an introduction to bayesian inference
with an application to network analysis

jake hofman
http://jakehofman.com

january 13, 2010
motivation

- would like models that:
  - provide predictive and explanatory power
  - are complex enough to describe observed phenomena
  - are simple enough to generalize to future observations
motivation

- would like models that:
  - provide predictive and explanatory power
  - are complex enough to describe observed phenomena
  - are simple enough to generalize to future observations

claim: bayesian inference provides a systematic framework to infer such models from observed data
motivation

- principles behind Bayesian interpretation of probability and Bayesian inference are well established (Bayes, Laplace, etc., 18th century)

- recent advances in mathematical techniques and computational resources have enabled successful applications of these principles to real-world problems
motivation: a bayesian approach to network modularity
1. **principles (what we’d like to do)**
   - background: joint, marginal, and conditional probabilities
   - bayes’ theorem: inverting conditional probabilities
   - bayesian probability: unknowns as random variables
   - bayesian inference: bayesian probability + bayes’ theorem

2. **practice (what we’re able to do)**
   - monte carlo methods: representative samples
   - variational methods: bound optimization
   - references

3. **application: bayesian inference for network data**
joint, marginal, and conditional probabilities

**Joint Distribution**

$p_{XY}(X = x, Y = y)$: probability $X = x$ and $Y = y$

**Conditional Distribution**

$p_{X|Y}(X = x | Y = y)$: probability $X = x$ given $Y = y$

**Marginal Distribution**

$p_X(X)$: probability $X = x$ (regardless of $Y$)
### Sum Rule

Sum out settings of irrelevant variables:

\[
p(x) = \sum_{y \in \Omega_Y} p(x, y)
\]  
(1)

### Product Rule

The joint as the product of the conditional and marginal:

\[
p(x, y) = p(x|y) p(y)
\]  
(2)

\[
p(x, y) = p(y|x) p(x)
\]  
(3)
1. **principles (what we’d like to do)**
   - background: joint, marginal, and conditional probabilities
   - bayes’ theorem: inverting conditional probabilities
   - bayesian probability: unknowns as random variables
   - bayesian inference: bayesian probability + bayes’ theorem

2. **practice (what we’re able to do)**
   - monte carlo methods: representative samples
   - variational methods: bound optimization
   - references

3. **application: bayesian inference for network data**
inverting conditional probabilities

equate far right- and left-hand sides of product rule

\[ p(y|x) \cdot p(x) = p(x,y) = p(x|y) \cdot p(y) \] (4)

and divide:

**bayes’ theorem (bayes and price 1763)**

the probability of \( Y \) given \( X \) from the probability of \( X \) given \( Y \):

\[ p(y|x) = \frac{p(x|y) \cdot p(y)}{p(x)} \] (5)

where \( p(x) = \sum_{y \in \Omega} p(x|y) \cdot p(y) \) is the normalization constant
example: diagnoses a la bayes

- population 10,000
- 1% has (rare) disease\(^1\)
- test is 99% (relatively) effective, i.e.
  - given a patient is sick, 99% test positive
  - given a patient is healthy, 99% test negative

\(^1\)subtlety: assuming this fraction is *known*
example: diagnoses a la bayes

- population 10,000
- 1% has (rare) disease
- test is 99% (relatively) effective, i.e.
  - given a patient is sick, 99% test positive
  - given a patient is healthy, 99% test negative
- given positive test, what is probability the patient is sick?\(^1\)

\(^1\)follows wiggins (2006)
example: diagnoses a la bayes

- 99 sick patients test positive, 99 healthy patients test positive
example: diagnoses a la bayes

- 99 sick patients test positive, 99 healthy patients test positive
- given positive test, 50% probability that patient is sick
example: diagnoses a la bayes

- know probability of testing positive/negative given sick/healthy
- use bayes’ theorem to “invert” to probability of sick/healthy given positive/negative test

\[
p(sick|test+) = \frac{\left(\frac{99}{100}\right) \cdot \frac{1}{100}}{\left(\frac{99}{100}\right) + \left(\frac{99}{100}\right)} = \frac{99}{198} = \frac{1}{2}
\]
example: diagnoses a la Bayes

- know probability of testing positive/negative given sick/healthy
- use Bayes’ theorem to “invert” to probability of sick/healthy given positive/negative test

\[
p(sick | test^+) = \frac{p(test^+ | sick) p(sick)}{p(test^+)} = \frac{\frac{99}{100} \times \frac{1}{100}}{\frac{99}{100} + \frac{99}{100}} = \frac{99}{198} = \frac{1}{2} \tag{6}
\]

- most “work” in calculating denominator (normalization)
**1 principles (what we’d like to do)**
- background: joint, marginal, and conditional probabilities
- bayes’ theorem: inverting conditional probabilities
- bayesian probability: unknowns as random variables
- bayesian inference: bayesian probability + bayes’ theorem

**2 practice (what we’re able to do)**
- monte carlo methods: representative samples
- variational methods: bound optimization
- references

**3 application: bayesian inference for network data**
interpretations of probabilities
(just enough philosophy)

- frequentists: limit of relative frequency of events for large number of trials
- bayesians: measure of a state of knowledge, quantifying degrees of belief (jaynes 2003)
interpretations of probabilities
(just enough philosophy)

- frequentists: limit of relative frequency of events for large number of trials
- bayesians: measure of a state of knowledge, quantifying degrees of belief (Jaynes 2003)
- key difference: bayesians permit assignment of probabilities to unknown/unobservable hypotheses (frequentists do not)
e.g., inferring model parameters $\Theta$ from observed data $\mathcal{D}$:

- frequentist approach: calculate parameter setting that maximizes likelihood of observed data (point estimate),

$$\hat{\Theta} = \arg \max_{\Theta} p(\mathcal{D}|\Theta)$$  \hspace{1cm} (7)

- bayesian approach: calculate distribution over parameter settings given data,

$$p(\Theta|\mathcal{D}) = ?$$  \hspace{1cm} (8)
interpretations of probabilities
(just enough philosophy)

- using Bayes' rule ≠ “being Bayesian”
interpretations of probabilities
(just enough philosophy)

- using bayes' rule ≠ “being bayesian”
- s/bayesian/subjective probabilist/g
**Outline**

1. **Principles (what we’d like to do)**
   - Background: joint, marginal, and conditional probabilities
   - Bayes’ theorem: inverting conditional probabilities
   - Bayesian probability: unknowns as random variables
   - Bayesian inference: Bayesian probability + Bayes’ theorem

2. **Practice (what we’re able to do)**
   - Monte Carlo methods: representative samples
   - Variational methods: bound optimization
   - References

3. **Application: Bayesian inference for network data**
Bayesian inference:

- Treat unknown quantities as random variables
- Use Bayes’ theorem to systematically update prior knowledge in the presence of observed data

\[
p(\Theta|D) = \frac{p(D|\Theta)p(\Theta)}{p(D)}
\]  \hspace{1cm} (9)
example: coin flipping

- observe independent coin flips (bernoulli trials)
- infer distribution over coin bias
example: coin flipping

prior $p(\Theta)$ over coin bias before observing flips
example: coin flipping

observe flips: HTHHHTTHHHH
example: coin flipping

update posterior $p(\Theta|D)$ using Bayes’ theorem
example: coin flipping

observe flips: HHHHHHHHHHHHHTHHHHHHHHHHHHH
HHHHHHHHHHHHHHHHHHHHHHHHHHHH
HHHHHHHTHHHHHHHHHHHHHHHHHHHH
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH HH
update posterior $p(\Theta|D)$ using bayes’ theorem

![Graph showing posterior over coin fairness.

\[ \alpha = 100, \beta = 5 \]
“naive” “bayes” for document classification

- model presence/absence of each word as independent coin flip

\[
p(\text{word}|\text{class}) = \text{Bernoulli}(\theta_{wc})
\]

\[
p(\text{words}|\text{class}) = p(\text{word}_1|\text{class})p(\text{word}_2|\text{class})... \tag{11}
\]
“naive” “bayes” for document classification

- model presence/absence of each word as independent coin flip

\[
p(\text{word}|\text{class}) = \text{Bernoulli}(\theta_{wc}) \quad (10)
\]

\[
p(\text{words}|\text{class}) = p(\text{word}_1|\text{class})p(\text{word}_2|\text{class})\ldots (11)
\]

- maximum likelihood estimates of probabilities from word and class counts

\[
\hat{\theta}_{wc} = \frac{N_{wc}}{N_c} \quad (12)
\]
model presence/absence of each word as independent coin flip

\[ p(\text{word}|\text{class}) = \text{Bernoulli}(\theta_{wc}) \quad (10) \]

\[ p(\text{words}|\text{class}) = p(\text{word}_1|\text{class}) p(\text{word}_2|\text{class}) \ldots (11) \]

maximum likelihood estimates of probabilities from word and class counts

\[ \hat{\theta}_{wc} = \frac{N_{wc}}{N_c} \quad (12) \]

use bayes’ rule to calculate distribution over classes given words

\[ p(\text{class}|\text{words}, \Theta) = \frac{p(\text{words}|\text{class}, \Theta) p(\text{class}, \Theta)}{p(\text{words}, \Theta)} \quad (13) \]
“naive” “bayes” for document classification

- example: spam filtering for enron email using one word

\[ \hat{\theta}_{wc} = \frac{N_{wc} + \alpha}{N_{c} + \alpha + \beta} \]  

\[ \text{code: } \text{http://github.com/jhofman/ddm/blob/master/2009/lecture_03/enron_naive_bayes.sh} \]
“naive” “bayes” for document classification

- example: spam filtering for enron email using *one word*

```bash
$ ./enron1.sh meeting
1500 spam examples
3672 ham examples
16 spam examples containing meeting
153 ham examples containing meeting

estimated P(spam) = 0.2900
estimated P(ham) = 0.7100
estimated P(meeting|spam) = 0.0106
estimated P(meeting|ham) = 0.0416

P(spam|meeting) = 0.0923
```

---


jake hofman | an introduction to bayesian inference
"naive" "bayes" for document classification

- example: spam filtering for enron email using one word

```
$ ./enron1.sh meeting
1500 spam examples
3672 ham examples
16 spam examples containing meeting
153 ham examples containing meeting

estimated P(spam) = .2900
estimated P(ham) = .7100
estimated P(meeting|spam) = .0106
estimated P(meeting|ham) = .0416

P(spam|meeting) = .0923
```
```
$ ./enron1.sh money
1500 spam examples
3672 ham examples
194 spam examples containing money
50 ham examples containing money

estimated P(spam) = .2900
estimated P(ham) = .7100
estimated P(money|spam) = .1293
estimated P(money|ham) = .0136

P(spam|money) = .7957
```

\[ \hat{\theta}_{wc} = \frac{N_{wc} + \alpha}{N_c + \alpha + \beta} \]  

\(^{2}\text{code: http://github.com/jhofman/ddm/blob/master/2009/lecture_03/enron_naive_bayes.sh}\)
“naive” “bayes” for document classification

- example: spam filtering for Enron email using one word

\[
\hat{\theta}_{wc} = \frac{N_{wc} + \alpha}{N_c + \alpha + \beta}
\]

```
$ ./enron1.sh meeting
1500 spam examples
3672 ham examples
16 spam examples containing meeting
153 ham examples containing meeting

estimated P(spam) = .2900
estimated P(ham) = .7100
estimated P(meeting|spam) = .0106
estimated P(meeting|ham) = .0416

P(spam|meeting) = .0923

$ ./enron1.sh money
1500 spam examples
3672 ham examples
194 spam examples containing money
50 ham examples containing money

estimated P(spam) = .2900
estimated P(ham) = .7100
estimated P(money|spam) = .1293
estimated P(money|ham) = .0136

P(spam|money) = .7957

$ ./enron1.sh enron
1500 spam examples
3672 ham examples
0 spam examples containing enron
1478 ham examples containing enron

estimated P(spam) = .2900
estimated P(ham) = .7100
estimated P(enron|spam) = 0
estimated P(enron|ham) = .4025

P(spam|enron) = 0
```


jake hofman
an introduction to bayesian inference
example: spam filtering for Enron email using one word

\[
\hat{\theta}_{wc} = \frac{N_{wc} + \alpha}{N_c + \alpha + \beta}
\]  

naive bayes for document classification is neither naive nor bayesian!

- not so naive: works well in practice, not in theory
- not bayesian: point estimates $\hat{\theta}_{wc}$ for parameters rather than distributions over parameters
quantities of interest

- Bayesian inference maintains full posterior distributions over unknowns.
- Many quantities of interest require expectations under these posteriors, e.g. posterior mean and predictive distribution:

\[
\bar{\Theta} = \mathbb{E}_{p(\Theta|D)}[\Theta] = \int d\Theta \Theta p(\Theta|D) \quad (15)
\]

\[
p(x|D) = \mathbb{E}_{p(\Theta|D)}[p(x|\Theta, D)] = \int d\Theta p(x|\Theta, D) p(\Theta|D) \quad (16)
\]
quantities of interest

- Bayesian inference maintains full posterior distributions over unknowns.
- Many quantities of interest require expectations under these posteriors, e.g., posterior mean and predictive distribution:

\[
\bar{\Theta} = \mathbb{E}_{p(\Theta | D)} [\Theta] = \int d\Theta \Theta p(\Theta | D)
\]

(15)

\[
p(x | D) = \mathbb{E}_{p(\Theta | D)} [p(x | \Theta, D)] = \int d\Theta p(x | \Theta, D) p(\Theta | D)
\]

(16)

- Often can’t compute posterior (normalization), let alone expectations with respect to it → approximation methods.
principles (what we’d like to do)
- background: joint, marginal, and conditional probabilities
- bayes’ theorem: inverting conditional probabilities
- bayesian probability: unknowns as random variables
- bayesian inference: bayesian probability + bayes’ theorem

practice (what we’re able to do)
- monte carlo methods: representative samples
- variational methods: bound optimization
- references

application: bayesian inference for network data
representative samples

- general approach: approximate intractable expectations via sum over representative samples\(^3\)

\[
\Phi = \mathbb{E}_{p(x)}[\phi(x)] = \int dx \phi(x) p(x) \tag{17}
\]

\(^3\)follows mackay (2003)
representative samples

- general approach: approximate intractable expectations via sum over representative samples\(^3\)

\[
\Phi = \mathbb{E}_{p(x)} [\phi(x)] = \int dx \phi(x) p(x)
\]

[arbitrary function] [target density]  \hspace{1cm} (17)

down

\[
\hat{\Phi} = \frac{1}{R} \sum_{r=1}^{R} \phi(x^{(r)})
\]

(18)

\(^3\)follows mackay (2003)
representative samples

- general approach: approximate intractable expectations via sum over representative samples$^3$

\[
\Phi = \mathbb{E}_{p(x)} [\phi(x)] = \int dx \phi(x) p(x)
\]

arbitrary function \hspace{1cm} target density

\[
\Phi = \frac{1}{R} \sum_{r=1}^{R} \phi(x^{(r)})
\]

- shifts problem to finding “good” samples

$^3$follows mackay (2003)
further complication: in general we can only evaluate the target density to within a multiplicative (normalization) constant, i.e.

\[ p(x) = \frac{p^*(x)}{Z} \quad (19) \]

and \( p^*(x^{(r)}) \) can be evaluated with \( Z \) unknown
sampling methods

- monte carlo methods
  - uniform sampling
  - importance sampling
  - rejection sampling
  - ...

- markov chain monte carlo (mcmc) methods
  - metropolis-hastings
  - gibbs sampling
  - ...

an introduction to bayesian inference
1. principles (what we’d like to do)
   - background: joint, marginal, and conditional probabilities
   - Bayes’ theorem: inverting conditional probabilities
   - Bayesian probability: unknowns as random variables
   - Bayesian inference: Bayesian probability + Bayes’ theorem

2. practice (what we’re able to do)
   - Monte Carlo methods: representative samples
   - Variational methods: bound optimization
   - References

3. application: Bayesian inference for network data
Bound Optimization

- General approach: replace integration with optimization
- Construct auxiliary function upper-bounded by log-evidence, maximize auxiliary function

\[ \theta_{\text{old}} \quad \theta_{\text{new}} \]

\[ \ln p(X|\theta) \]

\[ \mathcal{L}(q, \theta) \]

variational bayes

- bound log of expected value by expected value of log using jensen’s inequality\(^5\):

\[
- \ln p(D) = - \ln \int d\Theta \ p(D|\Theta)p(\Theta) \\
= - \ln \int d\Theta \ \frac{p(D|\Theta)p(\Theta)}{q(\Theta)} q(\Theta) \\
\leq - \int d\Theta \ \ln \left[ \frac{p(D|\Theta)p(\Theta)}{q(\Theta)} \right] q(\Theta)
\]

- for sufficiently simple (i.e. factorized) approximating distribution \(q(\Theta)\), right-hand side can be easily evaluated and optimized

\(^5\)image from feynman (1972)
iterative coordinate ascent algorithm provides controlled analytic approximations to posterior and evidence
approximate posterior $q(\Theta)$ minimizes kullback-leibler distance to true posterior
resulting deterministic algorithm is often fast and scalable
Variational Bayes

- Iterative coordinate ascent algorithm provides controlled analytic approximations to posterior and evidence
- Approximate posterior \( q(\Theta) \) minimizes Kullback-Leibler distance to true posterior
- Resulting deterministic algorithm is often fast and scalable
- Complexity of approximation often limited (to, e.g., mean-field theory, assuming weak interaction between unknowns)
- Iterative algorithm requires restarts, no guarantees on quality of approximation
1 principles (what we’d like to do)
   - background: joint, marginal, and conditional probabilities
   - bayes’ theorem: inverting conditional probabilities
   - bayesian probability: unknowns as random variables
   - bayesian inference: bayesian probability + bayes’ theorem

2 practice (what we’re able to do)
   - monte carlo methods: representative samples
   - variational methods: bound optimization
   - references

3 application: bayesian inference for network data
- “Bayesian data analysis”, Gelman et al. (2003)
- “Graphical models, exponential families, and variational inference”, Wainwright & Jordan (2006)
- “What is Bayes’ theorem …”, Wiggins (2006)
- Bayesian inference view on cran
- Variational-bayes.org
- Variational Bayesian inference for network modularity
application: bayesian inference for network data
Example: A Bayesian Approach to Network Modularity
example: a bayesian approach to network modularity
example: a bayesian approach to network modularity

inferred topological communities correspond to sub-disciplines
Thanks.

Questions? \(^6\)

\(^6\) jake@jakehofman.com, @jakehofman