Bayes Networks and Probalistic Programming

Adam Massmann

Water Center NN Meetings

Week 5: Oct. 17th, 2017

Introduction - Information Theory (see section 1.6 in Bishop)

- Consider the amount of information gained/learned by an event or observation (we'll call x). We would receive more new information from a very surprising event than an event we expect (because we already know something about the expected event).
- So if we want to quantify this "amount of information" contained in an event we should use a function of the probability of the event ($p(x)$). The amount of information we'll call a function $h(\cdot)$, which will be a function of $p(x)$.

Guidance for the functional from of $h(\cdot)$

- If two events x and y are independent, then the amount of information gained by both events should be $h(x, y) = h(x) + h(y)$.
- We also know that the joint probability of x and y's occurrence would be: $p(x, y) = p(x)\, p(y)$.
- So the question is, what function $\hat{h}$ satisfies: $h(x, y) = \hat{h}(p(x)\, p(y)) = \hat{h}(p(x)) + \hat{h}(p(y))$?

Information Entropy

- $\hat{h}(\cdot) = \log(\cdot)$ satisfies $h(x, y) = \hat{h}(p(x)\ p(y)) = \hat{h}(p(x)) + \hat{h}(p(y))$, so $h(\cdot) = \log(p(\cdot))$.
- It's desirable for h to be positive, so because $0 \leq p \leq 1$, lets make it $h(\cdot) = -\log p(\cdot)$.
- Now say we have a bunch of random variables x for which we want to know the average amount of information (i.e. expectation of $h(x)$). This would be given by:

$$H[x] = -\sum_x p(x) \log p(x)$$

- This is known as the *entropy* of x.
- Extending this to continuous variables gives the *differential entropy*:

$$H[x] = -\int p(x)\ \log p(x)\, dx$$

[Bishop 2006, Shannon 1948]

So what does information entropy look like?

- From thermodynamics and statistical mechanics we have some idea of entropy as a measure of the disorder or randomness in a system. For information theory it is similar.[1]

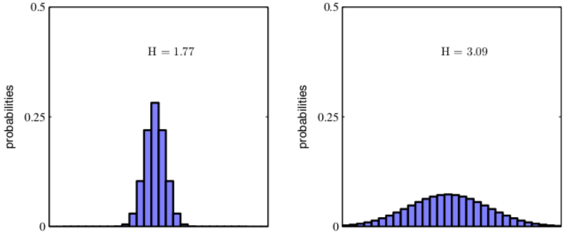

Figure 1: From Bishop 2006: Histograms of two probability distributions over thirty bins illustrating the higher value of entropy H for the broader distribution. The largest entropy would arise from a uniform distribution that would give $H = -\ln 1/30 = 3.40$

[1] von Neumann told Shannon he should also call it entropy because "nobody knows what entropy really is, so in any discussion you will always have an advantage."

Kullback-Leibler divergence

- Why should we even care about information theory or entropy?
  - Because we can use entropy ideas to perfrom inference on a probalistic model (e.g. a Bayes network), given data.
  - Say we have some phenomenon with a true probability distribution $p(x)$, which we are approximating with some [possibly parametric] distibution $q(x)$.
  - Then the additional necesary information required to communicate the value of x as consequence of using $q(x)$ would be:

$$
\begin{aligned}
KL(p||q) &= -\int p(x) \ln q(x) dx - \left(-\int p(x) \ln p(x) dx \right) \\
&= -\int p(x) \ln \frac{q(x)}{p(x)}
\end{aligned} \tag{1}
$$

This is known as relative entropy or Kullback-Leibler (KL) divergence (Kullback and Leibler 1951).

Properties of Kullback-Leibler divergence

- Note that it is not a symmetrical quantity (e.g. $KL(p||q) \neq KL(q||p)$).
- Also, $KL(p||q) \geq 0$,
- and $KL(p||q) = 0$ only if p and q are identical (see Bishop 2006 for proof).
- So practically speaking KL divergence is very useful as a cost function quantifying the similarity between two probability distributions.

Using KL divergence for inference

- Say we have N observations of data x_n from some unknown probaility distribution $p(x)$.
- We want to try to approximate $p(x)$ with a parametric disribution $q(x|\theta)$, by minimizing the KL diverence which can be approximated by:

$$KL(p||q) \simeq \frac{1}{N} \sum_{n=1}^{N} [-\ln q(x_n|\theta) + ln(p(x_n))] \tag{2}$$

- The second term is not a function of θ, and the first term is just the negative log liklihood. So for this example, minimizing KL-divergence is the same as minimizing the negative log liklihood, which we saw in Week 1!

References

See section 1.6 in Bishop

Bishop, Christopher M (2006). *Pattern recognition and machine learning*. springer.

Kullback, Solomon and Richard A Leibler (1951). "On information and sufficiency". In: *The annals of mathematical statistics* 22.1, pp. 79–86.

Shannon, Claude E (1948). "A mathematical theory of communication, Part I, Part II". In: *Bell Syst. Tech. J.* 27, pp. 623–656.