An Introduction to Variational Inference

Carolyn Augusta

Department of Mathematics and Statistics
University of Guelph

May 19th 2015

1. Bayesian statistics in a nutshell

2. Bayesian linear regression

General philosophy

Frequentist statisticians don't assign a probability to an event before they perform an experiment

E.g. if they'd never seen a die before, a frequentist would roll it over and over and after some large number of rolls, would say the probability of seeing a '1' is roughly 1/6

Bayesian statisticians assign a probability to an event, then perform an experiment

E.g. look at the die before you roll it, see it has 6 faces, guess the probability of '1' is roughly 1/6. Then start rolling.

Why Bayesian statistics?

Frequentist statisticians look at long-range frequencies of events to derive probabilities

E.g. roll that die 24,000 times and count the number of '1's you see.

Bayesian statisticians can give you a rough idea of the probability of an event before they see any results

E.g. The probability that it will rain this afternoon

Notice the weather is not a repeatable event - you can't see the weather this afternoon more than once. So the frequentist paradigm doesn't really make sense to use here. You can't see this afternoon's weather 24,000 times and count how many times it rained.

Note - that's a controversial statement. It's just an example to get you thinking along these lines.

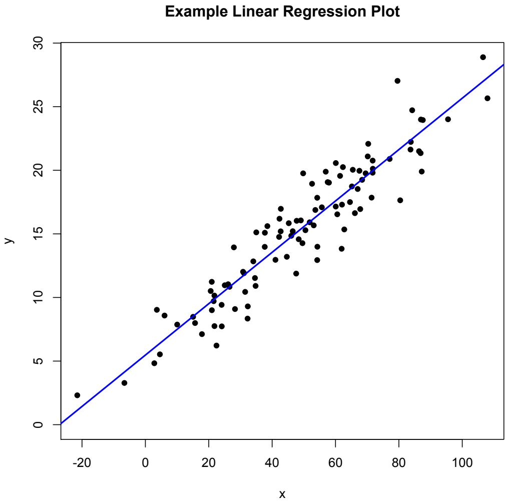
Also note that's just one use of Bayesian statistics, but it's the one I find most intuitive.

Bayesian inference

Often when we conduct an experiment, we're interested in finding a model describes the data

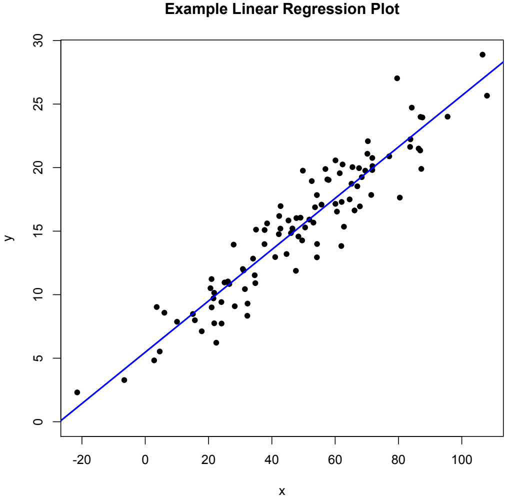
Suppose we expect a simple linear regression model will be a good fit to our data: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$. We want the line of best fit (find β_0 and β_1)

Let $\theta = (\beta_0, \beta_1)$



$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

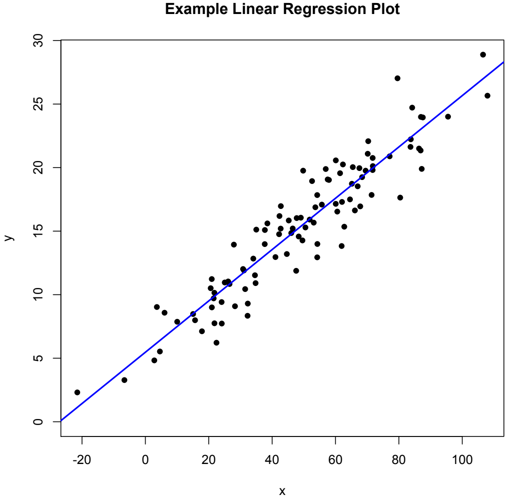
We want to find the best values of θ, given what we see in our data. So we want to infer θ from the conditional distribution of θ given our data. This is called the **posterior distribution**, and is denoted $\mathbf{P}(\theta \mid \mathbf{y})$



$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

posterior: $\mathbf{P}(\theta \mid \mathbf{y})$

Since we're working in a Bayesian framework, we have some guess of what β_0 and β_1 should be, before we even see the data y (maybe from previous work). So we have rough initial guess of the values of β_0 and β_1, and how they vary. This is called the **prior distribution** of θ, and is denoted $\mathbf{P}(\theta)$



$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

posterior: $\mathbf{P}(\theta \mid \mathbf{y})$
prior: $\mathbf{P}(\theta)$

We have data y, and we have some "best guess so far" values θ. We can look at the probability that the data we see was generated by a model with our "best guess" parameter values. This is called the **likelihood**, and is denoted $\mathbf{P}(\mathbf{y} \mid \theta)$

Now we need a relationship among all these distributions.

Bayes' Theorem from Conditional Probability

$$P(\theta \mid y) = \frac{P(y, \theta)}{P(y)} \tag{1}$$

by definition of conditional probability

$$P(y \mid \theta) = \frac{P(y, \theta)}{P(\theta)} \tag{2}$$

also by definition of conditional probability

$$P(y, \theta) = P(y \mid \theta)P(\theta) \tag{3}$$

rearranging (2)

$$P(\theta \mid y) = \frac{P(y \mid \theta)P(\theta)}{P(y)} \tag{4}$$

plugging (3) into (1) - Bayes' Theorem!

So what's the problem?

$$P(\theta \mid y) = \frac{P(y \mid \theta)P(\theta)}{P(y)}$$

Look again at the normalizing constant, $P(y)$

To get that value, we'd have to marginalize over all values θ could take: $P(y) = \int_\theta P(y \mid \theta)P(\theta)d\theta$

In a lot of cases, the number of parameters in our model is large, so we'd have to integrate over a huge space. The normalizing constant becomes computationally intractable, which makes the posterior intractable.

We need **approximate inference** methods to make conclusions about the parameters θ, based on the data y.

Approximate inference intro

There are two main frameworks for approximate Bayesian inference: Markov chain Monte Carlo (MCMC) and variational Bayes (VB) methods.
Today we'll go over VB with an example using Bayesian linear regression, so you can see how the mechanics work.
Then a more complicated example, when inference really is intractable.

1. Bayesian statistics in a nutshell

2. Bayesian linear regression

How do we make simple linear regression Bayesian?

Assume we have continuous data y and a single linear predictor x, like before

E.g. Say our experiment had x as temperature. A phenomenon that depends on temperature likely also depends on humidity, but we haven't put that in the model.

So our data y will vary by some amount ϵ, even if we hold x constant.

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{5}$$

where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

How do we make simple linear regression Bayesian?

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

Is the same as saying

$$Y_i \mid x_i = \beta_0 + \beta_1 x_i + \epsilon_i \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

$$E(Y_i \mid x_i) = E(\beta_0 + \ \beta_1 x_i + \epsilon_i) \quad = E(\beta_0) + E(\beta_1 x_i) + E(\epsilon_i) = \beta_0 + \beta_1 x_i$$

$$Var(Y_i \mid x_i) = Var(\beta_0 + \beta_1 x_i + \epsilon_i) \qquad = Var(\epsilon_i) = \sigma^2$$

$$P(Y_i \mid x_i) = N(\beta_0 + \beta_1 x_i, \sigma^2)$$

Outline

Bayesian philosophy

Bayes' Theorem

The posterior distribution

The normalizing constant

The marginal likelihood

The general idea

The algorithm

Where the algorithm came from

Bayesian mixture models

A step-by-step example using Bayesian mixture models

SIR Compartmental Model

Disease transmission and recovery can be described via a pure-birth process called an SIR model.

Suppose all individuals are susceptible to the disease (they are in state "S")

They become infectious according to a certain probability (move to state "I")

After γ time periods, they are removed from the population due to recovery, acquired immunity, etc. (state "R")

Individual-Level Models

P(infection in [t, t+1)) $= 1 - exp(-\omega_{it})$

$\omega_{it} = \alpha \sum_{j \in I(t)} \mathrm{d_{ij}}^{-\beta}$, $j \in I(t)$ if $t - \gamma \leq \mathrm{inf(j)} < t$

α: susceptibility, set prior: $\alpha \sim U(0, 50)$

$I(t)$: set of infectious individuals at time t

d_{ij}: (Euclidean) distance between susceptible i and infectious j

β: spatial effect, set prior: $\beta \sim U(-5, 50)$

γ: infectious period, set prior: $\gamma \sim DU[1, 10]$

References

Deardon, R. et al. (2010). Inference for individual-level models of infectious diseases in large populations. Statistica Sinica, 20, 239-261.

Gold, J. (2015). Computational inference for network-based individual-level models of infectious disease transmission (Unpublished doctoral dissertation). University of Guelph, Guelph Ontario Canada.