Bayesian Methods with Financial Applications

Ruei-Chi Lee

Trading Valley

November 29, 2018

Outline

Introduction

- The frequentist/Bayesian divide is fundamentally a question of philosophy: the definition of probability

Probability

- Objective view of probability (Frequentists)
  - The relative frequency of an outcome of an experiment over repeated runs of the experiment.
  - The observed proportion in a population.
- Subjective view of probability (Bayesians)
  - Individuals degree of belief in a statement
  - Can be influenced in many ways (personal beliefs, prior evidence)

- The Frequentist approach:
  - Observe data which were generated randomly
  - We made assumptions on the generating process (e.g., i.i.d., Gaussian data, linear regression function, etc...)
  - The generating process was associated to some object of interest (e.g., a parameter, a density, etc...)
  - This object was unknown but fixed and we wanted to find it: we either estimated it or tested a hypothesis about this object

- The Bayesian approach:
  - Now, we still observe data, assumed to be randomly generated by some process. Under some assumptions (e.g., parametric distribution), this process is associated with some fixed object.
  - We have a prior belief about it.
  - Using the data, we want to update that belief and transform it into a posterior belief.

Bayes' Theorem

$$P(A|B) = \frac{P(B \cap A)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

- Often compute denominator P(B) using the law of total probability
- links the degree of belief in a proposition before and after accounting for evidence
- For proposition A and evidence B:
  - $P(A)$, the prior, is the initial degree of believe in A
  - $P(A|B)$, the posterior, is the degree of belief having accounted for B

Example: Monte Hall Problem

- Rules:
  - One door hides a car, two hides goats.
  - The contestant chooses any door.
  - The host always opens a different door with a goat. (He can do this because he knows where the car is.)
  - The contestant is then allowed to switch doors if he/she wants.
- Question: what is the best strategy for winning a car?
  (a) Switch (b) Don't switch

Example: Monte Hall Problem

How to incorporate Bayes' theorem into this problem?

Now assume the contestant chose the 1st door, and the host opened the 2nd door with a goat.

- Prior: $P(A_1) = P(A_2) = P(A_3) = \frac{1}{3}$
- Evidence: $P(B|A_1) = \frac{1}{2}, \quad P(B|A_2) = 0, \quad P(B|A_3) = 1$
- Posterior: $P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}, \quad i = 1, 2, 3$

$$P(A_1|B) = \frac{1}{3}, \quad P(A_2|B) = 0, \quad P(A_3|B) = \frac{2}{3}$$

Learning parameters given in a model

- Assume we are working with a single model (e.g. a regression model with a particular set of explanatory variables) which depends on parameters θ.
- So we want to figure out properties of the posterior $P(\theta|y)$
- It is convenient to use Bayes' theorem to write the posterior in a different way by replacing A with θ and B with y.

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

- Bayesians treat $P(\theta|y)$ as being of fundamental interest. That is, it directly addresses the question "Given the data, what do we know about θ ?".
- The treatment of θ as a random variable is controversial among frequentist who says that θ is not a random variable.
- For estimation we can ignore the term $P(y)$ since it does not involve θ:

$$P(\theta|y) \propto P(y|\theta)P(\theta)$$

Multi-Armed Bandit

- What is Multi-Armed Bandit?
- One-Armed Bandit = Slot Machine

- How to know the winning rate of one-armed bandit if our budget is limited?
- Under frequentist approach, assume $X_1, ..., X_T \sim Ber(\theta)$
- Assume $P(\theta) \sim Beta(a_0, b_0), \quad a_0, b_0 > 0, \quad 0 \leq \theta \leq 1$

- Posterior:

$$\begin{aligned} P(\theta|X_1,...,X_T) &\propto P(X_1,...,X_T|\theta)P(\theta) \\ &\propto \theta^{\sum X_t}(1-\theta)^{T-\sum X_t}\theta^{a_0-1}(1-\theta)^{b_0-1} \\ &\propto \theta^{a_0+\sum X_t-1}(1-\theta)^{b_0-\sum X_t+T-1} \\ &\propto \theta^{a_1-1}(1-\theta)^{b_1-1} \end{aligned}$$

where $a_1 = a_0 + \sum X_t, \quad b_1 = b_0 - \sum X_t + T$

- We know this is also the kernel of beta distribution. Therefore we can write: $P(\theta|X_1,...,X_T) \sim Beta(a_1, b_1)$ (conjugate prior)

Interpretation of Bayesian Formulae for the Beta-Binomial model

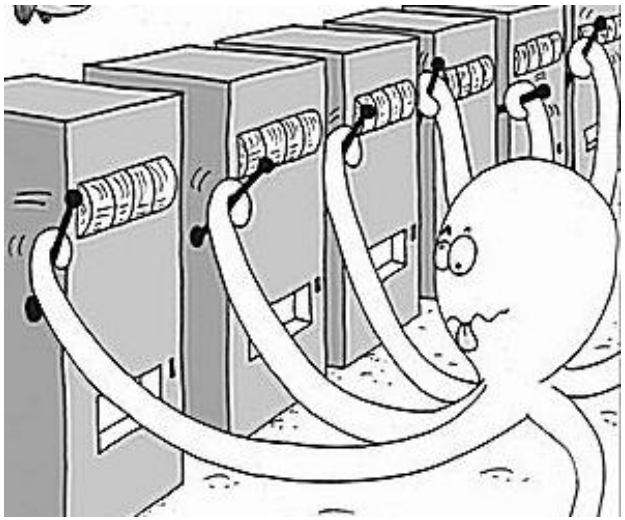
- Recall mean of beta distribution:

$$E(\theta|X_1,...,X_T) = \frac{a_1}{a_1+b_1} = \frac{a_0+\sum X_t}{a_0+b_0+T} = \frac{a_0+b_0}{a_0+b_0+T}\frac{a_0}{a_0+b_0} + \frac{T}{a_0+b_0+T}\frac{\sum X_t}{T}$$

- It's obvious that if we win this time, the more confidence we will have next time.

Multi-Armed Bandit

In reality, we may encounter more than one slot machine in casino, then how to update our prior belief among them?



Problem setting

- # of arms K, # of rounds T
- For each round t=1,...,T:
  - The player chooses an arm and receives the reward $r_{k,t}$.
  - The reward, $r_{k,t}$, follows the probability distribution P_k with the mean μ_k.
- We call this partially observable property as the bandit setting.
- Hence, the player should find the arm with the highest μ_k in the limited rounds.

Exploration-Exploitation Dilemma

- Exploration vs Exploitation:
  - exploration gather more information
  - exploitation make the best decision with given information
- Two naive strategies:
  - Random (full exploration): choose arm randomly
  - Greedy (full exploitation): choose the empirical best arm
- Question: How to balance between exploration and exploitation?

Bayesian Linear Regression

Recall

$$y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I_T)$$

- In the standard approach we write down the likelihood function $P(y|\beta, \sigma^2) = (2\pi)^{\frac{-T}{2}} |\sigma^2 I_T|^{\frac{-1}{2}} exp\left[\frac{-1}{2}(y - X\beta)'(\sigma^2 I_T)^{-1}(y - X\beta)\right]$
- Then we obtain data and maximize $P(y|\beta, \sigma^2)$, which gives the standard estimators

$$\hat{\beta} = (X'X)^{-1}X'y,$$

- Question: How to set prior $\theta = \{\beta\}$ and derive its posterior distributions? (For simplicity, here we assume σ^2 is known)

Conjugate prior distribution

- For an arbitrary prior distribution, there may be no analytical solution for the posterior distribution. Therefore, we will consider a so-called conjugate prior for which the posterior distribution can be derived analytically.
- Prior β is conjugate to the likelihood function if it has the same function form with respect to the β. Since the likelihood of β is exponential quadratic which is kernel of normal distribution, this suggests normal distribution for the prior

- Starts with a prior belief about the coefficient β. The prior belief is in the form of a distribution:

$$P(\beta) \sim N(\beta_0, \Lambda_0^{-1})$$

- Collect data and write down the likelihood function as before $P(y|\beta, X)$.
- Update prior belief on the basis of the information in the data. Combine the prior distribution $P(\beta)$ and the likelihood function $P(y|\beta)$ to obtain the posterior distribution $P(\beta|y)$

- Posterior:

$$
\begin{aligned}
P(\beta|y,X) &\propto P(y|X,\beta)P(\beta) \\
&\propto exp(\frac{-1}{2\sigma^2}(y-X\beta)'(y-X\beta))exp(\frac{-1}{2}(\beta-\beta_0)'\Lambda_0(\beta-\beta_0)) \\
&\propto exp(\frac{-1}{2}\left[\frac{1}{\sigma^2}\beta'X'X\beta+\beta'\Lambda_0\beta-2(\frac{1}{\sigma^2}\beta'X'y+\beta'\Lambda_0\beta_0)\right]) \\
&\propto exp(\frac{-1}{2}(\beta-\beta_1)'\Lambda_1(\beta-\beta_1))
\end{aligned}
$$

where $\Lambda_1 = (\frac{1}{\sigma^2}X'X+\Lambda_0)$

$\beta_1 = \Lambda_1^{-1}(\frac{1}{\sigma^2}X'y+\Lambda_0\beta_0)$

- This is the kernel of a normal distribution. Therefore we can write: $P(\beta|y,X) \sim N(\beta_1, \Lambda_1^{-1})$

Interpretation of Bayesian Formulae for the Linear Regression

$$\beta_1 = (\frac{1}{\sigma^2}X'X + \Lambda_0)^{-1}(\frac{1}{\sigma^2}X'y + \Lambda_0\beta_0)$$

- Posterior mean, β_1, is a weighted average of OLS estimator and prior mean, β_0
- The weights are inversely proportional to the precision of prior and data information.
- If we set $\Lambda_0 = \frac{1}{\sigma^2}\lambda I$ and $\beta_0 = 0$, then β_1 will become ridge regression: $\beta_1 = (X'X + \lambda I)^{-1}X'y$

$$\beta_1 = (\frac{1}{\sigma^2}X'X + \Lambda_0)^{-1}(\frac{1}{\sigma^2}X'y + \Lambda_0\beta_0)$$

- If we set variance of prior, Λ_0^{-1}, large, then great prior uncertainty, lack of informative prior, noninformative prior.
- Hence, Λ_0 represents the prior of believe.
- However, if we have larger and larger data, the influence of prior will become insignificant even thought we have strong belief in prior. Eventually, $\beta_1 \approx (X'X)^{-1}X'y$.

Bayesian Computation

- In practice, we may encounter models which have more than one parameter needed to estimate.
- Bayesian propose a method called MCMC (Markov-Chain Monte Carlo) to address this computation problem.
- Gibbs sampler (one particular MCMC estimation method) can be implemented if marginal (conditional) posterior distributions are closed form.

Gibbs Sampler

- For simplicity, here we assume linear regression which has only two parameters β_0 (intercept) and β_1 (slope)

- Followed from previous discussion, we already know full posterior is a bivariate normal distribution, then their marginal posterior distributions (given the other) are:
  $\beta_0|\beta_1, x, y \sim N(\Lambda_{10}^{-1}\left(\beta_{00}\Lambda_{00} + \frac{1}{\sigma^2}\sum(y_t - \beta_1 x_t)\right), \Lambda_{10}^{-1})$
  $\beta_1|\beta_0, x, y \sim N(\Lambda_{11}^{-1}\left(\beta_{01}\Lambda_{01} + \frac{1}{\sigma^2}\sum(y_t - \beta_0)x_t\right), \Lambda_{11}^{-1})$
  where $\Lambda_{10} = \Lambda_{00} + \frac{T}{\sigma^2}, \quad \Lambda_{11} = \Lambda_{01} + \frac{\sum x_i^2}{\sigma^2}$

- Gibbs sampling works as follow: suppose we have two parameters β_0 and β_1. Our goal is to find the posterior distribution of $P(\beta_0, \beta_1 | X, y)$.

- To do this in a Gibbs sampling regime we need to work out the marginal distributions. The Gibbs updates are then:
  - Pick initial $\beta_0^{(i)}$ and $\beta_1^{(i)}$
  - Sample $\beta_0^{(i+1)} \sim P(\beta_0 | \beta_1^{(i)}, X, y)$
  - Sample $\beta_1^{(i+1)} \sim P(\beta_1 | \beta_0^{(i+1)}, X, y)$
  - ...
  - Until they converge to stable posterior distributions.

Bayesian Variable Selection

- Bayesian variable selection methods come equipped with natural measures of uncertainty, such as the posterior probability of each model and the marginal inclusion probabilities of the individual explanatory variables.
- Stochastic variable search is extremely flexible and has been applied in a wide range of applications.
- Given the model (likelihood, prior, and posterior), there are formal justifications for choosing a particular model.

Stochastic Search Variable Selection

- Maybe the most naive method is to compute all $2^K - 1$ model and compare their $\overline{R}^2$.
- Stochastic variable selection is an alternative.
- Here we use MCMC to draw samples from the model space so models with high posterior probability will be visited more often than low probability models.

- Recall what we have learned so far?
- Bayes Theorem, Multi-Armed Bandit, Bayesian linear regression...
- Also learned if we restrict variance of the prior distribution, it will narrow search space while implementing Gibbs sampler.
- How to combine them to do variable selection?

Recall

$$y = X\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I_T)$$

- SSVS is a hierarchical prior = prior expressed in terms of parameters which in turn have a prior of their own
- $P(\beta_k|\gamma_k) \sim \gamma_k N(0, c_k^2\tau_k^2) + (1-\gamma_k)N(0, \tau_k^2)$
  $P(\gamma_k|w_k) \sim Ber(w_k)$
  $P(w_k) \sim Beta(c_o, d_0)$
- The constant τ_i is small, so that if $\gamma_k = 0$, β_k could be safely estimated by 0.
- The constant c_k is small, so that if $\gamma_k = 0$, β_k could be safely estimated by 0.

- Full Posterior:

$$
\begin{aligned}
P(\beta, \gamma, w|y, X) &\propto P(y|X, \beta) \prod P(\beta|\gamma_k) P(\gamma_k|w_k) P(w_k) \\
&\propto N(y|X\beta, \sigma^2 I) \prod \left[\gamma_k N(\beta_k; 0, c_k^2 \tau_k^2) + (1 - \gamma_k) N(\beta_k; 0, \tau_k^2)\right] \\
&\quad \times w_k^{\gamma_k} (1 - w_k)^{1-\gamma_k} w_k^{c_0 - 1} (1 - w_k)^{d_0 - 1}
\end{aligned}
$$

Gibbs sampler

- Marginal Posterior:
  - $P(w_k|c_0, d_0, \gamma_k) \sim Beta(c_0 + \gamma_k, 1 - \gamma_k + d_0)$
  - $P(\gamma_k|\beta, \gamma_{-k}, w_k) \sim Ber(\frac{a}{a+b})$
    where $a = P(\beta_k|\gamma_k = 1)w_k, \quad b = P(\beta_k|\gamma_k = 0)(1 - w_k)$
  - $P(\beta|\gamma, y, X) \sim N_K\left((X'X + \sigma^2 D_\gamma^{-2})^{-1}X'y, \sigma^2(X'X + \sigma^2 D_\gamma^{-2})^{-1}\right)$
    where D_γ is a diagonal matrix with k^{th} diagonal element equal to $\gamma_k c_k \tau_k + (1 - \gamma_k)\tau_k$

- The Gibbs updates are then:
  - Pick initial $\beta^{(i)}$, $\gamma^{(i)}$ and $w^{(i)}$
  - Sample $\beta^{(i+1)} \sim P(\beta|\gamma^{(i)}, y, X)$
  - Sample $\gamma_k^{(i+1)} \sim P(\gamma_k|\beta^{(i+1)}, \gamma_{-k}, w_k^{(i)}), \quad k = 1, .., K$
  - Sample $w_k^{(i+1)} \sim P(w_k|c_0, d_0, \gamma_k^{(i+1)}), \quad k = 1, ..., K$
  - ...
  - Until they converge to stable posterior distributions.

- SSVS approach provides a more efficiency way to search model space and parameter space.
- The Bayesian approach to variable selection has several attractive features, including straight-forward quantification of variable importance and model uncertainty and the flexibility to handle missing data and non-Gaussian distributions.
- Outstanding issues include computational efficiency and concerns about priors.

Conclusion

Pros:

- Incorporate prior knowledge
- Do inference even in small data
- Provide more information (posterior)
- Estimate much more difficult models

Cons:

- It's hard(er)
- Computationally intensive
- No guarantee of MCMC convergence

The End