ML Preparation Notes

March 28, 2021

Chapter 1

Basic Statistics

In this section, we review some basic concepts from Statistics. The material in this section is based on [4].

1.1 Covariance and Correlation

The *correlation* between two sets of data is a measure of the strength of the relationship between them. In particular, Pearson's correlation coefficient is a measure of linear relationship between two sets of data. Let X and Y be two random variables. Then Pearson's correlation coefficient $\rho(X, Y)$ is defined as:

$$\rho(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \, \sigma_Y} \tag{1.1}$$

Two important facts about the Pearson's correlation coefficient [3]:

1. $-1 \leq \rho(X, Y) \leq 1$
2. $|\rho(X, Y)| = 1$ iff there exists $a \neq 0$ and b such that $Y = aX + b$.

1.2 Tail Inequalities and the Law of Large Numbers

We first look at Markov and Chebyshev inequalities and then apply them to random samples.

The Markov inequality tells us how much probability mass can be at large values given the mean of the distribution.

Theorem 1.1. *Let X be a random variable that takes on positive values only. Then for $t > 0$,*

$$\text{P}(X \geq t) \leq \frac{\text{E}[X]}{t}. \tag{1.2}$$

Proof. Let us assume that X is a continuous r.v. with pdf f_X. By definition,

$$\begin{aligned} \text{E}[X] &= \int_0^\infty x f_X(x) dx \\ &= \int_0^t x f_X(x) dx + \int_t^\infty x f_X(x) dx \\ &\geq t \int_t^\infty f_X(x) dx. \end{aligned}$$

Since $t > 0$, dividing both sides of the last inequality yields the Markov inequality. □

This inequality is useful when $t > \mathrm{E}[X]$; when $t \leq \mathrm{E}[X]$, it merely bounds $\mathrm{P}(X \geq t)$ by 1.

The Chebyschev inequality involves both the mean and the variance of the distribution. It bounds the probability of how far a random variable can be from its mean as a function of the variance.

Theorem 1.2. *Let X be a random variable for which $\mathrm{Var}(X)$ exists. Then*

$$\mathrm{P}(|X - \mathrm{E}[X]| > t) \leq \frac{\mathrm{Var}(X)}{t^2}. \tag{1.3}$$

Proof. Define $Y \coloneqq (X - \mathrm{E}[X])^2$ so that $\mathrm{E}[Y] = \mathrm{Var}(X)$. Applying the Markov inequality to the r.v. Y, we obtain:

$$\mathrm{P}(Y \geq t^2) \leq \frac{\mathrm{Var}(X)}{t^2}.$$

But $\mathrm{P}(Y \geq t^2) = \mathrm{P}(|X - \mathrm{E}[X]| > t)$ and so this proves the Chebyschev inequality too. □

To talk about the law of large numbers, one has to talk about the notion of a sequence of random variables converging to a real number. We say that a sequence $Z_1, Z_2, \ldots$ of random variables converges to the number b if the distribution of Z_n as $n \to \infty$ becomes more and more concentrated around this single number.

Definition 1.1 (Convergence in Probability)**.** A sequence $Z_1, Z_2, \ldots$ of random variables converges in probability to the number b if for every $\epsilon > 0$, the following condition holds:

$$\lim_{n \to \infty} \mathrm{P}(|Z_n - b| < \epsilon) = 1. \tag{1.4}$$

This fact is denoted by $Z_n \xrightarrow{p} b$.

Theorem 1.3 (The Law of Large Numbers)**.** *Let $X_1, \ldots, X_n$ be a random sample from a distribution with mean μ and finite variance σ^2. Then the sample mean $\bar{X}_n$ converges in probability to μ.*

Proof. We have $\mathrm{E}[\bar{X}_n] = \mu$ and $\mathrm{Var}(\bar{X}_n) = \sigma^2/n$. Use the Chebyschev inequality to obtain:

$$\mathrm{P}(|\bar{X}_n - \mu| \leq t) \geq 1 - \frac{\sigma^2}{nt^2}.$$

For every fixed $t > 0$, take limits as $n \to \infty$ to obtain: $\lim_{n\to\infty} \mathrm{P}(|\bar{X}_n - \mu| \leq t) = 1$. Hence $\bar{X} \xrightarrow{p} \mu$. □

If we know that a sequence of random variables converges in probability, what can we say about continuous functions of that sequence of random variables?

Theorem 1.4 (Continuous Functions of Random Variables)**.** *Suppose a sequence $Z_1, Z_2, \ldots$ converges in probability to b and that g is a function continuous at b. Then $g(Z_n) \xrightarrow{p} g(b)$.*

Proof. We have to show $\lim_{n\to\infty} \mathrm{P}(|g(Z_n) - g(b)| < \epsilon) = 1$. Since g is continuous at b, given $\epsilon > 0$ there exists $\delta > 0$ such that $|z_n - b| < \delta$ implies $|g(z_n) - g(b)| < \epsilon$. Now consider the events $\mathcal{E}_1 = \{|Z_n - b| < \delta\}$ and $\mathcal{E}_2 = \{|g(Z_n) - g(b)| < \epsilon\}$. Since $\mathcal{E}_1$ implies $\mathcal{E}_2$, we have that $\mathrm{P}(\mathcal{E}_2) \geq \mathrm{P}(\mathcal{E}_1)$. Since $\lim_{n\to\infty} \mathrm{P}(\mathcal{E}_1) = 1$, we must have $\lim_{n\to\infty} \mathrm{P}(\mathcal{E}_2) = 1$. □

1.3 Maximum Likelihood Estimation

Let the random variables $X_1, \ldots, X_n$ form a random sample from a distribution with pdf $f(x \mid \theta)$. Recall that this means that $X_i \overset{\text{iid}}{\sim} f(\cdot \mid \theta)$ for all $1 \leq i \leq n$. Let $f_n(\vec{x} \mid \theta)$ denote the value of the joint pdf of the random vector $(X_1, \ldots, X_n)'$ at the point $\vec{x} = (x_1, \ldots, x_n)'$. The *likelihood function* is the joint pdf of the observations of a random sample viewed as a function of θ for a given set of values of the sample. The maximum likelihood estimate of θ is that value of θ for which $f_n(\vec{x} \mid \theta)$ is maximized.

Example 1.1. Suppose that $X_1, \ldots, X_n$ form a random sample from a distribution with pdf $f(x \mid \theta)$ defined as follows:

$$f(x \mid \theta) = \begin{cases} \theta x^{\theta-1} & \text{for } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Let us assume that $\theta > 0$. The joint distribution of the vector $(X_1, \ldots, X_n)'$ at the point $\vec{x} = (x_1, \ldots, x_n)'$ is

$$f_n(\vec{x} \mid \theta) = \prod_{i=1}^{n} \theta x_i^{\theta-1},$$

where we assume that $0 < x_i < 1$ for all i. Taking logs, we obtain that $\log f_n(\vec{x} \mid \theta) = n \log \theta + (\theta - 1) \sum_i \log x_i$. Take the derivative wrt θ and set to 0:

$$\frac{\partial \log f_n(\vec{x} \mid \theta)}{\partial \theta} = \frac{n}{\theta} + \sum_i \log x_i \overset{\text{set}}{=} 0$$

to obtain that $\theta = n / \sum_i \log \frac{1}{x_i}$.

Example 1.2. Suppose that $X_1, \ldots, X_n$ form a random sample from a distribution with pdf $f(x \mid \theta)$ defined as follows:

$$f(x \mid \theta) = \frac{1}{2} e^{-|x-\theta|} \quad \text{for } -\infty < x < \infty.$$

Suppose that θ is unknown and that $-\infty < \theta < \infty$. In this case, the joint probability distribution is easily seen to be:

$$f_n(\vec{x} \mid \theta) = \frac{1}{2^n} e^{-\sum_i |x_i - \theta|}.$$

Take logs to obtain: $\log f_n(\vec{x} \mid \theta) = \log \frac{1}{2^n} - \sum_i |x_i - \theta|$. Maximizing f_n is equivalent to minimizing $\sum_i |x_i - \theta|$. This is equivalent to obtaining a point on the real line that minimizes the sum of the distances to the points $x_1, \ldots, x_n$. This happens when θ is the median of $x_1, \ldots, x_n$.

Example 1.3. Suppose that $X_1, \ldots, X_n$ form a random sample from the uniform distribution on the interval $[\theta_1, \theta_2]$, where both θ_1 and θ_2 are unknown ($-\infty < \theta_1 < \theta_2 < \infty$). In this case, the log pdf of the joint distribution of $(X_1, \ldots, X_n)'$ is given by

$$\log f_n(\vec{x} \mid \theta_1, \theta_2) = \log \prod_{i=1}^{n} \frac{1}{\theta_2 - \theta_1} = -n \log(\theta_2 - \theta_1).$$

Maximizing the likelihood is equivalent to minimizing $\log(\theta_2 - \theta_1)$. The minimum possible value of θ_2 is $\max\{x_1, \ldots, x_n\}$ and the maximum possible value of θ_1 is $\min\{x_1, \ldots, x_n\}$.

Example 1.4. Suppose that a certain large population contains k different types of individuals ($k \geq 2$), and let θ_i denote the proportion of people of type i, for $1 \leq i \leq k$. Here, $0 \leq \theta_i \leq 1$ and $\sum_{i=1}^{k} \theta_i = 1$.

Suppose also that in a random sample of n individuals from this population there are exactly n_i individuals of type i so that $n = n_1 + \cdots + n_k$.

In this setting, for $1 \leq i \leq k$, define X_i to be the number of individuals of type i in a random sample of size n. Then the probability that $\bigwedge_{i=1}^{k} X_i = n_i$ is given by $\theta_1^{n_1} \cdots \theta_k^{n_k}$. The log pdf of the joint distribution is given by:

$$\log f_k((n_1, \ldots, n_k) \mid \theta_1, \ldots, \theta_k, n) = \sum_{i=1}^{k} n_i \log \theta_i.$$

Note that there are actually $k-1$ variables here since we may write $\theta_k = 1 - \sum_{i=1}^{k-1} \theta_i$. Differentiating wrt θ_i for $1 \leq i \leq k-1$, we obtain:

$$\frac{\partial \log f_k}{\partial \theta_i} = \frac{n_i}{\theta_i} - \frac{n_k}{\theta_k}.$$

Setting this to 0, we get that $\theta_i/\theta_k = n_i/n_k$. Sum this up from $1 \leq i \leq k-1$, to obtain:

$$\frac{1-\theta_k}{\theta_k} = \frac{n-n_k}{n_k},$$

which yields $\theta_k = n_k/n$. Substitute this in $\theta_i/\theta_k = n_i/n_k$ to obtain $\theta_i = n_i/n$.

Example 1.5 (Nonexistence of an MLE)**.** An obvious disadvantage of the technique of maximum likelihood estimation is when the maximum does not exist. Consider again Example 1.3 where we let $\theta_1 = 0$ and $\theta = \theta_2$ The pdf of the uniform distribution is defined as:

$$f(x \mid \theta) = \begin{cases} \frac{1}{\theta} & 0 \leq x \leq \theta \\ 0 & \text{otherwise.} \end{cases}$$

Let us modify the definition of the pdf so that we use strict inequalities $0 < x < \theta$ above. Given a sample $x_1, \ldots, x_n$, the log pdf is $-n \log \theta$ as before and the MLE technique would require us to minimize $\log \theta$. However, in this case, there is no $\theta > \max\{x_1, \ldots, x_n\}$ that mimimizes $\log \theta$ and the MLE does not exist.

Example 1.6 (Non-uniqueness of an MLE)**.** Consider a random sample $X_1, \ldots, X_n$ from the uniform distribution over the interval $[\theta, \theta+1]$. In this case, the joint pdf is given by:

$$f_n(\vec{x} \mid \theta) = \begin{cases} 1 & \text{for } \theta \leq x_i \leq \theta + 1 \quad (1 \leq i \leq n) \\ 0 & \text{otherwise.} \end{cases}$$

In this case, the condition $\theta \leq x_i \leq \theta + 1$ for $1 \leq i \leq n$ may be written using the two conditions:

$$\theta \leq \min\{x_1, \ldots, x_n\} \text{ and } \max\{x_1, \ldots, x_n\} - 1 \leq \theta.$$

Any value of θ in the interval $[\max\{x_1, \ldots, x_n\} - 1, \min\{x_1, \ldots, x_n\}]$ is valid, but there is no unique value of θ.

1.4 Bayesian Statistics and MCMC

This section is based on Chapters 12–15 from [6]. Bayes' rule gives us a recipe for calculating the posterior probability density.

$$\mathrm{P}(\Theta \mid \text{data}) = \frac{\mathrm{P}(\text{data} \mid \Theta) \cdot \mathrm{P}(\Theta)}{\mathrm{P}(\text{data})}. \tag{1.5}$$

Consider a case in which we have a sample of N data points $x_1, \ldots, x_N$. We assume that the likelihood is a Poisson distribution with mean λ and that the prior for λ is a log-normal$(1, 1)$ distribution. To calculate the probability of the data P(data), we must evaluate the integral:

$$\text{P(data)} = \int_0^\infty \prod_{i=1}^{N} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} \cdot \frac{1}{\sqrt{2\pi}\lambda} e^{-\frac{1}{2}(\log \lambda - 1)^2} d\lambda. \tag{1.6}$$

While this integral is not too difficult, it explains the problem of calculating posteriors analytically. As the number of parameters (the length of Θ) increases, calculating the probability of the data requires evaluating integrals in higher dimensional spaces. This is why we use alternative methods to derive approximate versions of the posterior.

Chapter 2

Timeseries

The material in the section follows [2]. A *time series* is a sequence of random variables $\{X_1, X_2, \ldots\}$. A complete model of such a time series would require specifying the joint probability distributions of all random vectors $(X_1, \ldots, X_n)'$ for all $n \geq 1$. In practice, this might be impossible to do unless the time series is generated by some simple well-understood mechanism. Instead one typically specifies the first- and second-order moments of the joint distibutions, that is, $\mathrm{E}[X_t]$ and $\mathrm{E}[X_t X_{t+h}]$ for all $t \geq 1$ and for all $h \geq 0$.

The next important concepts are those of stationarity and the auto-correlation function. Roughly speaking, a time series $\{X_t\}_{t=-\infty}^{\infty}$ is stationary if its "statistical properties" are similar to those of the time-shifted series $\{X_{t+h}\}_{t=-\infty}^{\infty}$ for every integer "lag" h. By statistical properties, we mean the first- and second-order moments of $\{X_t\}$.

Formally, a time series $\{X_t\}$ is weakly stationary if

1. $\mathrm{E}[X_t]$ is independent of t
2. $\mathrm{Cov}(X_t, X_{t+h})$ is independent of t for every fixed lag h.

In contrast, a time series $\{X_t\}$ is strictly stationary if the random vectors $(X_{t_1}, \ldots, X_{t_n})'$ and $(X_{t_1+h}, \ldots, X_{t_n+h})'$ have the same joint distributions for all sets of indices $\{t_1, \ldots, t_n\}$, for all $h \geq 0$ and all $n \geq 1$. This is written as:

$$(X_{t_1}, \ldots, X_{t_n})' \stackrel{d}{=} (X_{t_1+h}, \ldots, X_{t_n+h})'$$

Strict stationarity implies the following:

1. The random variables X_t are identically distributed.
2. Pairs of random variables (X_t, X_{t+h}) have the same distribution as (X_1, X_{1+h}) (set $n = 2$).
3. Strict stationarity implies weak stationarity: $\mathrm{E}[X_t] = \mathrm{E}[X_1]$ and $\mathrm{Cov}(X_t, X_{t+h}) = \mathrm{Cov}(X_1, X_{1+h})$ for all $t \geq 1$ and all $h \geq 0$. Both terms are independent of t.
4. Weak stationarity does *not* imply strong stationarity. We show this by an example. Let $Z_i \stackrel{\text{iid}}{\sim} N(0, 1)$ for all i. Define X_t as:

$$X_t = \begin{cases} Z_t & \text{if } t \text{ is even} \\ 2Z_t & \text{if } t \text{ is odd.} \end{cases}$$

Then $\mathrm{E}[X_t] = 0$ for all t. Also,

$$\mathrm{Cov}(X_t, X_{t+h}) = \begin{cases} 0 & \text{if } h > 0 \\ 1 & \text{if } h = 0. \end{cases}$$

This follows from the assumed independence of X_t and X_{t+h} when $h > 0$. However, X_ts do not have the same distribution, a requirement of strict stationarity.

The autocovariance function of a stationary time series $\{X_t\}$ at lag h is defined as $\mathrm{Cov}(X_{t+h}, X_t) = \mathrm{Cov}(X_h, X_0)$. The autocorrelation of $\{X_t\}$ at lag h is defined as

$$\frac{\mathrm{Cov}(X_{t+h}, X_t)}{\mathrm{Var}(X_t)} = \frac{\mathrm{Cov}(X_h, X_0)}{\mathrm{Var}(X_0)}.$$

Example 2.1 (iid Noise)**.** Suppose $\{X_t\}$ is iid noise with zero mean and $\mathrm{E}[X_t^2] = \sigma^2 < \infty$. Then $\mathrm{E}[X_t] = 0$ for all t and for every fixed lag h and all t:

$$\mathrm{Cov}(X_t, X_{t+h}) = \begin{cases} 0 & \text{if } h > 0 \\ \sigma^2 & \text{if } h = 0. \end{cases}$$

Thus iid noise is stationary. We denote such a series as $\{X_t\} \sim \mathrm{IID}(0, \sigma^2)$.

Example 2.2 (White Noise)**.** A sequence $\{X_t\}$ of *uncorrelated* random variables with zero mean and finite second moment σ^2 is called *white noise*. Clearly, the covariance function at lag h is the same as that of iid noise and, as such, white noise is stationary. We denote such a series as $\{X_t\} \sim \mathrm{WN}(0, \sigma^2)$. Unlike iid noise, the components of white noise need not be independent (recall that independence implies zero correlation but not the other way around). In particular, every $\mathrm{IID}(0, \sigma^2)$ sequence is a $\mathrm{WN}(0, \sigma^2)$ sequence but not the other way around.

Example 2.3 (Random Walk)**.** A *random walk* $\{S_t\}$ is a sequence obtained by cumulatively summing iid random variables. A random walk with zero mean is obtained by defining $S_0 = 0$ and

$$S_t = X_1 + \cdots + X_t$$

for all $t > 0$, where $\{X_t\} \sim \mathrm{IID}(0, \sigma^2)$. In this case, $\mathrm{E}[S_t] = 0$ and $\mathrm{E}[S_t^2] = t\sigma^2 < \infty$ for all t. For all lags $h \geq 0$ and all t,

$$\begin{aligned} \mathrm{Cov}(S_t, S_{t+h}) &= \mathrm{Cov}(S_t, S_t + X_{t+1} + \cdots + X_{t+h}) \\ &= \mathrm{Cov}(S_t, S_t) + \mathrm{Cov}(S_t, X_{t+1}) + \cdots + \mathrm{Cov}(S_t, X_{t+h}) \\ &= t\sigma^2. \end{aligned}$$

The last equality follows since $\mathrm{Cov}(S_t, X_{t+i}) = 0$ for all $i \geq 1$. Hence $\mathrm{Cov}(S_t, S_{t+h})$ depends on t and $\{S_t\}$ is not stationary.

Example 2.4 (First-Order Moving Average)**.** Consider the series defined by

$$X_t = Z_t + \theta Z_{t-1}, \quad t = 0, \pm 1, \pm 2, \ldots, \tag{2.1}$$

where $Z_t \sim \mathrm{WN}(0, \sigma^2)$ and θ is a real-valued constant. Now $\mathrm{E}[X_t] = 0$ and

$$\begin{aligned} \mathrm{E}[X_t^2] &= \mathrm{E}[(Z_t + \theta Z_{t-1})^2] \\ &= \mathrm{E}[Z_t^2 + 2\theta Z_t Z_{t-1} + \theta^2 Z_{t-1}^2] \\ &= \sigma^2 + 2\theta\,\mathrm{E}[Z_t Z_{t-1}] + \theta^2\sigma^2 \\ &= (1 + \theta^2)\sigma^2. \end{aligned}$$

The last equality follows since, Z_t and Z_{t-1} being uncorrelated, satisfy $\mathrm{Cov}(Z_t, Z_{t-1}) = 0$. Recall that $\mathrm{Cov}(Z_t, Z_{t-1}) = \mathrm{E}[Z_t Z_{t-1}] - \mathrm{E}[Z_t]\,\mathrm{E}[Z_{t-1}]$ and that being uncorrelated is sufficient for the expectation

of the product of two random variables to be equal to the product of their expectations. One can easily verify that for all t

$$\mathrm{Cov}(X_t, X_{t+h}) = \begin{cases} (1+\theta^2)\sigma^2 & \text{if } h = 0 \\ \theta\sigma^2 & \text{if } h = \pm 1 \\ 0 & \text{if } |h| \geq 2. \end{cases} \tag{2.2}$$

Thus the conditions of weak stationarity hold and the sequence $\{X_t\}$ is weakly stationary. The autocorrelation function is given by:

$$\rho(h) = \frac{\mathrm{Cov}(X_0, X_h)}{\mathrm{Var}(X_0)} = \begin{cases} 1 & \text{if } h = 0 \\ \frac{\theta}{1+\theta^2} & \text{if } h = \pm 1 \\ 0 & \text{if } |h| \geq 2. \end{cases} \tag{2.3}$$

Example 2.5 (First-Order Autoregression)**.** Let $\{X_t\}$ be a stationary series satisfying the equation:

$$X_t = \phi X_{t-1} + Z_t \quad t = 0, \pm 1, \pm 2, \ldots, \tag{2.4}$$

where $\{Z_t\} \sim \mathrm{WN}(0, \sigma^2)$, $|\phi| < 1$ and Z_t is uncorrelated with X_s for all $s < t$. As can be seen, $\mathrm{E}[X_t] = 0$ and for $h > 0$:

$$\begin{aligned} \mathrm{Cov}(X_{t+h}, X_t) &= \mathrm{Cov}(\phi X_{t+h-1} + Z_{t+h}, X_t) \\ &= \phi \,\mathrm{Cov}(X_{t+h-1}, X_t) + \mathrm{Cov}(Z_{t+h}, X_t) \\ &= \phi \,\mathrm{Cov}(X_{t+h-1}, X_t). \end{aligned}$$

From this, one can show that $\mathrm{Cov}(X_{t+h}, X_t) = \phi^h \,\mathrm{Cov}(X_t, X_t)$. By assumption, $\{X_t\}$ is stationary and hence $\mathrm{Cov}(X_{t+h}, X_t) = \phi^h \,\mathrm{Cov}(X_0, X_0)$ Next suppose that $h < 0$. Let $s = t + h = t - |h|$ so that $t = s + |h|$. Then $\mathrm{Cov}(X_{t+h}, X_t) = \mathrm{Cov}(X_s, X_{s+|h|}) = \phi^{|h|} \,\mathrm{Cov}(X_s, X_s) = \phi^{|h|} \,\mathrm{Cov}(X_0, X_0)$. The autocorrelation function at lag h is

$$\rho(h) = \frac{\phi^{|h|} \,\mathrm{Cov}(X_0, X_0)}{\mathrm{Cov}(X_0, X_0)} = \phi^{|h|}.$$

One can also obtain a closed-form expression for $\mathrm{Cov}(X_0, X_0)$. Since $\mathrm{Cov}(X_0, X_0) = \mathrm{Cov}(X_t, X_t)$ and

$$\mathrm{Cov}(X_t, X_t) = \mathrm{Cov}(\phi X_{t-1} + Z_t, \phi X_{t-1} + Z_t) = \phi^2 \,\mathrm{Cov}(X_0, X_0) + \mathrm{Cov}(Z_t, Z_t),$$

we obtain that $\mathrm{Cov}(X_0, X_0) = \sigma^2/(1 - \phi^2)$.

Chapter 3

Trees, Boosting and Random Forests

Chapter 4

Neural Networks

Bibliography

[1] Joseph K. Blitzstein, Jessica Hwang. *Introduction to Probability*, Second Edition, Chapman and Hall, 2019.

[2] Peter J. Brockwell, Richard A. Davis. *Introduction to Time Series and Forecasting*, Third Edition, Springer, 2016.

[3] George Casella, Roger L. Berger. *Statistical Inference*, Second Edition, Duxbury Advanced Series, 2001.

[4] Morris DeGroot, Mark J. Schervish. *Probability and Statistics*, Fourth Edition, Pearson, 2012.

[5] Rob J. Hyndman and George Athanasopoulos. *Forecasting: Principles and Practice*, Second Edition. Online book at: `https://otexts.com/fpp2/`

[6] Ben Lambert. *A Student's Guide to Bayesian Statistics*, Sage Publications, 2018.