10-708: Probabilistic Graphical Models 10-708, Spring 2014

22 : Hilbert Space Embeddings of Distributions

Lecturer: Eric P. Xing *Scribes: Sujay Kumar Jauhar and Zhiguang Huo*

1 Introduction and Motivation

The application of classical optimization techniques to Graphical Models has led to specialized derivations of powerful paradigms such as the class of EM algorithms, variational inference, max-margin and maximum entropy learning. This view has also sustained a conceptual bridge between the research communities of Graphical Models, Statistical Physics and Numerical Optimization. The optimization perspective has many advantages, based on a mature and diverse field that allows for problems to be easily formulated, efficiently solved, and approximated in a principled manner via convex relaxations. But it poses several challenges too, which are particularly limiting to problems that involve non-parametric continuous variables or non-convex objectives.

An alternate view, based on principles of linear algebra is a solution to this problem, and the central topic of this lecture. In contrast to the optimization perspective, it offers a basis for dealing with Graphical Models that involve non-Gaussian continuous variables and local minima free learning of latent variable models. It also bridges the theoretical gap between Graphical Models, Kernels in machine learning, and Tensor Algebra. The representational power of this alternative view, however, comes at the cost of expressive intuitiveness.

In the optimization paradigm, parametric distributions in general, and Gaussians in particular, present several modelling advantages. They permit characterization using very simple sufficient statistics. They also allow for easy manipulation on marginalization and conditioning operations that lead to convenient closed-form solutions. An arbitrary distribution is harder to characterize. We could potentially calculate its mean. But many different distributions have the same mean. It might be still better to additionally calculate its variance. While the first two moments completely define a Gaussian distribution, these statistics may not be sufficient for arbitrary distributions. Higher order moments bring increasingly greater resolution power for characterizing arbitrary distributions, which leads to the intuition that an infinite dimensional vector consisting of moments will absolutely capture any distribution.

This is, of course, practically infeasible, since storing or manipulating a vector of infinite dimensions is impossible. It however motivates the use of Hilbert Space embeddings, and the kernel trick to solve this infinite dimensional representation scenario. In particular, the Hilbert Space envisaged for this problem needs to satisfy two basic desiderata: it should create a one-to-one mapping between distributions and the embeddings of their statistics, and it should be cleverly constructed so that the kernel trick can be applied for practical and efficient computation.

In what follows, we introduce Hilbert Spaces and show how distributions can be embedded in them. We also derive several useful linear algebraic techniques to deal with joint as well as conditional distributions. Mathematical detail, as well as intuitive insight are presented as appropriate.

2 Hilbert Spaces

Hilbert spaces are now formally presented and their characteristics and suitability to embedding distributions is discussed.

2.1 Definition of Hilbert Space

A Hilbert space, named after David Hilbert, is an extension of a vector space. Regular vector spaces are sets of objects that are closed under linear combination. That is, given a vector space $\mathcal{V}$, we have $v, w \in \mathcal{V} \implies \alpha v + \beta w \in \mathcal{V}$. While one normally thinks of these objects as finite dimensional vectors, they could potentially be infinite dimensional vectors, and as such should be treated as functions.

A Hilbert Space is a complete vector space equipped with an inner product, which yields a number when input with two functions from the space. An example of an inner product might be: $< f, g >= \int f(x)g(x)dx$. It should be noted that this is just an example, and an inner product need not have an integral.

2.1.1 Basic Properties

The inner product $< f, g >$ in a Hilbert Space must respect the following properties:

1. Symmetry: $< f, g >=< g, f >$
2. Linearity: $< \alpha f_1 + \beta f_2, g >=< \alpha f_1, g > + < \beta f_2, g >$
3. Non-negativity: $< f, f >\geq 0$
4. Zero: $< f, f >= 0 \Rightarrow f = 0$

2.1.2 Operators, Adjoints and the Outer Product

Given this basic definition of a Hilbert Space, we can now define a fundamental concept that is an operator. An operator C maps a function f in one Hilbert Space to another function g in the same or another Hilbert Space. Mathematically this corresponds to:

$$g = Cf$$

This operator has the following property:

$$C(\alpha f + \beta g) = \alpha Cf + \beta Cg$$

As an analogous intuition one can think of functions as vectors and operators as matrices. In linear algebraic terms a matrix typically projects a set of vectors or bases to another set of vectors or bases. Therefore the effect of the operator is to transform a function in a Hilbert Space to another function in another Hilbert Space.

We can similarly define an adjoint (or transpose) of an operator. Formally, the adjoint $C^\top : \mathcal{G} \to \mathcal{F}$ of an operator $C : \mathcal{F} \to \mathcal{G}$ is defined such that the following always holds:

$$< g, Cf >=< C^\top g, f >, \forall f \in \mathcal{F}, g \in \mathcal{G}$$

This is analogous to the transpose or conjugate transpose for real or complex matrices:

$$w^\top M v = (M^\top w) v$$

Finally, also consider the Hilbert Space Outer Product $f \otimes g$, which is implicitly defined such that:

$$f \otimes g(h) =< g, h > f$$

.

Again, as an analogy, consider the vector space outer product which is simply defined by:

$$v w^\top (z) =< w, z > v$$

3 Reproducing Kernel Hilbert Spaces

We now introduce Reproducing Kernel Hilbert Spaces, that are special Hilbert Spaces with even more nice properties. A Reproducing Kernel Hilbert space (RKHS) is a Hilbert space, in which each point of the space is a continuous linear function. It is a infinite dimensional vector space where even more things behave like the finite case.

An RKHS is constructed on the basis of a Mercer Kernel. A Mercer Kernel $K(x, y)$ is a function of two variables, such that:

$$\int \int K(x, y) f(x) f(y) dx dy > 0, \forall f$$

This is a generalization of a positive definite matrix:

$$x^\top A x > 0, \forall x$$

The most common kernel that we will use is the Gaussian RBF Kernel:

$$K(x, y) = \exp\big(\frac{\|x - y\|_2^2}{\sigma^2}\big)$$

Consider holding one element of the kernel fixed. The result is a function of one variable which we call a feature function. The collection of feature functions is called the feature map.

$$\phi_x := K(x, :)$$

For a Gaussian Kernel the feature functions are unnormalized Gaussians. Here are two examples:

$$\phi_1(y) = \exp\big(\frac{\|1 - y\|_2^2}{\sigma^2}\big)$$

$$\phi_{1.5}(y) = \exp\big(\frac{\|1.5 - y\|_2^2}{\sigma^2}\big)$$

The inner product of feature functions in an RKHS is defined as:

$$\langle \phi_x, \phi_y \rangle = \langle K(x, \cdot), K(y, \cdot) \rangle = K(x, y) \tag{1}$$

Intuitively, this quantity is the dot product between two feature vectors, and is thus a scalar. Because of the symmetric property of kernels it is easy to see that $\phi_x(y) = \phi_y(x) = K(x, y)$.

Having defined feature functions, consider the space composed of all functions that are a linear combination of these feature functions. In effect let:

$$\mathcal{F}_0 = \left\{ f(z) : \sum_{j=1}^{k} \alpha_j \phi_{x_j}(z), \forall k \in \mathbb{N}_+ \text{ and } x_j \in \mathcal{X} \right\} \tag{2}$$

Then, define a Reproducing Kernel Hilbert Space $\mathcal{F}$ to be the completion of the set $\mathcal{F}_0$ defined above. The feature functions thus form a spanning set basis (albeit over-complete) for this space $\mathcal{F}$. Indeed, any object in the RKHS can be obtained as a linear combination of these feature functions, by definition.

With this definition in place, the space $\mathcal{F}$ exhibits the nice Reproducing property, from which the RKHS derives its name. Mathematically this is denoted by: $f(x) = \langle f, \phi_x \rangle$, where f is some function. What this means is that to evaluate a function at some point in infinite dimension, one does not explicitly have to operate in infinite dimensions but can instead simply take the inner product of that function with the feature function mapping of the point. The proof of this property is as follows:

$$\begin{aligned}
\langle f, \phi_x \rangle &= \langle \sum_j \alpha_j \phi_{x_j}, \phi_x \rangle \text{ by definition of the RKHS} \\
&= \sum_j \alpha_j \langle \phi_{x_j}, \phi_x \rangle \text{ because of the linearity of the inner product} \\
&= \sum_j \alpha_j K(x_j, x) \text{ by definition of a kernel in an RKHS} \\
&= \sum_j \alpha_j \phi_{x_j}(x) \text{ by the property of kernels} \\
&= f(x)
\end{aligned} \tag{3}$$

Recall how this property is used to great advantage in SVMs, where data points are symbolically mapped to RKHS feature functions. However, operationally, they are only evaluated with inner products, so that this symbolic mapping never has to be explicited.

4 Embedding Distributions in Reproducing Kernel Hilbert Spaces

We now turn to the problem of embedding entire distributions in RKHS. The theory will be developed for embedding distributions of single variables as well as joint distributions and conditional distributions for two variables. Analogies to linear algebra in the case of finite distributions will be drawn when appropriate.

4.1 The Mean Map – Embedding Distributions of One Variable

We first show how to embed distributions of one variable in RKHS. Consider the mean map defined as:

$$\mu_X(\cdot) = \mathbb{E}_{X \sim \mathcal{D}}[\phi_X] = \int p_{\mathcal{D}}(X) \phi_X(\cdot) dX \tag{4}$$

This is effectively the statistic computed over the feature function mappings of the distribution into the RKHS. It corresponds, intuitively to the "Empirical Estimate" of the data. In the finite case this is simply the first moment: $\hat{\mu}_X = \frac{1}{N}\sum_{n=1}^{N}\phi_{x_n}$. It can be shown that when the kernels are universal, the mapping from distributions to embeddings is one-to-one. The Gaussian RBF Kernel and the Laplacian Kernel are examples of universal kernels.

As an illustrative example consider the finite dimensional case of an RKHS embedding for a distribution that takes on discrete values from 1 to 4. In its explicit form, the moments of this distribution can be computed directly from the data, but leads to loss of information. The nature of the discrete distribution implies that distinction between values is lost. Moreover, the computed statistic is simply a number. Now consider an RKHS mapping of the data into $\mathbb{R}^4$. The feature functions then become:

$$\phi_1 = \begin{pmatrix}1\\0\\0\\0\end{pmatrix}, \phi_2 = \begin{pmatrix}0\\1\\0\\0\end{pmatrix}, \phi_3 = \begin{pmatrix}0\\0\\1\\0\end{pmatrix}, \phi_4 = \begin{pmatrix}0\\0\\0\\1\end{pmatrix},$$

Given this mapping it is easy to see that the mean map is:

$\mu_X = \mathbb{E}_X[\phi_X] = \mathbb{P}[X=1]\phi_1 + \mathbb{P}[X=2]\phi_2 + \mathbb{P}[X=3]\phi_3 + \mathbb{P}[X=4]\phi_4$

Or equivalently in vectorial form:

$$\mu_X = \begin{pmatrix}\mathbb{P}[X=1]\\\mathbb{P}[X=2]\\\mathbb{P}[X=3]\\\mathbb{P}[X=4]\end{pmatrix}$$

which is the marginal probability vector in the discrete case. It is evident that the mean map thus defined is a more expressive statistic than the empirical mean calculated in the simplistic case.

The mean map can be conveniently evaluated by using an inner product as well – rather than dealing with the symbolic infinite dimensional vector. Indeed: $\mathbb{E}_{X\sim\mathcal{D}}[f(X)] = \langle f, \mu_X\rangle$. The proof is as follows:

$$\begin{aligned}\langle f, \mu_X\rangle &= \langle f, \mathbb{E}_{X\sim\mathcal{D}}[\phi_X]\rangle \text{ by definition of the mean map}\\ &= \mathbb{E}_{X\sim\mathcal{D}}[\langle f, \phi_x\rangle] \text{ because of the linearity of expectation}\\ &= \mathbb{E}_{X\sim\mathcal{D}}[f(X)] \text{ by virtue of the reproducing property in RKHS}\end{aligned} \tag{5}$$

4.2 Cross-Covariance – Embedding Joint Distributions of Two Variables

Now consider the problem of embedding joint distributions of two variables in RKHS. Begin by implicitly defining the cross-covariance operator C_{YX} such that $C_{YX} = \mathbb{E}_{YX}[f(X)g(Y)]$. We will show that this definition leads to the following property:

$$\langle g, C_{YX} f\rangle = \mathbb{E}_{YX}[f(X)g(Y)]\ \forall f \in \mathcal{F}, \forall g \in \mathcal{G} \tag{6}$$

Note that the two Hilbert spaces $\mathcal{F}$ and $\mathcal{G}$ no longer need to be analogous, and indeed are likely to be different. It is hypothesized that C_{YX} is then the joint embedding of the distribution over X and Y. C_{YX}

now has the shape of an infinite dimensional matrix, and thus intuitively corresponds to the joint probability table associated with the mapping of the distribution of $P(X, Y)$. In linear algebraic terms, it is an operator that converts from one basis to another.

Let us flesh out these ideas more clearly and formally. Suppose we have $\phi_X \in \mathcal{F}$ and $\psi_Y \in \mathcal{G}$, which are the feature functions of the two RKHS. In the discrete case, the uncentered covariance of two variables is: $COV(X, Y) = \mathbb{E}_{YX}[YX]$. In the infinite dimensional case, this translates to: $C_{YX} = \mathbb{E}_{YX}[\psi_Y \otimes \phi_X]$, where $\otimes$ is the tensor product operator. This operator effectively creates a new space by taking the cross-product of the feature functions in the spaces $\mathcal{F}$ and $\mathcal{G}$.

Intuitively, in the finite case, this corresponds to taking the outer (or cross) product of the elements of two finite sets to obtain a third set that is their tensor product. As a warning, this example should not be taken too literally, because finite sets don't necessarily form vector spaces. Nevertheless this leads to a formal characterization of the tensor product of the two Hilbert spaces:

$$\mathcal{H} = \{h : \exists f \in \mathcal{F}, g \in \mathcal{G} \text{ such that } h = f \otimes g\} \tag{7}$$

The expectation of this new space is then the cross-covariance operator. The proof of correctness of the cross covariance operator property is now given below:

$$\begin{aligned}
\langle g, C_{YX} f\rangle &= \langle g, \mathbb{E}_{YX}[\psi_Y \otimes \phi_X] f\rangle \text{ by definition of the cross covariance operator} \\
&= \mathbb{E}_{YX}[\langle g, [\psi_Y \otimes \phi_X] f] \text{ by moving the expectation outside the inner product} \\
&= \mathbb{E}_{YX}[\langle g, \langle \phi_X, f\rangle \psi_Y] \text{ by the definition of the outer product} \\
&= \mathbb{E}_{YX}[\langle g, \psi_Y\rangle \langle f, \phi_X\rangle] \text{ by redistributing the arguments in the inner product} \\
&= \mathbb{E}_{YX}[f(X)g(Y)] \text{by the reproducing property in RKHS}
\end{aligned} \tag{8}$$

4.2.1 Autocovariance

As an important note consider the cross-covariance of in RKHS of a distribution with itself. This leads to the auto covariance operator defined as: $C_{XX} = \mathbb{E}_{XX}[\phi_X \otimes \phi_X]$. In the discrete case, this would be the uncentered variance of a random variable X: $E[X^2]$. Similarly as C_{YX} was intuitively a matrix, C_{XX} can also be thought of as a matrix. It is, however, a diagonal matrix, whose diagonal elements characterize the distribution of X.

4.3 Product of Cross-Covariances – Embedding Conditional Distributions of 2 Variables

Given what we know about cross-covariance and auto-covariance, we can now proceed to explicit a form for the embedding of conditional distributions of two variables. In simple probabilistic terms we have $P(X, Y) = \dfrac{P(Y|X)}{P(X)}$. In linear algebraic operations, the conditional distribution then emerges as: $P(Y|X) = P(Y, X) \times Diag(P(X))^{-1}$.

But we already know that the embedding of a joint distribution $P(X, Y)$ is cross covariance operator CYX and the embedding of a distribution $P(X)$ in a diagonalized matrix form is the auto covariance operator C_{XX}. It follows that the embedding of a conditional distribution is then also an operator. Specifically we have:

$$C_{Y|X} = C_{YX} C_{XX}^{-1} \tag{9}$$

It can be shown that this operator has the following property: $\mathbb{E}_{Y|x}[\phi_Y|x] = C_{Y|X}\phi_x$, which can be thought of as the slicing operation on a conditional probability table of two variables. That is just selecting a single row of interest from this table: $P(Y|X = x) = P(Y|X)\delta_x$

5 Conclusion and Future Material

In summary this lecture showed how Hilbert Space embeddings (specifically RKHS embeddings) can be effectively used to compute sufficient statistics for arbitrary distribution without the need to make any parametric assumptions about them. A theory for embedding marginal, joint and conditional distributions has been developed and explicited. Analogies with the finite case, specially as regards linear algebraic operations have been noted.

In future lectures this theory will be further developed to allow for defining graphical models on these embedded distributions, as well as perform probabilistic inference on them. Two important results will be shown: the fact that the probabilistic sum rule and chain rule have counterparts in RKHS embeddings, and they can be neatly represented as combinations of functions and operators. Indeed for the sum rule it will be shown that:

$$P(X) = \int_Y P(X, Y) = \int_Y P(X|Y)P(Y) \Longleftrightarrow \mu_x = C_{X|Y}\mu_Y \tag{10}$$

And for the chain rule it will be shown that:

$$P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X) \Longleftrightarrow C_{YX} = C_{Y|X}C_{XX} = C_{X|Y}C_{YY} \tag{11}$$