Juliette + Chris CS 109

Mutual Information from an Ideal Classifier

Suppose we have data where Y is a **medical record** and X is a **summary** that was written from that record. We want to measure how much information the summary X contains about the medical record Y. A natural quantity for this is the *mutual information*

$$\text{MutualInfo}(X;Y) := \sum_x \sum_y P(x,y)\,\log\left(\frac{P(y \mid x)}{P(y)}\right).$$

This section shows that if we assume we have access to an *ideal classifier* that identifies which medical record corresponds to a given summary, then the classifiers average log-probability of being correct forms a good approximation to this mutual information.

1. **A Special Classifier**

   For each summary X, we construct a list of N candidate medical records:

   $$(Y_1, Y_2, \ldots, Y_N).$$

   Exactly one of these (call it Y_I) is the *true* medical record. The other $N-1$ candidates are randomly drawn distractors from the marginal $P(Y)$, independently of X.

   The index $I \in \{1, \ldots, N\}$ is the correct answer. A classifier must output a probability distribution

   $$P(I = 1 \mid x, y_1, \ldots, y_N), \quad \ldots, \quad P(I = N \mid x, y_1, \ldots, y_N).$$

2. **What an Ideal Classifier Looks Like**

   We compute the true conditional probability $P(I = i \mid x, y_1, \ldots, y_N)$. Using conditional probability,

   $$P(I = i \mid x, y_1, \ldots, y_N) = \frac{P(x, y_1, \ldots, y_N, i)}{P(x, y_1, \ldots, y_N)}.$$

   Because the distractors Y_j for $j \neq i$ are drawn independently from $P(Y)$ and independently of X,

   $$P(x, y_1, \ldots, y_N, i) = \frac{1}{N}\,P(x, y_i)\,\prod_{j \neq i} P(y_j).$$

   The denominator is a sum over all possible positions of the true record:

   $$P(x, y_1, \ldots, y_N) = \sum_{k=1}^{N} \frac{1}{N}\,P(x, y_k)\,\prod_{j \neq k} P(y_j).$$

Canceling the common factors $\frac{1}{N}\prod_j P(y_j)$,

$$P(I = i \mid x, y_1, \ldots, y_N) = \frac{P(x, y_i)/P(y_i)}{\sum_{k=1}^{N} P(x, y_k)/P(y_k)} = \frac{P(y_i \mid x)P(y_i)}{\sum_{k=1}^{N} P(y_k \mid x)/P(y_k)}$$

3. **Expected Log Likelihood**

If we have a big enough dataset, the natural MLE-style score is the *average* log-probability weighted by the probability of each datapoint:

$$\begin{aligned}
LL_{\text{exp}} &= \sum_{x,y_1,\ldots,y_N,i} P(x, y_1, \ldots, y_N, i)\ \log P(I = i \mid x, y_1, \ldots, y_N) \\
&= \sum_{x,y_1,\ldots,y_N,i} P(x, y_1, \ldots, y_N, i)\ \log\left(\frac{\frac{P(y_i|x)}{P(y_i)}}{\sum_{k=1}^{N} \frac{P(y_k|x)}{P(y_k)}}\right) \\
&= \underbrace{\sum_{x,y} P(x, y)\ \log\left(\frac{P(y \mid x)}{P(y)}\right)}_{A} - \underbrace{\sum_{x,y_1,\ldots,y_N} P(x, y_1, \ldots, y_N)\ \log\left(\sum_{k=1}^{N} \frac{P(y_k \mid x)}{P(y_k)}\right)}_{B}.
\end{aligned}$$

Here the term involving the true index i and the sampled distractors reduces to $P(x, y)$ because, under the sampling process, the probability that the true pair (x, y) appears in position i within $(y_1, \ldots, y_N)$ is exactly $P(x, y)$, independent of which index i holds the true record.

Part A. The first term is exactly the mutual information: $A = \text{MutualInfo}(X; Y)$

Part B. By Jensens inequality (not part of cs109) applied to the concave function log, $B \leq \log N$

Thus $LL_{\text{exp}} \ \geq\ \text{MutualInfo}(X; Y) - \log N$

4. **Just a Sample Perspective**

For a single example where $p^\star$ is the probability assigned to the correct index $\star$, x is the summary and $y^\star$ is the corresponding medical record:

$$\begin{aligned}
\log p^\star &= \log\left(\frac{P(y^\star \mid x)}{P(y^\star)}\right) - \log\left(\sum_{k=1}^{N} \frac{P(y_k \mid x)}{P(y_k)}\right) \\
&= \text{PointwiseMutualInfo}(x, y^\star) - \log\left(\sum_{k=1}^{N} e^{\text{PMI}(x, y_k)}\right) \\
&\geq\ \text{PointwiseMutualInfo}(x, y^\star) - \log N.
\end{aligned}$$