The Information Bottleneck Method

N. Tishby, F. Pereira, and W. Bialek
Proc. 37th Allerton Conf. on Communication and Computation
1999

Presented by
Jun Wang
Feb 12, 2003
Basic concepts

- **Entropy** -- measures the average (i.e., expected) uncertainty about a random variable
  \[
  H(X) = -\sum_x p(x) \log p(x)
  \]

- **Conditional Entropy** -- measures the amount of remaining uncertainty about $X$ after observing $Y$
  \[
  H(X \mid Y) = \sum_y p(y) H(X \mid Y = y)
  \]

- **Relative Entropy** (or Kullback-Leibler distance)
  -- measures the distance between two distributions
  \[
  D_{KL}[p(x) \mid \mid q(x)] = \sum_x p(x) \log \frac{p(x)}{q(x)}
  \]
Mutual Information

The "Mutual Information" of two random variables $X$ and $Y$ measures the amount of uncertainty about $X$ that is reduced by observing $Y$

$$I(X, Y) = H(X) - H(X \mid Y)$$

$$= \sum_{x,y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$

$$I(X, Y) = I(Y, X)$$
### A Simple Example...

From Tishby NIPS2001 Workshop (see references)

<table>
<thead>
<tr>
<th>Doc-Word occurrences</th>
<th>Israel</th>
<th>Health</th>
<th>www</th>
<th>Drug</th>
<th>Jewish</th>
<th>Dos</th>
<th>Doctor</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc1</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Doc2</td>
<td>0</td>
<td>9</td>
<td>2</td>
<td>11</td>
<td>1</td>
<td>0</td>
<td>6</td>
<td>...</td>
</tr>
<tr>
<td>Doc3</td>
<td>0</td>
<td>10</td>
<td>1</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>20</td>
<td>...</td>
</tr>
<tr>
<td>Doc4</td>
<td>9</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>Doc5</td>
<td>0</td>
<td>3</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>10</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>Doc6</td>
<td>1</td>
<td>11</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>1</td>
<td>7</td>
<td>...</td>
</tr>
<tr>
<td>Doc7</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>0</td>
<td>2</td>
<td>12</td>
<td>2</td>
<td>...</td>
</tr>
<tr>
<td>Doc8</td>
<td>15</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>Doc9</td>
<td>0</td>
<td>12</td>
<td>1</td>
<td>16</td>
<td>0</td>
<td>1</td>
<td>12</td>
<td>...</td>
</tr>
<tr>
<td>Doc10</td>
<td>1</td>
<td>0</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>11</td>
<td>2</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
# Simple Example

After permutation

<table>
<thead>
<tr>
<th></th>
<th>Israel</th>
<th>Jewish</th>
<th>Health</th>
<th>Drug</th>
<th>Doctor</th>
<th>www</th>
<th>Dos</th>
<th>....</th>
</tr>
</thead>
<tbody>
<tr>
<td>Doc1</td>
<td>12</td>
<td>8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>Doc4</td>
<td>9</td>
<td>7</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>Doc8</td>
<td>15</td>
<td>10</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>Doc2</td>
<td>0</td>
<td>1</td>
<td>9</td>
<td>11</td>
<td>6</td>
<td>2</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>Doc3</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>6</td>
<td>20</td>
<td>1</td>
<td>0</td>
<td>...</td>
</tr>
<tr>
<td>Doc6</td>
<td>1</td>
<td>0</td>
<td>11</td>
<td>6</td>
<td>7</td>
<td>0</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>Doc9</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>16</td>
<td>12</td>
<td>1</td>
<td>1</td>
<td>...</td>
</tr>
<tr>
<td>Doc5</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>10</td>
<td>...</td>
</tr>
<tr>
<td>Doc7</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>8</td>
<td>12</td>
<td>...</td>
</tr>
<tr>
<td>Doc10</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>9</td>
<td>11</td>
<td>...</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A new compact representation

<table>
<thead>
<tr>
<th>Cluster1</th>
<th>36</th>
<th>25</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>0</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster2</td>
<td>1</td>
<td>1</td>
<td>42</td>
<td>39</td>
<td>45</td>
<td>4</td>
<td>2</td>
<td>...</td>
</tr>
<tr>
<td>Cluster3</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>0</td>
<td>4</td>
<td>26</td>
<td>33</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

The document clusters preserve the relevant information between the documents and words
$X_1 \quad X_2 \quad \ldots \quad X_n$

$I(X, \hat{X}) \quad I(\hat{X}, Y)$

$\hat{x}_1 \quad \hat{x}_2 \quad \ldots \quad \hat{x}_k$

$y_1 \quad y_2 \quad \ldots \quad y_m$
Information Bottleneck problem

- Given joint distribution $p(x, y)$, and let $|\hat{X}|$ be fixed (usually $|\hat{X}| \ll |X|$)

- Find conditional prob. distribution $p(\hat{x} | x)$ (probabilistic mapping, or soft clustering), s.t.,
  - $I(X, \hat{X})$ is minimized
    - compress $X$ as much as possible
  - $I(\hat{X}, Y)$ is maximized
    - preserve information about $Y$ as much as possible
Balance two objectives

- We have two objectives:
  \[ I(X, \hat{X}) \quad \text{and} \quad I(\hat{X}, Y) \]

- Using Lagrangian multiplier

  \[ L[p(\hat{x} \mid x)] = I(X, \hat{X}) - \beta I(\hat{X}, Y) \]

  - where \( \beta \) is the Lagrangian multiplier, playing the role of trade-off.
Find \( p(\hat{x} \mid x) \), s.t. \( L[p(\hat{x} \mid x)] \) is optimal

Solve \( \frac{\partial L[p(\hat{x} \mid x)]}{\partial p(\hat{x} \mid x)} = 0 \)

\[ p(\hat{x} \mid x) = \frac{p(\hat{x})}{Z(x, \beta)} e^{-\beta D_{KL}[p(y|x) \parallel p(y|\hat{x})]} \] (1)

- Where \( Z(x, \beta) \) is a normalization function
- and \( D_{KL} \) is the Kullback-Leibler distance
Two more equations

\[ p(\hat{x} \mid x) = \frac{p(\hat{x})}{Z(x, \beta)} e^{-\beta D_{KL}[p(y \mid x) \parallel p(y \mid \hat{x})]} \]  \hspace{1cm} (1)

We also need equations for \( p(\hat{x}) \) and \( p(y \mid \hat{x}) \)

\[ p(\hat{x}) = \sum_x p(\hat{x} \mid x) p(x) \]  \hspace{1cm} (2)

\[ p(y \mid \hat{x}) = \frac{p(\hat{x} \mid y) p(y)}{p(\hat{x})} = \frac{1}{p(\hat{x})} \sum_x p(\hat{x} \mid x) p(y \mid x) p(x) \]  \hspace{1cm} (3)
Solve the above Equations: (1),(2),(3)

IB iterative algorithm (like EM)

Denote by $t$ the iteration step

\[
\begin{aligned}
    p_{t+1}(\hat{x} | x) &= \frac{p_t(\hat{x})}{Z_t(x, \beta)} e^{-\beta D_{KL}[p(y|x)\|p_t(y|\hat{x})]} \\
    p_{t+1}(\hat{x}) &= \sum_x p(x) p_t(\hat{x} | x) \\
    p_{t+1}(y | \hat{x}) &= \frac{1}{p_t(\hat{x})} \sum_x p(y | x) p_t(\hat{x} | x) p(x)
\end{aligned}
\]

Convergence can be proven (see the paper).
In the previous algorithm, we assume we know:

- $|\hat{X}|$ is fixed
- $\beta$ is fixed

How to find the best $|\hat{X}|$ and $\beta$?

- Using deterministic annealing approach.
Information Bottleneck (IB) vs. Rate Distortion Theory (RDT)

- IB is a natural generalization of RDT with similar convergence and optimality proofs.
  - RDT: only considers $X$ and $\hat{X}$
  - IB: also takes $Y$ into consideration.
Information Bottleneck VS. Neural Nets

- Auto association: forcing compact representations
- $\hat{X}$ is a relevant code of $X$ w.r.t. $Y$

From Tishby NIPS2001 Workshop (see references)
Recent Progresses

- Multivariate IB
- And many applications
  - document clustering, classification
  - bioinformatics
  - spectral analysis
  - ...
References