Topic Modeling
Lecture 9: October 9, 2013

CS886-2 Natural Language Understanding

University of Waterloo

Information Retrieval Example #1

Search Query: Batman

The Batman, a masked hero in the tradition of Zorro and The Scarlet Pimpernel, first appeared in Detective Comics #27, dated May, 1939

Content A

Banished from their primitive village, they set off on an epic journey through the ancient world

Content B

Information Retrieval Example #1

Search Query: Batman

The Batman, a masked hero in the tradition of Zorro and The Scarlet Pimpernel, first appeared in Detective Comics #27, dated May, 1939

Content A

Banished from their primitive village, they set off on an epic journey through the ancient world

Content B

Solution: Keyword Usage

Since Content A contains the word "Batman" and Content B does not, the engine can easily choose which one to rank.

Information Retrieval Example #2

Search Query: Chief Wiggum

Clumsily scraping up the remains of the dried Squishee from the floor, Wiggum hastily shoved them into his mouth.

Content A

If you chose to do that, first let him know that you have spoken to the chief.

Content B

Information Retrieval Example #2

Search Query: Chief Wiggum

Clumsily scraping up the remains of the dried Squishee from the floor, Wiggum hastily shoved them into his mouth.

Content A

If you chose to do that, first let him know that you have spoken to the chief.

Content B

Solution: TF*IDF

The search engine can use TF*IDF (Term Frequency x Inverse Document Frequency) to determine that "Wiggum" is a much less common word than "chief" and thus, Content A is more relevant to the query than Content B.

Information Retrieval Example #3

Search Query: Pianist

Dropping his meeting notes at the door, he jiggled the keys into the lock but found it wouldn't budge.

Content A

Her hands mercilessly pounded the keys, notes cascading into the surrounding stairway.

Content B

Information Retrieval Example #3

Search Query: Pianist

Dropping his meeting notes at the door, he jiggled the keys into the lock but found it wouldn't budge.

Content A

Her hands mercilesslly pounded the keys, notes cascading into the surrounding stairway.

Content B

Solution: Topic Modeling

As humans reading both sentences, we can infer that Content B is obviously about the musical instrument - a piano - and the woman playing it. But a search engine armed with only the methods we described above will struggle since both sentences use the words "keys" and "notes," some of the only clues to the puzzle.

Latent Semantic Analysis

- Idea: singular value decomposition
  - Infer latent space in which documents or words can be described more succinctly
- Issues:
  - How do we interpret this latent space?
  - How many dimensions should it have?
  - How can we represent uncertainty/ambiguities?

Latent Dirichlet Allocation

- Idea: probabilistic generative model for documents
  - Latent variables often correspond to topics
  - Some machine learning techniques can automatically infer the # of topics
  - Probabilistic framework allows us to quantify uncertainty/ambiguities

Graphical Model

- Picture

Plate Model

- Picture

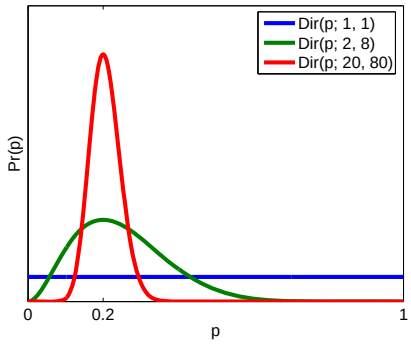
Dirichlet

- Definition

$$Dir(p; \alpha_1, \alpha_2) = kp^{\alpha_1 - 1}(1-p)^{\alpha_2 - 1}$$

- p: probability of head
- $\alpha_1 - 1$: # of heads
- $\alpha_2 - 1$: # of tails

- Mean: $\alpha_1/(\alpha_1 + \alpha_2)$



Conjugate Prior

- Bayesian learning
  - Prior: $\Pr(p) = Dir(p; \alpha_1, \alpha_2)$
  - Posterior: $\Pr(p|HTHH)$
- Bayes theorem:

$$\begin{aligned}\Pr(p|HTHH) &\propto \Pr(p)\Pr(HTHH|p) \\ &= Dir(p; \alpha_1, \alpha_2)p^3(1-p) \\ &= kp^{\alpha_1+3-1}(1-p)^{\alpha_2+1-1} \\ &= Dir(p; \alpha_1+3, \alpha_2+1)\end{aligned}$$

Topic Modeling

- Task
  - Infer topics and parameters: $\Pr(T_{1:n}, \theta, \phi|W_{1:n})$
- Two common approaches
  - Gibbs sampling
    - Simple, but stochastic and slow
  - Variational Bayes (variant of EM)
    - Complex, but deterministic and fast

Sampling Techniques

- Direct sampling
- Rejection sampling
- Likelihood weighting
- Importance sampling
- Markov chain Monte Carlo (MCMC)
  - Gibbs Sampling
  - Metropolis-Hastings
- Sequential Monte Carlo sampling (a.k.a. particle filtering)

Approximate Inference by Sampling

- Expectation: $E_P[f(x)] = \int_x P(x)f(x)dx$
  - Approximate integral by sampling:
    $E_P[f(x)] \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)$ where $x_i \sim P(x)$

- Inference query: $\Pr(\boldsymbol{X}|e) = \sum_{\boldsymbol{Y}} \Pr(\boldsymbol{X}, \boldsymbol{Y}|e)$
  - Approximate exponentially large sum by sampling:
    $\Pr(\boldsymbol{X}|e) = \frac{1}{n}\sum_{i=1}^{n} \Pr(\boldsymbol{X}|\boldsymbol{y}_i, e)$ where $\boldsymbol{y}_i \sim P(\boldsymbol{Y}|e)$

Direct Sampling (a.k.a. forward sampling)

- Unconditional inference queries (i.e., $\Pr(V = t)$)
- Bayesian networks only
  - Idea: sample each variable given the values of its parents according to the topological order of the graph.

Direct Sampling Algorithm

Sort the variables by topological order

For $i = 1$ to n do (sample n particles)

For each variable V_j do

Sample $v_j^{(i)} \sim \Pr(V|\boldsymbol{pa}_V)$

- Approximation: $\Pr(V_k = t) \approx \frac{1}{n}\sum_{i=1}^{n} \delta(v_k^{(i)} = t)$

Example

Analysis

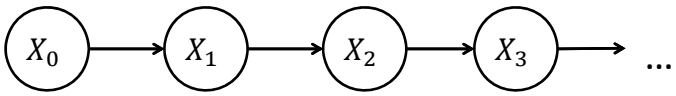
- Complexity: $O(n|\boldsymbol{V}|)$ where $|\boldsymbol{V}| = \#\text{variables}$
- Accuracy
  - Absolute error ϵ: $\mathrm{P}\left(\left|\widehat{\mathrm{P}}(\mathrm{V}) - \mathrm{P}(\mathrm{V})\right| > \epsilon\right) \leq \delta = 2e^{-2n\epsilon^2}$
    - Sample size $n \geq \frac{\ln\left(\frac{2}{\delta}\right)}{2\epsilon^2}$
  - Relative error ϵ: $P\left(\frac{\hat{P}(V)}{P(V)} \notin [1-\epsilon, 1+\epsilon]\right) \leq \delta = 2e^{-\frac{nP(V)\epsilon^2}{3}}$
    - Sample size $n \geq \frac{3\ln\left(\frac{2}{\delta}\right)}{2P(V)\epsilon^2}$

Markov Chain Monte Carlo

- Iterative sampling technique that converges to the desired distribution in the limit
- Idea: set up a Markov chain such that its stationary distribution is the desired distribution

Markov Chain

- Definition: A Markov chain is a linear chain Bayesian network with a stationary conditional distribution known as the transition function



- Initial distribution: $\Pr(X_0)$
- Transition distribution: $\Pr(X_t|X_{t-1})$

Asymptotic Behaviour

- Let $\Pr(X_t)$ be the distribution at time step t

$$\Pr(X_t) = \sum_{\boldsymbol{X}_{0..t-1}} \Pr(\boldsymbol{X}_{0..t})$$
$$= \sum_{X_{t-1}} \Pr(X_{t-1}) \Pr(X_t | X_{t-1})$$

- In the limit (i.e., when $t \to \infty$), the Markov chain may converge to stationary distribution $\pi(x) = \Pr(X_\infty = x)$

$$\pi(x) = \Pr(X_\infty = x)$$
$$= \sum_{X_{\infty-1}} \Pr(X_{\infty-1} = x') \Pr(X_\infty = x | X_{\infty-1} = x')$$
$$= \sum_{x'} \pi(x') \Pr(x | x')$$

Stationary distribution

- Let $T_{x|x'} = \Pr(x|x')$ be a matrix that represents the transition function

- If we think of π as a column vector, then π is an eigenvector of T with eigenvalue 1

$$T\pi = \pi$$

Ergodic Markov Chain

- Definition: A Markov chain is ergodic when there is a non-zero probability of reaching any state from any state in a finite number of steps
- When the Markov chain is ergodic, there is a unique stationary distribution

- Sufficient condition: detailed balance

$$\pi(x)\Pr(x'|x) = \pi(x')\Pr(x|x')$$

Detailed balance → ergodicity → unique stationary dist.

Markov Chain Monte Carlo

- Idea: set up an ergodic Markov chain such that the unique stationary distribution is the desired distribution

- Since the Markov chain is a linear chain Bayes net, we can use direct sampling (forward sampling) to obtain a sample of the stationary distribution

Generic MCMC Algorithm

Sample $x_0 \sim \Pr(X_0)$

For $i = 1$ to n do (sample n particles)

Sample $x_t \sim \Pr(X_t | x_{t-1})$

- Approximation: $\pi(x) \approx \frac{1}{n} \sum_{t=1}^{n} \delta(x_t = x)$
- In practice, ignore the first k samples for a better estimate (burn-in period):

$$\pi(x) \approx \frac{1}{n-k} \sum_{t>k}^{n} \delta(x_t = x)$$

Choosing a Markov Chain

- Different Markov chains lead to different algorithms
  - Gibbs sampling
  - Metropolis Hastings

Gibbs Sampling

- Suppose $\Pr(\boldsymbol{X})$ defined by a graphical model (Bayes net or Markov net)
- Inference query: $\Pr(\boldsymbol{Y}|\boldsymbol{e})$? Where $\boldsymbol{Y} \subseteq \boldsymbol{X}$

- Idea: randomly assign values to all non-evidence variables, then repeatedly sample each non-evidence variable given the assigned values for all other variables

Gibbs Sampling Algorithm

Randomly assign $v_j^{(0)}$ to all non-evidence variables V_j

For $i = 1$ to n do (sample n particles)

For each non-evidence variable V_j do

Sample $v_j^{(i)} \sim \Pr\left(V_j \middle| \boldsymbol{v}_{\sim j}^{(i-1)}, \boldsymbol{e}\right)$

- Approximation: $\Pr(V_k = t|\boldsymbol{e}) \approx \frac{1}{n}\sum_{i=1}^{n} \delta(v_k^{(i)} = t)$

Example

Practical Consideration

- Burn-in period: ignore first k samples:

$$\Pr(V_k = t|\boldsymbol{e}) \approx \frac{1}{n-k}\sum_{i>k}^{n} \delta(v_k^{(i)} = t)$$

- Use most recent values to sample $V_j^{(i)}$

$$v_j^{(i)} \sim \Pr(V_j^{(i)}|\boldsymbol{v}_{1\ldots j-1}^{(i)}, \boldsymbol{v}_{j+1\ldots|V|}^{(i-1)})$$

- Use conditional independence to restrict parent variables to the Markov blanket

$$v_j^{(i)} \sim \Pr(V_j^{(i)}|\boldsymbol{v}_{\forall k<j, k\in mb(j)}^{(i)}, \boldsymbol{v}_{\forall k>j, k\in mb(j)}^{(i-1)})$$

Convergence

- Let $\Pr(\boldsymbol{V}^{(i)}|\boldsymbol{V}^{(i-1)}, \boldsymbol{e})$ be the transition function of the Markov chain associated with Gibbs sampling
- **Theorem:** Gibbs sampling converges to $\Pr(\boldsymbol{V}|\boldsymbol{e})$ when all potentials are strictly positive.
- Proof: $\Pr(\boldsymbol{V}^{(i)}|\boldsymbol{V}^{(i-1)}, \boldsymbol{e})$ satisfies detailed balance
  i.e. $\Pr(\boldsymbol{V}|\boldsymbol{e}) \Pr(\boldsymbol{V}'|\boldsymbol{V}, \boldsymbol{e}) = \Pr(\boldsymbol{V}'|\boldsymbol{e}) \Pr(\boldsymbol{V}|\boldsymbol{V}', \boldsymbol{e})$