Generative Model

- Two stages: model definition and inference
- Model definition:
  - Define a model mimicking the real world
  - "tell your story in a probabilistic way"
- Inference
  - Infer unknown but interesting variables based on evidence/observed data

Example: life of a Tufts student

- A student either keeps an early schedule or late schedule
- Students with an early schedule arrives at campus at $T_0$ within a range of 1 hour
- Students with an late schedule get up at $T_1$ within a range of 1 hour

Formal definition: each student $n$ has schedule $s_n \in \{0, 1\}$ and arrival time $t_n$

$s_n \sim Bernoulli(0.5)$
$t_n|s_n = 1 \sim Normal(T_0, 1)$
$t_n|s_n = 1 \sim Normal(T_1, 1)$

The data:

<table>
<thead>
<tr>
<th>Student</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arrival time</td>
<td>6.8</td>
<td>7.3</td>
<td>9.2</td>
<td>8.0</td>
<td>7.7</td>
<td>8.8</td>
<td>7.1</td>
<td>9.5</td>
<td></td>
</tr>
</tbody>
</table>

Questions:
- What is the value of $T_0$ and $T_1$?
- Does student 5 take an early schedule?

Example: a grading problem

- The exam has 50 questions
- The class has 100 students taking the exam
- The professor collect students' answers
- The problem: rate students' performances without correct answers of the 50 questions

Define a generative model

- Specify random variables
- Define conditional probabilities
Example: a grading problem

Define variables:
- Problems with binary answers \( \{ t_m \in \{0,1\}, m = 1, \ldots, M \} \)
- Student’s performance \( \{ s_n \in \{0,1\}, n = 1, \ldots, N \} \)
- Student’s answers \( \{ a_{nm} \in \{0,1\}, n = 1, \ldots, N, m = 1, \ldots, M \} \)

Example: a grading problem

Define distributions:
- For each question \( m \), choose a true answer \( t_m \sim \text{Bernoulli}(\mu) \).
- For each student \( n \), choose a score \( s_n \sim \text{Uniform}(0,1) \).
- For each question \( m \) and each student \( n \), choose an answer \( a_{nm} \sim \text{Bernoulli} \left( (s_n)^m (1 - s_n)^{(1-t_m)} \right) \).

The joint probability

- The joint probability \( p(t, s, a) = p(t)p(s)p(a|s, t) \).
- What is the performance \( s_n \) of the student \( n \) given my observations?
  - Calculate \( p(s_n|a) \).
  - This is inference about!

Define a model

The real world

The model

- The model mimics the real world.
- Keeps important relations.
- Neglects unimportant relations.
- And folds uncertainties into distributions.

Terminology

- The model
  - Define variables and their relations.
  - Simulate the real world at correct level.
- Variables
  - Observed \( X \), hidden \( Z \).
  - Represent things/facts that exist in the real world.
- Parameter \( \theta \)
  - Not exist in the real world.
  - Tuned to get good fidelity of the model.

The general form

- The general form
  - Probability: \( p(X, Z; \theta) \).
  - Parameters: \( 0.5 \) in Bernoulli dist.
  - Observed variables: students’ answers \( a \).
  - Hidden variables: problem answers \( t \) and students’ performances \( s \).
Parameter Learning

- Maximum Likelihood Estimation (MLE)
  - Maximize the data likelihood of $X$
  - EM algorithm (talk later)

$$\theta^* = \arg \max \log p(X, Z; \theta)$$

The inference problem

- Calculate the posterior of some interesting variable, e.g. $Z_i$
  $$p(Z_i | X; \theta^*)$$
- Many inference packages can do the calculation approximately

## Topic Modeling

**Generation of documents $w$ in a corpus $D$**

- $w = \{w_n\}$ is a list of words
- $z_n$ is the topic of each word
- $\theta$ is the topic component of document $w$
- $\beta$ is a probability matrix. Each column $\beta_j$ is the distribution of words in topic $j$

From “Latent Dirichlet Allocation” [Blei et al., 2003]:

1. Choose $N \sim \text{Poisson}(\kappa)$.
2. Choose $\theta \sim \text{Dir}(\alpha)$.
3. For each of the $N$ words $w_n$:
   a. Choose a topic $z_n \sim \text{Multinomial(} \theta)$.
   b. Choose a word $w_n$ from $p(w_n | z_n, \beta)$, a multinomial probability conditioned on the topic $z_n$.

**Discussion**

- What variables are hidden and what are observed?
- How to write the joint?
- Which probability to calculate the infer the topic component of a document?
- What is the distribution of words in the fifth topic?

**EM algorithm**

- MLE for models with hidden variables
  - Soft clustering / example of arrival time
    $$\theta^* = \arg \max \Sigma_n \log \Sigma_{s_n=0,1} p(t_n, s_n; \theta)$$
  - log cannot get into summation
  - Maximize a lower bound with Jensen’s inequality (log get into summation now)
    $$\log \Sigma_{s_n=0,1} p(t_n, s_n; \theta) \geq \Sigma_n q(s_n) \log p(t_n, s_n; \theta) - \Sigma_n q(s_n) \log p(s_n)$$
**EM algorithm**

- MLE for models with hidden variables
  - Soft clustering / example of arrival time
  \[ \theta^* = \arg \max \sum_n \log \sum_{s_n=0,1} p(t_n, s_n; \theta) \]
  - \( \log \) cannot get into summation
  - Maximize a lower bound with Jensen’s inequality (\( \log \) get into summation now!)
  \[ \log \sum_{s_n=0,1} p(t_n, s_n; \theta) \geq \sum_{s_n} q(s_n) \log p(t_n, s_n; \theta) - \sum_{s_n} q(s_n) \log p(s_n) \]

**EM lower bound**

\[
\log \sum_{s_n=0,1} p(t_n, s_n; \theta) \\
\geq \sum_{s_n} q(s_n) \log p(t_n, s_n; \theta) - \sum_{s_n} q(s_n) \log p(s_n)
\]

- The bound is tight (The lower bound is maximized) when
  \[ q(s_n) = \frac{p(s_n|t_n; \theta)}{p(s_n,t_n; \theta)} = \frac{p(s_n = 0|t_n, \theta) + p(s_n = 1|t_n, \theta)}{p(s_n = 0|t_n, \theta) + p(s_n = 1|t_n, \theta)} \]

**Maximize the lower bound**

Push up the lower bound to be tight; objective do not move

**Understand EM in soft clustering model**

- In M step,
  \[ T_\theta = \arg \max \sum_n \sum_{s_n} q(s_n) \log p(t_n, s_n; T_\theta) \]
  \[ = \arg \max \sum_n Q(s_n = 0) \log p(t_n, s_n = 0; T_\theta) + \log p(s_n = 0) \]
  \[ = \arg \max \sum_n Q(s_n = 0) (t_n - T_\theta)^2 + \text{constant} \]

\( T_\theta \) is the center of instances weighted by cluster memberships

**Generative Model**

- Identify interesting variables, including hidden ones \( Z \) and observed variables \( X \)
- Define joint probability \( p(X, Z; \theta) \) by define conditionals
- Learn parameter \( \theta \) if necessary
- Infer \( Z \) with observations \( X \)
Generative Model

- Powerful modeling tool
  - Can treat almost any data with appropriate model assumption
- Further topics: probabilistic programming