A brief introduction to uses of information theory in machine learning

Rate Distortion Theory, Deterministic Annealing, and Soft K-means.
The Information Bottleneck Method.
Overall goal

- Physical systems execute computations and produce information.

- This information is reflected in observations of the system.

- We make use of this information when we make a model of a system.

- Try to understand this process and to automate it.
Overall goal

- The observer (= learner) determines to some extend the data that it has at hand by deciding which measurements to take.

- Furthermore, actions that the learner takes may change the underlying process.

- Pay special attention to this feedback!
Lectures

1. Background: Rate distortion theory and related methods.
2. Predictive filtering and inference.
3. Interactive learning.
Goal of model making

- Want to find a “good” representation of the observed data.
- Need to define what we mean by “good”.
- Should capture the relevant aspects of the observations and filter out the noise.
- Should not be more complicated than necessary → find an efficient summary.
Approaches

- Supervised Learning if labeling of the data is available.
- Otherwise unsupervised Learning, cluster analysis:
  - Colloquially, the goal is to group data such that similar objects end up in one cluster.
  - Need a definition of similarity.
  - Important to realize that this is a form of lossy compression!
Lossy Compression

- Summarize data by keeping only relevant information and throwing away irrelevant information.

- Need a measure for information $\rightarrow$ Shannon's mutual information.

- Need a notion of relevance.
Relevance via distortion

Shannon (1948): Define a function that measures the distortion between the original signal and its compressed representation.

Note: This is related to the similarity measure in unsupervised learning/cluster analysis.
Distortion function

- Degree of freedom: which function to use is up to the experimenter.

- It is not always obvious what function should be used, especially if the data do not live in a metric space, and so there is no “natural” measure.

- Example: Speech.
Relevant information

Tishby, Pereira, Bialek (1999): Measure relevance directly via Shannon’s mutual information by defining:

Relevant information = the information about a variable of interest.

Then, there is no need to define a distortion function ad hoc, and the appropriate similarity measure arises naturally.
Learning and lossy data compression

When we build a model of a data set, we map observations to a representation that summarizes the data in an efficient way.

Example: K-means. Map N data points to K clusters, with centroids c. If K << N, we get a substantial compression!
Learning and lossy data compression

Entropy can be used to measure the compactness of the model. Sometimes called “statistical complexity” (J. P. Crutchfield.)

Good measure of complexity if we are searching for a deterministic map (“hard partition”).

In general, however, we may search over probabilistic assignments (“soft partition”). Then, we need a complexity measure that accounts for the “fuzziness” of the map. Use mutual information.
Rate distortion theory

- Shannon (1948)

- Find assignments $p(c|x)$ from data $x \in X$ to clusters $c = 1, \ldots, K$, and find class-representatives $x_c$ ("cluster centers"), such that the average distortion $D = \langle d(x, x_c) \rangle$ is small.

- Compress the data by summarizing it efficiently in terms of bit-cost: Minimize the coding rate, the information which the clusters retain about the raw data.

$$I = \langle \frac{p(x, c)}{p(x)p(c)} \rangle$$
Constrained optimization

\[
\begin{align*}
\min_{p(c|x)} \quad & \left[ I(x, c) + \beta \langle d(x, x_c) \rangle \right] \\
\text{Solution:} \quad & p(c|x) = \frac{p(c)}{Z(x, \beta)} \exp \left[ -\beta d(x, x_c) \right] \\
\langle \frac{d}{dx_c} d(x, x_c) \rangle_{p(x|c)} = 0 \\
\text{(centroid condition)}
\end{align*}
\]
Rate distortion curve

Family of optimal solutions, one for each value of the Lagrange multiplier beta. This parameter controls the trade-off between compression and fidelity.

\[
p(c|x) = \frac{p(c)}{Z(x, \beta)} \exp \left[ -\beta d(x, x_c) \right]
\]

Evaluate objective function at the optimum for each value of beta and plot I vs D. => Rate-distortion curve.
Remarks

- for squared error distortion, \( d = (x - x_c)^2 / 2 \), the centroid condition reduces to

\[
x_c = \langle x \rangle p(x|c)
\]

- "soft K-means" because assignments can be probabilistic (fuzzy):

\[
p(c|x) = \frac{p(c)}{Z(x, \beta)} \exp \left[ -\beta d(x, x_c) \right]
\]
in the zero temperature limit, $\beta \to \infty$, we have deterministic (or "hard") assignments because

$$c^* := \arg\min_c d(x, x_c); \quad D(x, c) := d(x, x_c) - d(x, x_{c^*}) > 0$$

$$p(c^* | x) = \frac{\exp(-\beta d(x, x_{c^*}))}{\sum_c \exp(-\beta d(x, x_c))} = (1 + \sum_{c \neq c^*} \exp(-\beta D(x, c)))^{-1} \to 1$$

the analogy to thermodynamics inspired Deterministic Annealing (Rose, 1990)
Soft K-means algorithm

- Choose a distortion measure.
- Fix the “temperature”, $T$, to a very large value (corresponds to small $\beta = 1/T$)
- Solve iteratively, until convergence:

Assignments: \[ p(c|x) = \frac{p(c)}{Z(x, \beta)} \exp[-\beta d(x, x_c)] \]

Centroids: \[ \langle \frac{d}{dx_c} d(x, x_c) \rangle_{p(x|c)} = 0 \]

- Lower temperature: $T \leftarrow aT$, and repeat.
- $a = “annealing rate”, a small, positive number.$
How to choose the distortion function?

Example: Speech

Cluster speech such that signals which encode the same word will group together.

May be extremely difficult to define a distortion function which achieves this.

Intelligibility criterion: should measure how well the meaning is preserved. Explicit form?
Information Bottleneck Method

- Tishby, Pereira, Bialek, 1999

- Instead of guessing a distortion function, define relevant information as information the data carries about a quantity of interest (Example: phonemes or words.)

- Data is compressed such that relevant information is kept maximally.

\[
\begin{align*}
\min_{c} I(x, c) & \quad \rightarrow \quad y \\
\max_{c} I(c, y) & \quad \rightarrow \quad x
\end{align*}
\]
Constrained optimization

\[
\max_{p(c|x)} \left[ I(y, c) - \lambda I(x, c) \right]
\]

Optimal assignment rule

\[
p(c|x) = \frac{p(c)}{Z(x, \lambda)} \exp \left( -\frac{1}{\lambda} D_{KL}[p(y|x)\|p(y|c)] \right)
\]

Kullback-Leibler divergence emerges as the distortion function:

\[
D_{KL}[p(y|x)\|p(y|c)] = \sum_y p(y|x) \log_2 \left[ \frac{p(y|x)}{p(y|c)} \right]
\]
Inference?

So far, we have assumed that the statistics of the underlying process, $p(x,y)$, are known!

The method produces a relevant and efficient summary: we do not want to keep all of the information! How much we keep is up to us (choice of $\lambda$).

If $p(x,y)$ is not known, but estimated from the data, then sampling errors set an upper bound on how much information we can keep without over-fitting (minimum value of $\lambda$).

This can be estimated using perturbation theory: S. Still and W. Bialek, 2004.