Inference and Representation: Lab 10
Network Models

Yacine Jernite

November 13, 2014
Lecture plan

- Further notes on MCMC and variational methods
- Review of last week’s methods
- Network models
Let \((X_i)\) be a sequence of samples from a Markov Chain with equilibrium distribution \(\pi\), and \(f : \mathbb{R} \to \mathbb{R}\). Let \(\bar{f}_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i)\), then, if \(\mathbb{E}_\pi[|f|] < \infty\), we have:

- \(\bar{f}_n \to \mathbb{E}_\pi[f]\) almost surely.
- \(\sqrt{n}(\bar{f}_n - \mathbb{E}_\pi[f])\) converges weakly to \(\mathcal{N}(0, \sigma(f)^2)\)


http://stat.rutgers.edu/home/rongchen/papers/tierney.pdf
Recall from class:

\[ p(x; \theta) = \exp\left(\sum_{c \in C} \theta_c(x_c) - \ln(Z(\theta))\right) \]

Where

\[ Z(\theta) = \sum_{x'} \exp\left(\sum_{c \in C} \theta_c(x'_c)\right) \]

Then, \( \forall q \) distributions,

\[ D_{KL}(q\|p) = \ln(Z(\theta)) - \sum_{c \in C} \mathbb{E}_q[\theta_c(x_c)] - H(q) \]

See class slides 2-3: computing \( Z \) as an optimization problem, getting approximate marginals.
For a model with hidden variables, instead of $Z$, compute likelihood

$$p(x; \theta) = \sum_z p(x, z; \theta)$$

Start from the observation:

$$\forall z, p(x; \theta) = \frac{p(x, z; \theta)}{p(z|x; \theta)}$$

Hence, for any distribution $q$, $\forall z'$:

$$\ln(p(x; \theta)) = \ln\left(\frac{p(x, z'; \theta)}{p(z'|x; \theta)}\right) = \sum_z q(z) \ln\left(\frac{p(x, z; \theta)}{p(z|x; \theta)}\right)$$
This gives us:

\[
\ln(p(x; \theta)) = \sum_z q(z) \ln(p(x, z; \theta)) - \sum_z q(z) \ln(p(z|x; \theta)) \\
= \sum_z q(z) \ln(p(x, z; \theta)) - \sum_z q(z) \ln(q(z)) \\
+ \sum_z q(z) \ln \left( \frac{q(z)}{p(x, z; \theta)} \right) \\
= \sum_z q(z) \ln(p(x, z; \theta)) + H(q) + D_{KL}(q||p(\cdot|x; \theta))
\]

Computing the likelihood as an optimization problem (and getting a lower bound), obtaining approximate posterior distribution.
Variational Methods: optimizing over moments

Going back to the partition function case:

\[
\ln(Z(\theta)) = \max_{q \in \text{Distr}} \left( \sum_{c \in C} \mathbb{E}_q[\theta_c(x_c)] + H(q) \right) = \max_{q \in \text{Distr}} f(q)
\]

\[
= \max_{q \in \text{Distr}} \left( \sum_{c \in C} \sum_{x_c} q_c(x_c) \theta_c(x_c) + H(q) \right)
\]

\[
= \max_{q \in \text{Distr}} \left( \sum_{c \in C} \sum_{x_c} \mu_{q,c,x_c} \theta_c(x_c) + H^*(\mu_q) \right)
\]

\[
= \max_{q \in \text{Distr}} g(\mu_q) = \max_{\mu_q \in \mathcal{M}} g(\mu_q) = \max_{\mu \in \mathcal{M}} g(\mu)
\]

Where \(\mathcal{M}\) (the marginal polytope) is exactly the set of moments that could have been generated by some distribution \(q\).

**Problem:** the marginal polytope is difficult to optimize over (see class slide 6).
The mean field approximation optimizes over a set of distributions $q$ that is **strictly smaller** than Distr. We will call it $\text{Distr}_{mf} \subsetneq \text{Distr}$, then:

$$\ln(Z(\theta)) = \max_{q \in \text{Distr}} f(q) \geq \max_{q \in \text{Distr}_{mf} \subsetneq \text{Distr}} f(q)$$

Let $\mathcal{M}_{mf}$ (the marginal polytope) be the set of moments that could have been generated by some distribution $q$ in $\text{Distr}_{mf}$. Then:

$$\mathcal{M}_{mf} \subsetneq \mathcal{M}$$

And:

$$\max_{q \in \text{Distr}_{mf}} f(q) = \max_{\mu \in \mathcal{M}_{mf}} g(q) \leq \max_{\mu \in \mathcal{M}} g(q) = \ln(Z(\theta))$$
Similarly, if we optimize \( \mu \) over some bigger set \( \mathcal{M}^{LC} \supseteq \mathcal{M} \), we get the upper bound:

\[
\max_{\mu \in \mathcal{M}^{LC}} g(\mu) \geq \max_{\mu \in \mathcal{M}} g(\mu) = \ln(Z(\theta))
\]

We chose \( \mathcal{M}^{LC} \) to be the **local consistency polytope** defined by:

\[
\sum_{x_{\bar{s}}} \mu_c(x_{c\setminus s}, x_s) = \mu_c\setminus s(x_{c\setminus s})
\]

In other words: pseudo-marginals consistent with each other.

Proof that \( \mathcal{M}^{LC} \supseteq \mathcal{M} \): triangle with \( \mu_{i,j}(x_i, x_j) = 0.5 \times \mathbb{I}[x_i \neq x_j] \)
Other notes

- Maximum entropy optimization
- Bethe free energy
- Tree Re-Weighted approximation
Review of last week: modifying a model

Figure: Starting from a generic model

\[ \alpha \]
\[ Z \]
\[ X \]
Review of last week: modifying a model

We saw extensions which allowed to:

- Add supervision to learning max-likelihood parameters
- Improve modelling of the latent variables distribution
- Add a model of time
- Add a model of influence
Review of last week: adding supervision

Figure: Adding $P(Y|Z)$
Review of last week: adding supervision

- Examples: discriminative LDA, MedLDA
- Gives $P(Y|Z)$, and directs the learning of $\alpha$
Review of last week: better latent distribution

Figure: Changing $P(Z|\alpha)$
Review of last week: better latent distribution

- Examples: Correlated Topic Model, Pachinko Allocation Model
- Can allow to find more specific models, (example, if two topics often cooccur)
Review of last week: modelling evolution

Figure: Adding time

\[ \alpha_{t-1} \xrightarrow{} \alpha_t \]
\[ Z_{t-1} \xrightarrow{} Z_t \]
\[ X_{t-1} \xrightarrow{} X_t \]
### Review of last week: Time

<table>
<thead>
<tr>
<th>Topic #2</th>
<th>Topic #4</th>
<th>Topic #19</th>
<th>Topic #20</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>2009-2010</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Electro</td>
<td>Hip hop</td>
<td>Indie rock</td>
</tr>
<tr>
<td>2</td>
<td>Tech house</td>
<td>Rap</td>
<td>Indie</td>
</tr>
<tr>
<td>3</td>
<td>Techno</td>
<td>Hardcore rap</td>
<td>Acoustic</td>
</tr>
<tr>
<td>4</td>
<td>Electronica</td>
<td>Soul</td>
<td>Indie rock</td>
</tr>
<tr>
<td>5</td>
<td>Deep house</td>
<td>Reggae</td>
<td>Singer-songwriter</td>
</tr>
<tr>
<td><strong>1999-2000</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Hip hop</td>
<td>Hip hop</td>
<td>Alternative rock</td>
</tr>
<tr>
<td>2</td>
<td>Downtempo</td>
<td>Rap</td>
<td>Punk</td>
</tr>
<tr>
<td>3</td>
<td>Electronica</td>
<td>Reggae</td>
<td>Alternative</td>
</tr>
<tr>
<td>4</td>
<td>Trip hop</td>
<td>Hardcore rap</td>
<td>Indie</td>
</tr>
<tr>
<td>5</td>
<td>Electro</td>
<td>Gangster rap</td>
<td></td>
</tr>
<tr>
<td><strong>1989-1990</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Hip hop</td>
<td>Hip hop</td>
<td>Alternative rock</td>
</tr>
<tr>
<td>2</td>
<td>Electro</td>
<td>Pop rock</td>
<td>Punk</td>
</tr>
<tr>
<td>3</td>
<td>Techno</td>
<td>Classic rock</td>
<td>Classic rock</td>
</tr>
<tr>
<td>4</td>
<td>Pop rap</td>
<td>Jazz</td>
<td>Pop rock</td>
</tr>
<tr>
<td>5</td>
<td>Downtempo</td>
<td>Funk</td>
<td>Classic rock</td>
</tr>
<tr>
<td><strong>1979-1980</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Funk</td>
<td>Disco</td>
<td>New wave</td>
</tr>
<tr>
<td>2</td>
<td>Jazz</td>
<td>Funk</td>
<td>Classic rock</td>
</tr>
<tr>
<td>3</td>
<td>Disco</td>
<td>New wave</td>
<td>Pop rock</td>
</tr>
<tr>
<td>4</td>
<td>Reggae</td>
<td>Classic rock</td>
<td>Jazz</td>
</tr>
<tr>
<td>5</td>
<td>Soul</td>
<td>Pop rock</td>
<td>Singers songwriter</td>
</tr>
<tr>
<td><strong>1969-1970</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Classic rock</td>
<td>Classic rock</td>
<td>Classic rock</td>
</tr>
<tr>
<td>2</td>
<td>Jazz</td>
<td>Blues</td>
<td>Psychedelic rock</td>
</tr>
<tr>
<td>3</td>
<td>Blues</td>
<td>Psychedelic rock</td>
<td>Blues</td>
</tr>
<tr>
<td>4</td>
<td>Pop rock</td>
<td>Classic rock</td>
<td>Classic rock</td>
</tr>
<tr>
<td>5</td>
<td>Psychedelic rock</td>
<td>Classic rock</td>
<td>Blues</td>
</tr>
</tbody>
</table>

---

Yacine Jernite

Inference and Representation: Lab 10 Network Models
Review of last week: modelling influence

Figure: Adding time and influence

\[ \alpha_{t-1} \rightarrow \alpha_t \]
\[ Z_{t-1} \rightarrow Z_t \]
\[ X_{t-1} \rightarrow X_t \]
Review of last week: time and evolution

- Keep time dependences in variational EM
  \[ \alpha_t \sim \mathcal{N}(\alpha_{t-1}, \Sigma) \]

- Influence model: influential topics “pull \( \alpha \) their way”
  \[ \alpha_t \sim \mathcal{N}(\alpha_{t-1} + \sum_{d=1}^{D} I(d)Z_d, \Sigma) \]
## Review of last week: Influence

<table>
<thead>
<tr>
<th>Artist name</th>
<th>Song names</th>
<th>Year</th>
<th>Epoch influence rank in Topic #</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob Dylan</td>
<td>Rainy Day Women #12 &amp; 35, Like a Rolling Stone</td>
<td>1966, 1965</td>
<td>ranked 1&lt;sup&gt;st&lt;/sup&gt;, 2&lt;sup&gt;nd&lt;/sup&gt; in topic #16</td>
</tr>
<tr>
<td>Killing Joke</td>
<td>The Wait</td>
<td>1979</td>
<td>ranked 1&lt;sup&gt;st&lt;/sup&gt; in topic #13</td>
</tr>
<tr>
<td>Beastie Boys</td>
<td>Paul Revere</td>
<td>1986</td>
<td>ranked 1&lt;sup&gt;st&lt;/sup&gt; in topic #4</td>
</tr>
<tr>
<td>Run-D.M.C.</td>
<td>Is It Live</td>
<td>1986</td>
<td>ranked 2&lt;sup&gt;nd&lt;/sup&gt; in topic #4, ranked 2&lt;sup&gt;nd&lt;/sup&gt; in topic #5</td>
</tr>
</tbody>
</table>
Modelling networks

- Examples of network datasets
- Building the Stochastic Block Model
- Adding topic information: Relational Topic Model
- Random Subgraph Model and ecclesiastic network (next week)
- Multiple classes: Mixed Membership versus Overlapping SBM (next week)
Sample datasets

A Survey of Statistical Network Models, Goldenberg et al. ,2009  

- Sampson monastery study
- Protein interaction network in budding yeast
- NIPS co-authorship dataset
- Blog hyperlinks network
- Medieval ecclesiastic councils
Sampson monastery study
Protein interaction network
NIPS co-authorship dataset

- **Irrelevant features and the subset selection problem**
  - We address the problem of finding a subset of features that allows a supervised induction algorithm to induce small high-accuracy concepts...

- **Utilizing prior concepts for learning**
  - The inductive learning problem consists of learning a concept given examples and nonexamples of the concept. To perform this learning task, inductive learning algorithms bias their learning method...

- **Learning with many irrelevant features**
  - In many domains, an appropriate inductive bias is the MIN-FEATURES bias, which prefers consistent hypotheses definable over as few features as possible...

- **Evaluation and selection of biases in machine learning**
  - In this introduction, we define the term bias as it is used in machine learning systems. We motivate the importance of automated methods for evaluating...

- **An evolutionary approach to learning in robots**
  - Evolutionary learning methods have been found to be useful in several areas in the development of intelligent robots. In the approach described here, evolutionary...

- **Improving tactical plans with genetic algorithms**
  - The problem of learning decision rules for sequential tasks is addressed, focusing on the problem of learning tactical plans from a simple flight simulator where a plane must avoid a missile...

- **Using a genetic algorithm to learn strategies for collision avoidance and local navigation**
  - Navigation through obstacles such as mine fields is an important capability for autonomous underwater vehicles. One way to produce robust behavior...
Sampson study: building a model

- What are the observed variables?
- What are the latent variables?
- What are the probability distributions?
What are the observed variables?
- Adjacency matrix
  \[ X_{i,j} \in \{0, 1\} \]

What are the latent variables?
- Categorical clusters
  \[ z_i \in [1, K] \]

What are the probability distributions?
- Matrix of Bernoulli parameters
  \[ P(X_{ij} = 1) = \prod_{z_i, z_j} \]
Sampson study: results

Yacine Jernite

Inference and Representation: Lab 10 Network Models
Relational Topic Model, Chang and Blei, 2009

- Usual formulation of LDA, with the added:

\[ p(X_{d_i,d_j} = 1) = \sigma(\langle \eta, z_{d_i} \circ z_{d_j} \rangle + \nu) \]
Adding Topic Model Information

Relational topic models

- Adapt fitting algorithm for sLDA with binary GLM response
- RTMs allow predictions about new and unlinked data.
- These predictions are out of reach for traditional network models.
### Competitive environments evolve better solutions for complex tasks

<table>
<thead>
<tr>
<th>Coevolving High Level Representations</th>
</tr>
</thead>
<tbody>
<tr>
<td>A Survey of Evolutionary Strategies</td>
</tr>
<tr>
<td>Genetic Algorithms in Search, Optimization and Machine Learning</td>
</tr>
<tr>
<td>Strongly typed genetic programming in evolving cooperation strategies</td>
</tr>
<tr>
<td>Solving combinatorial problems using evolutionary algorithms</td>
</tr>
<tr>
<td>A promising genetic algorithm approach to job-shop scheduling, rescheduling, and open-shop scheduling problems</td>
</tr>
<tr>
<td>Evolutionary Module Acquisition</td>
</tr>
<tr>
<td>An Empirical Investigation of Multi-Parent Recombination Operators in Evolution Strategies</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A New Algorithm for DNA Sequence Assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identification of protein coding regions in genomic DNA</td>
</tr>
<tr>
<td>Solving combinatorial problems using evolutionary algorithms</td>
</tr>
<tr>
<td>A promising genetic algorithm approach to job-shop scheduling, rescheduling, and open-shop scheduling problems</td>
</tr>
<tr>
<td>A genetic algorithm for passive management</td>
</tr>
<tr>
<td>The Performance of a Genetic Algorithm on a Chaotic Objective Function</td>
</tr>
<tr>
<td>Adaptive global optimization with local search</td>
</tr>
<tr>
<td>Mutation rates as adaptations</td>
</tr>
</tbody>
</table>

- RTM (ϕ) |
- LDA + Regression |
Dataset: council notes from Merovingian Gaul

Example datum:

King Alberic – Archdeacon Francis – Positive

For each character:
- Role
- Place
- Relations
- Other information
Random Subgraph Model

\[ \alpha \quad \eta \quad S \quad \gamma \quad Z_i \quad Z_j \quad a_{ij} \quad X_{ij} \quad \Pi \quad \Xi \quad \mu, \nu \]
Random Subgraph Model

- Model presence and type of relation

- Only type depends on clusters

- Cluster probabilities depend on subgraph