CMP 670
STATISTICAL NATURAL LANGUAGE PROCESSING

LECTURE 6: UNSUPERVISED LEARNING, NON-PARAMETRIC BAYESIAN MODELS

2018-2019 Spring
Frequentists vs Bayesians

• Two schools in statistics:
Frequentist school

School of Jerzy Neyman, Egon Pearson and Ronald Fisher.
Bayesian school

“School” of Thomas Bayes

\[ P(H|D) = \frac{P(D|H) \cdot P(H)}{\int P(D|H) \cdot P(H) dH} \]
Frequentists

Relative frequency of an event, $A$, is defined as

$$P(A) = \frac{\text{number of outcomes consistent with } A}{\text{number of experiments}}$$

What is the relative frequency of having a sunny weather on 4th April?
Frequentists

The definition restricts the things we can add probabilities to:

*What is the probability of there being life on Mars 100 billion years ago?*

We assume that there is an *unknown* but *fixed* underlying parameter, $\theta$, for a population (i.e., the mean height on Danish men).
Maximum Likelihood

• How will the frequentists estimate the parameter?
Maximum Likelihood

• How will the frequentist estimate the parameter?
  – Answer: maximum likelihood
Frequentists
An experiment
Frequentists
An experiment
Frequentists
An experiment
Frequentists
An experiment

167.2 cm
175.5 cm
Frequentists
An experiment
Frequentists
An experiment

167.2 cm
175.5 cm
187.7 cm
182.0 cm
Maximum Likelihood

**Basic idea**

Our best estimate of the parameter(s) are the one(s) that make our observed data most likely. We know what we have observed so far (our data). Our best "guess" would therefore be to select parameters that make our observations most likely.

Binomial distribution:

\[ P(Y = y) = \binom{n}{y} p^y (1 - p)^{n-y} \]

What type of data can we model with binomial distribution?
Bayesians

• Each investigator is entitled to his/hers personal belief, i.e. the prior information. No fixed values for parameters but a distribution.

• All distributions are subjective. Yours is as good as mine.
Bayesian Reasoning

• We have a model
  – This model encodes beliefs about the parameters as well
• We observe data
• We update our beliefs about the parameters according to the observed data
Example: speed of light

What is the speed of light in vacuum “really”? Results (m/s)

299792459.2
299792460.0
299792456.3
299792458.1
299792459.5
Example: frequentists’ solution

The average of our observations is an estimate of the true, fixed (but unknown) speed of light, \( \hat{\theta} = 299792458.6 \).

How about Bayesian solution?
Example: Bayesian solution

The observations are fixed realization from the underlying distribution of the true speed of light.

① “Guess” what the distribution of the speed of light is (the prior distribution).

② Use Bayes Theorem to modify/update the prior distribution based on the observed data.

③ The modified distribution is denoted the posterior distribution.

The posterior distribution holds the information about the true speed of light – and this distribution is entirely subjective.
Bayesian Inference

How to update knowledge, as data comes in? We use;

- **Prior distribution:** what you know about parameter $\beta$, excluding the information in the data – denoted $\pi(\beta)$
- **Likelihood:** based on modeling assumptions, how [relatively] likely the data $Y$ are *if* the truth is $\beta$ – denoted $f(Y|\beta)$

So how to get a **posterior distribution:** stating what we know about $\beta$, combining the prior with the data – denoted $p(\beta|Y)$? Bayes Theorem *used for inference* tells us to multiply;

$$p(\beta|Y) \propto f(Y|\beta) \times \pi(\beta)$$

**Posterior $\propto$ Likelihood $\times$ Prior.**

... and that’s it! (essentially!)
Simple Example

Task: Toss a (potentially biased) coin $N$ times. Compute $\theta$, the probability of heads.

Suppose we observe: $\{T, H, H, T\}$. What do we think $\theta$ is?
Simple Example

Task: Toss a (potentially biased) coin $\mathcal{N}$ times. Compute $\theta$, the probability of heads.

Suppose we observe: \{T, H, H, T\}. What do we think $\theta$ is? The maximum likelihood estimate is $\theta = 1/2$. Seems reasonable.
Simple Example

Task: Toss a (potentially biased) coin $N$ times. Compute $\theta$, the probability of heads.

Suppose we observe: \{T, H, H, T\}. What do we think $\theta$ is? The maximum likelihood estimate is $\theta = 1/2$. Seems reasonable.

Now suppose we observe: \{H, H, H, H\}. What do we think $\theta$ is?
Task: Toss a (potentially biased) coin $N$ times. Compute $\theta$, the probability of heads.

Suppose we observe: \{T, H, H, T\}. What do we think $\theta$ is? The maximum likelihood estimate is $\theta = 1/2$. Seems reasonable.

Now suppose we observe: \{H, H, H, H\}. What do we think $\theta$ is? The maximum likelihood estimate is $\theta = 1$. Seem reasonable?
Simple Example

Task: Toss a (potentially biased) coin $N$ times. Compute $\theta$, the probability of heads.

Suppose we observe: \{T, H, H, T\}. What do we think $\theta$ is? The maximum likelihood estimate is $\theta = 1/2$. Seems reasonable.

Now suppose we observe: \{H, H, H, H\}. What do we think $\theta$ is? The maximum likelihood estimate is $\theta = 1$. Seem reasonable?

Not really. Why?
Simple Example

When we observe \{H, H, H, H\}, why does $\theta = 1$ seem unreasonable?
Simple Example

When we observe \{H, H, H, H\}, why does $\theta = 1$ seem unreasonable?

Prior knowledge! We believe coins generally have $\theta \approx 1/2$. How to encode this? By using a Beta prior on $\theta$. 
Bayesian Approach to Estimating $\theta$

Place a Beta$(a, b)$ prior on $\theta$. This prior has the form

$$p(\theta) \propto \theta^{a-1}(1 - \theta)^{b-1}.$$ 

What does this distribution look like?
Bayesian Approach to Estimating $\theta$

After observing $X$, a sequence with $n$ heads and $m$ tails, the posterior on $\theta$ is:

$$p(\theta|X) \propto p(X|\theta)p(\theta)$$
$$\propto \theta^{a+n-1}(1-\theta)^{b+m-1}$$
$$\sim \text{Beta}(a+n, b+m).$$
Bayesian Approach to Estimating $\theta$

After observing $X$, a sequence with $n$ heads and $m$ tails, the posterior on $\theta$ is:

$$ p(\theta|X) \propto p(X|\theta)p(\theta) $$

$$ \propto \theta^{a+n-1}(1-\theta)^{b+m-1} $$

$$ \sim \text{Beta}(a + n, b + m). $$

If $a = b = 1$ and we observe 5 heads and 2 tails, Beta(6, 3) looks like

![Graph showing Beta(6, 3)]
Where do priors come from?

• Scientists & statisticians all make assumptions... even if they do not talk about them.
Bayesian vs Non-Bayesian

• The Non-Bayesian model:

\[ P(X \mid \theta) \]

• The Bayesian model:

\[ P(X, \theta) = P(\theta)P(X \mid \theta) \]
Multinomial-Dirichlet

\[ X = (0, \ldots, 1, \ldots, 0) \] representing a certain outcome

- Same derivation as we had with the Beta prior:

\[ p(\theta \mid \alpha) \propto \prod_{i=1}^{d} \theta_i^{{\alpha_i} - 1} \]
\[ p(X \mid \theta) \propto \prod_{i=1}^{d} \theta_i^{X_i} \]

- Multiply those together:

\[ p(\theta \mid X) \propto \left( \prod_{i=1}^{d} \theta_i^{X_i} \right) \times \left( \prod_{i=1}^{d} \theta_i^{\alpha_i - 1} \right) = \prod_{i=1}^{d} \theta_i^{X_i + \alpha_i - 1} \]

- This is a Dirichlet with parameters \( X + \alpha \)

What type of data can be modelled with Multinomial-Dirichlet?

Give some example NLP tasks.
Having a likelihood does not necessarily make it easy to work with. In Bayesian statistics the posterior distribution contains all relevant information about the parameters. Statistical inference is often calculated from summaries (integrals)

\[ J = \int L(x) \, dx \]

However, these evaluations are not necessarily easy.
• Think of every possible assignment to $y$ as a state in a huge (maybe infinite) Markov model.

• Gibbs sampling (and other MCMC methods, such as Metropolis-Hastings) equate to random walks on this highly connected graph.

• The theory of Markov chains tells us that under certain conditions, we will eventually reach the stationary distribution (which is the joint $p(y \mid x)$ we want)
Inference with Sampling

Gibbs Sampling

• Sampling directly from a complicated posterior distribution (always given the evidence, \(x\)) is hard.

• Consider all hidden variables (structure, parameters, etc.) as a set \([y_1, y_2, \ldots, y_n]\). Initialize everything.

• A single iteration:

  • For \(i = 1 \ldots n\):

    • Randomly sample \(y_i \sim p(y_i \mid x, y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)\).

• After “enough iterations,” you will be sampling from the joint distribution over all \(y_i\).
Integrating out Parameters

\[ p(X_i = x \mid x_{-i}, \beta) = \frac{\text{count}(x; x_{-i}) + \beta}{n - 1 + M \beta} \]

"all x excluding x_i"
Metropolis-Hastings Algorithm

- Sometimes it is not clear how to find the conditionals to do Gibbs sampling
- Or they could be intractable to compute
- Especially the normalization constant of the distribution
- Metropolis-Hastings is a way to do MCMC sampling when the normalization constant is hard to compute
Metropolis-Hastings

- Main idea: sample from a proposal distribution and correct by accepting/rejecting the samples to get samples from the real distribution

- To traverse the Markov chain:
  - At time step $t$, sample $y'$ from a proposal distribution
  - Sample $\alpha$ from uniform $[0,1]$
  - Set $y_{t+1} = y'$ if
    $$\alpha < \frac{p(y)q(y_t; y')}{p(y_t)q(y'; y_t)}$$
Convergence

• Sampling is notoriously slow

• In fact, you cannot necessarily be certain when the Markov chain has “mixed”

• There are several heuristics:
  • If you can evaluate the likelihood, check if it stabilizes
  • If there is an extrinsic evaluation, you can evaluate on that measure
  • Run several Markov chains, and compare a scalar parameter across all these chains
  • In NLP, in many cases, we run a sampler for a fixed number of iterations that is chosen ahead
NON-PARAMETRIC BAYESIAN MODELS
Modeling Data

*All models are wrong, but some are useful.*


> Models are never correct for real world data.
Nonparametric: Does NOT mean there are no parameters.
Nonparametric Bayesian Modeling

Now we know what nonparametric and Bayesian mean. What should we expect from nonparametric Bayesian methods?

- Complexity of our model should be allowed to grow as we get more data.
Nonparametric Bayesian Methods

- Dirichlet Process/Chinese Restaurant Process
  Latent class models - often used in the clustering context

- Beta Process/Indian Buffet Process
  Latent feature models

- Gaussian Process (No culinary metaphor - oh well)
  Regression

Today we focus on the Dirichlet Process!
\{(G(A_1), \ldots, G(A_n)) \sim \text{Dir}(\alpha_0 G_0(A_1), \ldots, \alpha_0 G_0(A_n))\}
Dirichlet Process

- $\alpha$ - The concentration parameter.

- $G_0$ - The base measure. A prior distribution for the cluster specific parameters.

The Dirichlet Process (DP) is a *distribution over distributions*. We write

$$G \sim DP(\alpha, G_0)$$

to indicate $G$ is a distribution drawn from the DP.

It will become clearer in a bit what $\alpha$ and $G_0$ are.
Dirichlet Process

\[ G \sim \text{DP}(\alpha, G_0) \]

The CRP describes the partitions of \( \theta \) when \( G \) is marginalized out.

Stick-Breaking Process (just the weights)
Dirichlet Process

Dirichlet process is a probability distribution over distributions:

\[ G \sim DP(\alpha, G_0) \]
\[ X_n \sim G \]
The first customer sits at the first table with probability $\frac{1}{5+\alpha}$, sits at the second table with probability $\frac{2}{5+\alpha}$, sits at the third table with probability $\frac{2}{5+\alpha}$, and can sit at an empty table with probability $\frac{\alpha}{5+\alpha}$.
Apply CRP for Morphological Segmentation

- Stem restaurant
- order \(3/(\alpha+6)\)
- jump \(1/(\alpha+6)\)
- danc \(2/(\alpha+6)\)
- Empty table \(\alpha/(\alpha+6)\)
- ...
Apply CRP for Morphological Segmentation

Suffix restaurant

ed \(\frac{3}{(\alpha+6)}\)

s \(\frac{1}{(\alpha+6)}\)

ing \(\frac{2}{(\alpha+6)}\)

Empty table \(\frac{\alpha}{(\alpha+6)}\)
Apply CRP for Morphological Segmentation

New word: drinking

Stem restaurant

order  jump  danc  Empty table  ...

$$p(\text{drink}|\text{drinking}) = \alpha/(\alpha+6)$$

Suffix restaurant

ed  s  ing  Empty table  ...

$$p(\text{ing}|\text{drinking}) = 2/(\alpha+6)$$

{danced, dancing, jumped, ordered, orders, ordering}
Apply CRP for Morphological Segmentation

\[ p(w = s_1 + s_2 + \cdots + s_n) = \prod_i p(s_i) \]

\[ G_s \sim DP(\alpha_s, H_s) \]

\[ s \sim G_s \]
REFERENCES

- Michael Paul, Natural Language Processing Class, Johns Hopkins University
- Roger Levy, Statistical NLP Class, Unsupervised Text Categorization
- Tom Griffiths, Topic Models
- Ken Rice, Bayesian Statistics
- Claus Ekstrøm, Frequentist and Bayesian Statistics
- Shay Cohen, The Bayesian approach
- Serkan Kumyol, MSc thesis
- Kurt Miller, Nonparametric Bayes, Berkeley University