MATH 829: Introduction to Data Mining and Analysis
A (very brief) introduction to Bayesian inference

Dominique Guillot

Departments of Mathematical Sciences
University of Delaware

May 13, 2016

Frequentist statistics:

- Compute *point* estimates (e.g. maximum likelihood).
- Define probabilities as the long-run frequency of events .

Bayesian vs. Frequentist

Frequentist statistics:

- Compute *point* estimates (e.g. maximum likelihood).
- Define probabilities as the long-run frequency of events .

Bayesian statistics:

- Probabilities are a “state of knowledge” or a “state of belief”.
- Parameters have a probability distribution.
- Prior knowledge is updated in the light of new data.

Example

You flip a coin 14 times. You get `head` 10 times. What is $p := P(head)$?

Example

You flip a coin 14 times. You get `head` 10 times. What is $p := P(head)$?

- Frequentist approach: estimate p using, say maximum likelihood:

$$p \approx \frac{10}{14} \approx 0.714.$$

Example

You flip a coin 14 times. You get `head` 10 times. What is $p := P(head)$?

- Frequentist approach: estimate p using, say maximum likelihood:

$$p \approx \frac{10}{14} \approx 0.714.$$

- Bayesian approach: we treat p as a *random variable*.

Example

You flip a coin 14 times. You get `head` 10 times. What is $p := P(head)$?

- Frequentist approach: estimate p using, say maximum likelihood:

$$p \approx \frac{10}{14} \approx 0.714.$$

- Bayesian approach: we treat p as a *random variable*.
  1. Choose a *prior* distribution for p, say $P(p)$.

Example

You flip a coin 14 times. You get `head` 10 times. What is $p := P(head)$?

- Frequentist approach: estimate p using, say maximum likelihood:

$$p \approx \frac{10}{14} \approx 0.714.$$

- Bayesian approach: we treat p as a *random variable*.
  1. Choose a *prior* distribution for p, say $P(p)$.
  2. Update the prior distribution using the data via *Bayes' theorem*:

$$P(p|data) = \frac{P(data|p)P(p)}{P(data)} \propto P(data|p)P(p).$$

Example (cont.)

Note: "$data|p$" $\sim \mathrm{Binomial}(14, p)$.

Example (cont.)

Note: "$data|p$" $\sim \text{Binomial}(14, p)$. Therefore:

$$P(data|p) = \binom{14}{10} p^{10}(1-p)^4.$$

Example (cont.)

Note: "$data|p$" $\sim \text{Binomial}(14, p)$. Therefore:

$$P(data|p) = \binom{14}{10} p^{10}(1-p)^4.$$

What should we choose for $P(p)$?

Example (cont.)

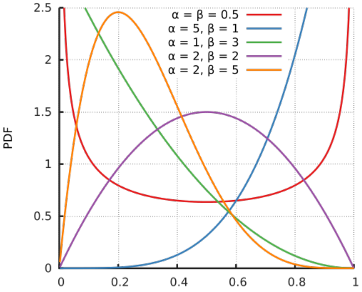
Note: "$data|p$" $\sim \text{Binomial}(14, p)$. Therefore:

$$P(data|p) = \binom{14}{10} p^{10}(1-p)^4.$$

What should we choose for $P(p)$?

The beta distribution $\text{Beta}(\alpha, \beta)$:

$$P(p; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \qquad (p \in (0,1)).$$



Source: Wikipedia.

Example (cont.)

- Suppose we decide to pick $p \sim \mathrm{Beta}(\alpha, \beta)$.

Example (cont.)

- Suppose we decide to pick $p \sim \mathrm{Beta}(\alpha, \beta)$. Then:

$$\begin{aligned}
P(p|data) &\propto P(data|p)P(p) \\
&= \binom{14}{10} p^{10}(1-p)^4 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \\
&\propto p^{10}(1-p)^4 p^{\alpha-1}(1-p)^{\beta-1} \\
&= p^{10+\alpha-1}(1-p)^{4+\beta-1}.
\end{aligned}$$

Example (cont.)

- Suppose we decide to pick $p \sim \mathrm{Beta}(\alpha, \beta)$. Then:

$$\begin{aligned} P(p|data) &\propto P(data|p)P(p) \\ &= \binom{14}{10} p^{10}(1-p)^4 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \\ &\propto p^{10}(1-p)^4 p^{\alpha-1}(1-p)^{\beta-1} \\ &= p^{10+\alpha-1}(1-p)^{4+\beta-1}. \end{aligned}$$

Remark: We don't need to worry about the *normalization constant* since it is uniquely determined by the fact that $P(p|data)$ is a probability distribution.

Example (cont.)

- Suppose we decide to pick $p \sim \mathrm{Beta}(\alpha, \beta)$. Then:

$$
\begin{aligned}
P(p|data) &\propto P(data|p)P(p) \\
&= \binom{14}{10} p^{10}(1-p)^4 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \\
&\propto p^{10}(1-p)^4 p^{\alpha-1}(1-p)^{\beta-1} \\
&= p^{10+\alpha-1}(1-p)^{4+\beta-1}.
\end{aligned}
$$

Remark: We don't need to worry about the *normalization constant* since it is uniquely determined by the fact that $P(p|data)$ is a probability distribution.

- Conclusion: $P(p|data) \sim \mathrm{Beta}(10+\alpha, 4+\beta)$.

Example (cont.)

- How should we choose α, β?

Example (cont.)

- How should we choose α, β?

According to our *prior knowledge* of p.

Example (cont.)

- How should we choose α, β?

According to our *prior knowledge* of p.

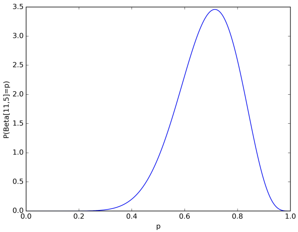
- Suppose we have no prior knowledge: use a *flat* prior: $\alpha = \beta = 1$ (Uniform distribution).

Example (cont.)

- How should we choose α, β?

According to our *prior knowledge* of p.

- Suppose we have no prior knowledge: use a *flat* prior: $\alpha = \beta = 1$ (Uniform distribution).
- The resulting *posterior distribution* is $p|data \sim \mathrm{Beta}(11, 5)$:

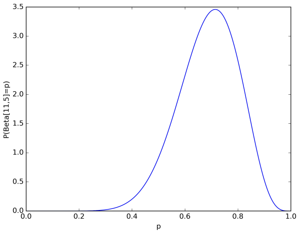


Example (cont.)

- How should we choose α, β?

According to our *prior knowledge* of p.

- Suppose we have no prior knowledge: use a *flat* prior: $\alpha = \beta = 1$ (Uniform distribution).
- The resulting *posterior distribution* is $p|data \sim \mathrm{Beta}(11, 5)$:



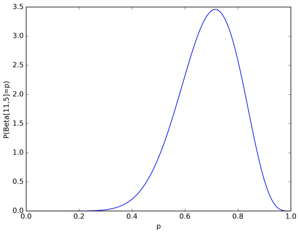
Our "knowledge" of p has now been updated using the observed data (or evidence).

Example (cont.)

- How should we choose α, β?

According to our *prior knowledge* of p.

- Suppose we have no prior knowledge: use a *flat* prior: $\alpha = \beta = 1$ (Uniform distribution).
- The resulting *posterior distribution* is $p|data \sim \mathrm{Beta}(11, 5)$:



Our “knowledge” of p has now been updated using the observed data (or evidence).
Impportant advantage: Our estimate of p comes with its own **uncertainty**.

Bayesian analysis

More generally: suppose we have a model for X that depends on some parameters θ. Then:

Bayesian analysis

More generally: suppose we have a model for X that depends on some parameters θ. Then:

1. Choose a prior $P(\theta)$ for θ.

Bayesian analysis

More generally: suppose we have a model for X that depends on some parameters θ. Then:

1. Choose a prior $P(\theta)$ for θ.
2. Compute the posterior distribution of θ using

$$p(\theta|X) \propto P(X|\theta) \cdot P(\theta).$$

Bayesian analysis

More generally: suppose we have a model for X that depends on some parameters θ. Then:

1. Choose a prior $P(\theta)$ for θ.
2. Compute the posterior distribution of θ using

$$p(\theta|X) \propto P(X|\theta) \cdot P(\theta).$$

Note: $\text{Posterior} = \text{Prior} \times \text{Likelihood}.$

Bayesian analysis

More generally: suppose we have a model for X that depends on some parameters θ. Then:

1. Choose a prior $P(\theta)$ for θ.
2. Compute the posterior distribution of θ using

$$p(\theta|X) \propto P(X|\theta) \cdot P(\theta).$$

Note: $\text{Posterior} = \text{Prior} \times \text{Likelihood}.$

Advantages:

- Mimics the scientific method: formulate hypothesis, run experiment, update knowledge.
- Can incorporate prior information (e.g. the range of variables).
- Automatically provides uncertainty estimates.

Bayesian analysis

More generally: suppose we have a model for X that depends on some parameters θ. Then:

1. Choose a prior $P(\theta)$ for θ.
2. Compute the posterior distribution of θ using
$$p(\theta|X) \propto P(X|\theta) \cdot P(\theta).$$

Note: $\text{Posterior} = \text{Prior} \times \text{Likelihood}.$

Advantages:

- Mimics the scientific method: formulate hypothesis, run experiment, update knowledge.
- Can incorporate prior information (e.g. the range of variables).
- Automatically provides uncertainty estimates.

Drawbacks:

- Not always obvious how to choose priors.
- Can be difficult to compute the posterior distribution.
- Can be computationally intensive to sample from the posterior distribution (when not available in closed form).

Conjugate priors

- In the previous example, the posterior distribution was from the same family as the prior.

Conjugate priors

- In the previous example, the posterior distribution was from the same family as the prior.
- A prior with this property is said to be a *conjugating prior*.

Conjugate priors

- In the previous example, the posterior distribution was from the same family as the prior.
- A prior with this property is said to be a *conjugating prior*.
- Conjugating priors are known for many common likelihood functions.

Likelihood	Conjugate prior
Binomial	Beta
Multinomial	Dirichlet
Poisson	Gamma
Normal	
μ unknown, σ^2 known	Normal
μ known, σ^2 unknown	Inverse Chi-Square
Multivariate Normal	
μ unknown, $\mathbf{V}$ known	Multivariate Normal
μ known, $\mathbf{V}$ unknown	Inverse Wishart

MCMC methods

- Markov chain Monte Carlo (MCMC) methods are popular ways of sampling from complicated distributions (e.g. the posterior distribution of a complicated model).

MCMC methods

- Markov chain Monte Carlo (MCMC) methods are popular ways of sampling from complicated distributions (e.g. the posterior distribution of a complicated model).
- Idea:

1. Construct a **Markov chain** with the desired distribution as its **stationary distribution** π.
2. Burn (e.g. forget) a given number of samples from the Markov chain (while the chain converges to its stationary distribution).
3. Generate a sample from the desired distribution (approximately).

MCMC methods

- Markov chain Monte Carlo (MCMC) methods are popular ways of sampling from complicated distributions (e.g. the posterior distribution of a complicated model).
- Idea:
  1. Construct a **Markov chain** with the desired distribution as its **stationary distribution** π.
  2. Burn (e.g. forget) a given number of samples from the Markov chain (while the chain converges to its stationary distribution).
  3. Generate a sample from the desired distribution (approximately).
- One generally then compute some *statistics* of the sample (e.g. mean, variance, mode, etc.).

Rejection sampling

A simple way to sample from a distribution:

Rejection sampling

A simple way to sample from a distribution:

- We want to sample from a distribution $f(x)$ (complicated).
- We know how to sample from another distribution $g(x)$ (simpler).
- We know that $f(x) \leq c \cdot g(x)$ for some (known) constant $c > 0$.

Rejection sampling

A simple way to sample from a distribution:

- We want to sample from a distribution $f(x)$ (complicated).
- We know how to sample from another distribution $g(x)$ (simpler).
- We know that $f(x) \leq c \cdot g(x)$ for some (known) constant $c > 0$.

Then

1. Draw $z \sim h(x)$ and $u \sim \mathrm{Uniform}[0, 1]$.
2. If $u < f(z)/(c \cdot g(z))$ accept the draw. Otherwise, discard z and repeat.

Rejection sampling

A simple way to sample from a distribution:

- We want to sample from a distribution $f(x)$ (complicated).
- We know how to sample from another distribution $g(x)$ (simpler).
- We know that $f(x) \leq c \cdot g(x)$ for some (known) constant $c > 0$.

Then

1. Draw $z \sim h(x)$ and $u \sim \mathrm{Uniform}[0, 1]$.
2. If $u < f(z)/(c \cdot g(z))$ accept the draw. Otherwise, discard z and repeat.

Works well in some cases, but the rejection rate is often large and the resulting algorithm can be very inefficient.

Metropolis–Hastings algorithm

- Nicolas Metropolis (1915–1999) was an American physicist. He worked on the first nuclear reactors at the Los Alamos National Laboratory during the second world war. Introduced the algorithm in 1953 in the paper

  Equation of State Calculations by Fast Computing Machines

  with A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller
- W. K. Hastings (Born 1930) is a Canadian statistician who extended the algorithm to the more general case in 1970.

Metropolis–Hastings algorithm (cont.)

- Suppose we want to sample from a distribution $P(x) = f(x)/K$, where $K > 0$ is some constant.

Metropolis–Hastings algorithm (cont.)

- Suppose we want to sample from a distribution $P(x) = f(x)/K$, where $K > 0$ is some constant.
Note: The normalization constant K is often unknown and difficult to compute.

Metropolis–Hastings algorithm (cont.)

- Suppose we want to sample from a distribution $P(x) = f(x)/K$, where $K > 0$ is some constant.
Note: The normalization constant K is often unknown and difficult to compute.
- The Metropolis–Hastings starts with an initial sample, and generate new samples using a *transition probability density* $q(x, y)$ (the *proposal distribution*).

- Suppose we want to sample from a distribution $P(x) = f(x)/K$, where $K > 0$ is some constant.
Note: The normalization constant K is often unknown and difficult to compute.
- The Metropolis–Hastings starts with an initial sample, and generate new samples using a *transition probability density* $q(x, y)$ (the *proposal distribution*).
- We assume
  - we can evaluate $f(x)$ at every x.
  - we can evaluate $q(x, y)$ at every x, y.
  - we can sample from the distribution $q(x, \cdot)$.

Metropolis–Hastings algorithm (cont.)

The Metropolis–Hastings algorithm: we start with x_0 such that $f(x_0) > 0$. For $i = 0, \ldots$

Metropolis–Hastings algorithm (cont.)

The Metropolis–Hastings algorithm: we start with x_0 such that $f(x_0) > 0$. For $i = 0, \ldots$

1. Generate a new value y according to $q(x, \cdot)$.

Metropolis–Hastings algorithm (cont.)

The Metropolis–Hastings algorithm: we start with x_0 such that $f(x_0) > 0$. For $i = 0, \ldots$

1. Generate a new value y according to $q(x, \cdot)$.
2. Compute the "Hastings" ratio:

$$R = \frac{f(y)q(y,x)}{f(x)q(x,y)}$$

Metropolis–Hastings algorithm (cont.)

The Metropolis–Hastings algorithm: we start with x_0 such that $f(x_0) > 0$. For $i = 0, \ldots$

1. Generate a new value y according to $q(x, \cdot)$.
2. Compute the "Hastings" ratio:

$$R = \frac{f(y)q(y,x)}{f(x)q(x,y)}$$

3. "Accept" the new sample y with probability $\min(1, R)$. If y is accepted, set $x_{i+1} := y$. Otherwise, $x_{i+1} = x_i$.

Metropolis–Hastings algorithm (cont.)

The Metropolis–Hastings algorithm: we start with x_0 such that $f(x_0) > 0$. For $i = 0, \ldots$

1. Generate a new value y according to $q(x, \cdot)$.
2. Compute the "Hastings" ratio:

$$R = \frac{f(y)q(y,x)}{f(x)q(x,y)}$$

3. "Accept" the new sample y with probability $\min(1, R)$. If y is accepted, set $x_{i+1} := y$. Otherwise, $x_{i+1} = x_i$.

Some difficulties:

- Choosing an efficient proposal distribution $q(x, y)$.
- How long should we wait for the Markov chain to converge to the desired distribution, i.e., how many samples should we burn?
- How long should we sample after convergence to make sure we sample in low probability regions?

Gibbs sampling

- Idea: use the conditional distribution of X to generate new samples.

Gibbs sampling

- Idea: use the conditional distribution of X to generate new samples.
- Note: only possible when the conditional distributions are "nice".

Gibbs sampling

- Idea: use the conditional distribution of X to generate new samples.
- Note: only possible when the conditional distributions are "nice".

Suppose $X = (X_1, \ldots, X_p)$ and $X^{(i)} = (x_1^{(i)}, \ldots, x_p^{(i)})$ is a given sample. Generate a new sample $X^{(i+1)} = (x_1^{(i+1)}, \ldots, x_p^{(i+1)})$ as follows:

Gibbs sampling

- Idea: use the conditional distribution of X to generate new samples.
- Note: only possible when the conditional distributions are "nice".

Suppose $X = (X_1, \ldots, X_p)$ and $X^{(i)} = (x_1^{(i)}, \ldots, x_p^{(i)})$ is a given sample. Generate a new sample $X^{(i+1)} = (x_1^{(i+1)}, \ldots, x_p^{(i+1)})$ as follows:

1. Generate $x_1^{(i+1)}$ according to the marginal

$$p(x_1 | x_2^{(i)}, \ldots, x_p^{(i)}).$$

Gibbs sampling

- Idea: use the conditional distribution of X to generate new samples.
- Note: only possible when the conditional distributions are "nice".

Suppose $X = (X_1, \ldots, X_p)$ and $X^{(i)} = (x_1^{(i)}, \ldots, x_p^{(i)})$ is a given sample. Generate a new sample $X^{(i+1)} = (x_1^{(i+1)}, \ldots, x_p^{(i+1)})$ as follows:

1. Generate $x_1^{(i+1)}$ according to the marginal
$$p(x_1 | x_2^{(i)}, \ldots, x_p^{(i)}).$$
2. Generate $x_2^{(i+1)}$ according to
$$p(x_2 | x_1^{(i+1)}, x_3^{(i)}, \ldots, x_p^{(i)}).$$

Gibbs sampling

- Idea: use the conditional distribution of X to generate new samples.
- Note: only possible when the conditional distributions are "nice".

Suppose $X = (X_1, \dots, X_p)$ and $X^{(i)} = (x_1^{(i)}, \dots, x_p^{(i)})$ is a given sample. Generate a new sample $X^{(i+1)} = (x_1^{(i+1)}, \dots, x_p^{(i+1)})$ as follows:

1. Generate $x_1^{(i+1)}$ according to the marginal
$$p(x_1|x_2^{(i)}, \dots, x_p^{(i)}).$$
2. Generate $x_2^{(i+1)}$ according to
$$p(x_2|x_1^{(i+1)}, x_3^{(i)}, \dots, x_p^{(i)}).$$
3. Generate $x_3^{(i+1)}$ accodring to
$$p(x_3|x_1^{(i+1)}, x_2^{(i+1)}, x_4^{(i)} \dots, x_p^{(i)}).$$
4. etc..