tau = 1/\sigma^2$ is called the precision, and the square root of the variance σ is called the standard deviation. The conjugate prior for μ is the Gaussian, and the conjugate prior for τ is the gamma distribution. If both μ and τ are unknown, their joint conjugate prior is the Gaussian-gamma distribution.

For a D-dimensional vector $\mathbf{x}$, the Gaussian is governed by a D-dimensional mean vector $\boldsymbol{\mu}$ and a $D \times D$ covariance matrix $\boldsymbol{\Sigma}$ that must be symmetric and

Bishop (2006)

Multinomial distribution

- Outputs a vector of binary outcomes (multivariate version of the Binomial distribution)

Multinomial

If we generalize the Bernoulli distribution to an K-dimensional binary variable $\mathbf{x}$ with components $x_k \in \{0, 1\}$ such that $\sum_k x_k = 1$, then we obtain the following discrete distribution

$$p(\mathbf{x}) = \prod_{k=1}^{K} \mu_k^{x_k} \tag{B.54}$$

$$\mathbb{E}[x_k] = \mu_k \tag{B.55}$$

$$\mathrm{var}[x_k] = \mu_k(1 - \mu_k) \tag{B.56}$$

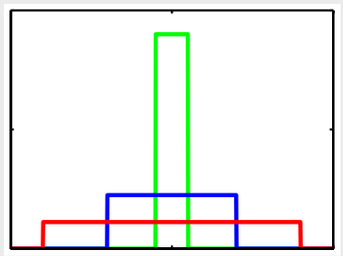
$$\mathrm{cov}[x_j x_k] = I_{jk}\mu_k \tag{B.57}$$

$$\mathrm{H}[\mathbf{x}] = -\sum_{k=1}^{M} \mu_k \ln \mu_k \tag{B.58}$$

Bishop (2006)

Uniform distribution

- Outputs a scalar number that falls within a defined interval



Uniform

This is a simple distribution for a continuous variable x defined over a finite interval $x \in [a, b]$ where $b > a$.

$$\mathrm{U}(x|a,b) = \frac{1}{b-a} \quad \text{(B.73)}$$

$$\mathbb{E}[x] = \frac{(b+a)}{2} \quad \text{(B.74)}$$

$$\mathrm{var}[x] = \frac{(b-a)^2}{12} \quad \text{(B.75)}$$

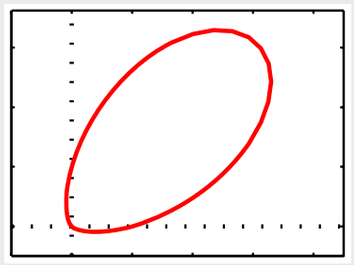
$$\mathrm{H}[x] = \ln(b-a). \quad \text{(B.76)}$$

If x has distribution $\mathrm{U}(x|0,1)$, then $a + (b-a)x$ will have distribution $\mathrm{U}(x|a,b)$.

Bishop (2006)

Von Mises distribution

- Outputs an angle (kind of like a Gaussian but for angles)



Von Mises

The von Mises distribution, also known as the circular normal or the circular Gaussian, is a univariate Gaussian-like periodic distribution for a variable $\theta \in [0, 2\pi)$.

$$p(\theta|\theta_0, m) = \frac{1}{2\pi I_0(m)} \exp\{m\cos(\theta - \theta_0)\} \quad \text{(B.77)}$$

where $I_0(m)$ is the zeroth-order Bessel function of the first kind. The distribution has period 2π so that $p(\theta + 2\pi) = p(\theta)$ for all θ. Care must be taken in interpreting this distribution because simple expectations will be dependent on the (arbitrary) choice of origin for the variable θ. The parameter θ_0 is analogous to the mean of a univariate Gaussian, and the parameter $m > 0$, known as the *concentration* parameter, is analogous to the precision (inverse variance). For large m, the von Mises distribution is approximately a Gaussian centred on θ_0.

Bishop (2006)

Wishart distribution

- Outputs a positive semi-definite matrix (e.g. a covariance matrix)

Wishart

The Wishart distribution is the conjugate prior for the precision matrix of a multivariate Gaussian.

$$\mathcal{W}(\mathbf{\Lambda}|\mathbf{W},\nu) = B(\mathbf{W},\nu)|\mathbf{\Lambda}|^{(\nu-D-1)/2}\exp\left(-\frac{1}{2}\mathrm{Tr}(\mathbf{W}^{-1}\mathbf{\Lambda})\right) \tag{B.78}$$

where

$$B(\mathbf{W},\nu) \equiv |\mathbf{W}|^{-\nu/2}\left(2^{\nu D/2}\,\pi^{D(D-1)/4}\prod_{i=1}^{D}\Gamma\left(\frac{\nu+1-i}{2}\right)\right)^{-1} \tag{B.79}$$

$$\mathbb{E}[\mathbf{\Lambda}] = \nu\mathbf{W} \tag{B.80}$$

$$\mathbb{E}\left[\ln|\mathbf{\Lambda}|\right] = \sum_{i=1}^{D}\psi\left(\frac{\nu+1-i}{2}\right) + D\ln 2 + \ln|\mathbf{W}| \tag{B.81}$$

$$\mathrm{H}[\mathbf{\Lambda}] = -\ln B(\mathbf{W},\nu) - \frac{(\nu-D-1)}{2}\mathbb{E}\left[\ln|\mathbf{\Lambda}|\right] + \frac{\nu D}{2} \tag{B.82}$$

where $\mathbf{W}$ is a $D \times D$ symmetric, positive definite matrix, and $\psi(\cdot)$ is the digamma function defined by (B.25). The parameter ν is called the *number of degrees of freedom* of the distribution and is restricted to $\nu > D-1$ to ensure that the Gamma function in the normalization factor is well-defined. In one dimension, the Wishart reduces to the gamma distribution $\mathrm{Gam}(\lambda|a,b)$ given by (B.26) with parameters $a = \nu/2$ and $b = 1/2W$.

Bishop (2006)

Distributions

- These distributions are like LEGO building blocks that can be combined to do increasingly more complicated things
- The output of one distribution can be the input to one (or more!) new distributions

Fitting models

- Sometimes Bayes' Theorem can be directly applied to your problem. This yields a posterior distribution over possible parameters and states, given the observed data.
- Often Bayes' Theorem cannot be used directly due to computational tractability issues

Fitting models

sum rule $$p(X) = \sum_Y p(X, Y)$$

product rule $$p(X, Y) = p(Y|X)p(X)$$

- The sum rule tells us how to compute p(X)— you need to sum over all possible values of Y (i.e. all possible combinations of parameters that we're trying to learn about). Remember: p(X) is the "normalizing term" in Bayes' Theorem.

- Often there are too many possible values of Y to reasonably sum over. So instead we use **approximate inference** (i.e. we try to approximate the posterior in a tractable way).