How to mix org and bib for research paper management

Philip Yang

April 11, 2014

Abstract

We demonstrate how to write a bibliography-enabled draft. This is done with the help of *biblatex*.

1 A major section

- clicking on a `note:some_paper` link in an org-mode document will jump to the corresponding bibliographic note about a particular paper.
- clicking on a `bib:some_paper` link in an org-mode document will jump to the corresponding bibliographic reference in the bibtex file.
- exporting an org document containing either of the above links to LaTeX will produce correct references cite{abibref} LaTeX code (see the results here : draft.pdf).

  So here is the solution. We process the link structure, but we don't really rely on it.

We'll also add into section all the notes relating to these articles. These notes will be identified by `CUSTOM_ID` properties which will contain the bibliographic reference of the papers.

Here is a citation of [4, Vishwanathan et al.]. Clicking the link in orgmode will give you our note on that paper, but the html / latex output will be just the citation.

[2, Hausser et al.] provides a way to estimate entropy for discrete distributions.

[3, Paninski et al.] provides ways to generate stuffs.

To simply add a citation here, we could use the *bibtex* syntax. We also want to thank Crutchfield *et al.* [1]

2 A section with formulae

We can also define mutual information of multiple variables.

$$\begin{aligned} I(X;Y;Z) &= \mathbb{E}_{X,Y,Z}\left[\log(\frac{p_M(X,Y,Z)}{p_M(X)p_M(Y)p_M(Z)})\right] \\ &= H(X) + H(Y) + H(Z) - H(X,Y,Z) \\ &= H(X) + H(Y) - H(X,Y) + H(X,Y) + H(Z) - H(X,Y,Z) \\ &= I(X;Y) + I(X,Y;Z). \end{aligned}$$

For a normal distribution, let $C = \frac{\log(2\pi)+1}{2}$

$$\begin{aligned} I(X,Y) &= H(X) + H(Y) - H(X,Y) \\ &= C + \log(|\Sigma_X|)/2 + C + \log(|\Sigma_Y|)/2 - 2C - \log(|\Sigma_{X,Y}|)/2 \\ &= -\frac{1}{2}\log(\frac{|\Sigma_{X,Y}|}{|\Sigma_X\Sigma_Y|}) \\ &= -\frac{1}{2}\log(1-\rho^2_{X,Y}). \end{aligned}$$

This shows that mutual information for multivariate normal distribution is a monotone transform of the correlation.

3 A list of stuffs

Kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{F}$ where $\mathbb{F}$ is a field, usually either the real field $\mathbb{R}$ or the complex field $\mathbb{C}$.

Kernel is used to define "similarity" between elements of $\mathcal{X}$, thus we expect "similar" elements to have higher value. There are some properties that k must satisfy:

- $k(x,y) = \overline{k(y,x)}$
- $k(x,x) \geq 0$

Kernel density estimate the first step to data analysis is usally done through looking at the histogram. Yet the bin size we choose might

affect the result. For many cases (especially when our data is real valued), we have no idea to determine it a priori.

Hilbert space complete space with an inner product. the decisive deature of Hilbert space is the inner product. In most cases, we use Hilbert space to study functions, to be specific, square integrable functions f with $\int_{-\infty}^{\infty} f\bar{f} < \infty$.

Complete space a metric space where every Cauchy sequence converges in it. The space of rational numbers is not complete, since we can construct a Cauchy sequence that converges to $\sqrt{2} \notin \mathbb{Q}$.

Cauchy sequence $\forall \delta > 0, \exists N, \forall n, m > N, \|a_n - a_m\| < \delta$. as for the case of a complete sequence, every Cauchy sequence converges, and the value it converges to is also in that space. As for example, $\mathbb{Q}$ is not a complete space since we can construct a Cauchy sequence which converges to some irrational number. In fact, this is pretty much how we construct the entire real line with Dedekind cut.

Inner product This definition is broader and more general than dot product and it's the right term to use whenever you are not sure. An inner product $\langle \cdot, \cdot \rangle : \mathcal{X} \times \mathcal{X} \to \mathbb{F}$ is a linear structure with the following properties.

- $\langle x, y \rangle = \overline{\langle y, x \rangle}$
- $\langle ax, y \rangle = a\langle x, y \rangle$
- $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$
- $\langle x, x \rangle \geq 0$ and the equaltity holds iff $x = 0$

These properties entails the famous Cauchy-Schwartz inequality

$$\langle x, y \rangle^2 \leq \langle x, x \rangle \langle y, y \rangle.$$

The inner product defined by the Hilbert space also induces a norm, $\|x\|_{\mathcal{H}} = \langle x, x \rangle^{1/2}$.

Metric in contrast to kernel, a metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is used to characterize "distance" between two elements in $\mathcal{X}$. properties include:

- $d(x, y) \geq 0$, where equality holds iff $x = y$
- $d(x, y) = d(y, x)$
- $d(x, y) \leq d(x, z) + d(y, z)$

We can define $d(x, y) = \|x - y\|_{\mathcal{H}}$ and verify that it is truly a metric.

Positive definite kernel with this property the kernel distance is a metric ... orz; any p.d. kernels induces a Hilbert space

Kernel distance when we compute the distance of two sets $\mathcal{P}, \mathcal{Q}$ given only the definition of similarity between $p \in \mathcal{P}$ and $q \in \mathcal{Q}$ via kernel $k(p, q)$,

Reproducing kernel Hilbert space (RKHS) a Hilbert space whose inner product "reproduces" the kernel. Almost all the properties of a Hilbert space is captured in its inner product (this could be seen from the isomorphism theorem). To be specific, given a kernel k, we want $\langle k(\cdot, x), k(\cdot, y)\rangle = k(x, y)$

Bochner's theorem It states that for any shift-invariant positive definite kernel k(z), its Fourier transform p(w) is a measure. We know that modern axiomatic probability theory is based upon measures. If we properly scale p(w) so that p(E) = 1 where E is the ground set, then p(w) is a probability (density function). In our case we wanted to find a finite feature space whose inner product $\langle \phi(x), \phi(y)\rangle$ approximates k(x, y). This is the same idea behind RKHS. Lucky for us, $k(x, y) = \int p(w) \exp(iw'(x - y))dw$, which then can be seen as an expectation $\mathbb{E}_w[\exp(iw'(x - y))]$. Thus sampling (or Monte Carlo algorithm) will suffice. This estimate should be more accurate in estimating kernel distance thanks to the concentration of measure. Chernoff-Hoeffding bound is typically used to prove it. If the kernel is epkov like, we know how to sample from it. Thus the connection via Fourier transform could go both ways.

Mercer's theorem the symmetric kernels has an eigen expansion where the eigenfunctions are in the sense of Hilbert-Schmidt integral operator $\mathcal{T}_k$.

$$(\mathcal{T}_k f)(x) = \int_\Omega k(x, z) f(z) \rho(z) dz$$

where our kernel could be rewritten as

$$k(x, y) = \sum_{n=0}^{\infty} \lambda_n \psi_n(x) \psi_n(y).$$

These eigen functions are orthonormal. So every function f in our space $\mathcal{H}$ could be rewritten as

$$f(\cdot) = \sum_{n=0}^{\infty} \alpha_n \psi_n(\cdot).$$

Notice that for $k(\cdot, x)$, $\alpha_n = \lambda_n \psi_n(x)$. If we define the inner product in a dot product as

$$\langle f, g \rangle = \sum_{n=0}^{\infty} \frac{a_n b_n}{\lambda_n},$$

then the "reproducing" property of the kernels still holds. This is definitely a convenient, though different from our approach, to look for a finite approximation in Euclidean space (there is only one Euclidean space for a given dimension).

For more details, on how these eigenfunctions look like, check section 6 of *Positive Definite Kernels: Past, Present and Future*. This could provide an alternative way to view the finite dimensional approximation.

Embedding given metric space (X, d) and (X, d') a map $f : X \to X'$ is called an embedding. An embedding is called distance-preserving or *isometric* if for all $x, y \in X$, $d(x, y) = d'(f(x), f(y))$. We call a finite metric (X, d) an l_p-metric if there exists an embedding from X into $\mathbb{R}^k$ for some k such that $\|f(i) - f(j)\|_p = d(x, y)$.

Distortion $\|f\|_{dist} = \max \frac{d(x,y)}{d'(f(x),f(y))} \max \frac{d'(f(z),f(w))}{d(z,w)}$. This measure is invariant under multiplicative factor of the corresponding metrics. The first term is called *contraction* and the second is called *expansion*. Equivalently, distortion is the smallest value $\alpha \geq 1$ such that

$$\forall x, y \; rd(x, y) \leq d'(f(x), f(y)) \leq \alpha r d(x, y).$$

Here r is the aforementioned scaling factor.

4 References with notes

4.1 Graph kernels

<2014-02-12 Wed> (from HSIC paper's reference)

The series regarding 'exotic' kernels. We will be looking at how the kernel machinery could be used in more general scenarios.

The work of the first author is worth going through. At any rate, this is a significant reference.

4.2 Expander graphs and their applications

for the inspiration.

The basics of expander graphs and their applications in both computer science and mathematics.

5 References

[1] James P Crutchfield and Melanie Mitchell. "The evolution of emergent computation". In: *Proceedings of the National Academy of Sciences* 92.23 (1995), pp. 10742–10746.

[2] Jean Hausser and Korbinian Strimmer. "Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks". In: *The Journal of Machine Learning Research* 10 (2009), pp. 1469–1484.

[3] Liam Paninski and Masanao Yajima. "Undersmoothed kernel entropy estimators". In: *Information Theory, IEEE Transactions on* 54.9 (2008), pp. 4384–4388.

[4] SVN Vishwanathan et al. "Graph kernels". In: *The Journal of Machine Learning Research* 11 (2010), pp. 1201–1242.