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

22 : Introduction to Hilbert Space Embeddings and Kernel GM

Lecturer: Eric P. Xing *Scribes: Kevin Lin*

1 Introduction and Motivation

The connection between optimization and graphical models has led to many amazing discoveries such as the class of EM algorithms, variational inference, max-margin and maximum entropy learning. The optimization perspective has many advantages such as being easy to formulate, having principled ways to derive convex relaxations and being able to take advantage of existing optimization tools. But it poses several challenges too, which are particularly limiting to problems that involve non-parametric continuous variables or non-convex objectives.

We study graphical models and its connection to linear algebra in this lecture. Compared to the optimization perspective, the linear algebra view often less intuitive to formulate. However, this perspective allows us to model 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, and tensor algebra. The representational power of this alternative view, however, comes at the cost of expressive intuitiveness. Typically, parametric distributions 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. For example, computing the mean of the distribution no longer uniquely identifies the distribution. 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.

Storing all the moments is, of course, practically infeasible. It however motivates the use of Hilbert Space embeddings, and the kernel trick to solve this infinite dimensional representation scenario. It can create a one-to-one mapping between distributions and the embeddings of their statistics, and it can be cleverly constructed so that the kernel trick can be applied for practical and efficient computation.

2 Hilbert Spaces

A Hilbert space 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 $f, g \in \mathcal{V} \rightarrow \alpha f + \beta g \in \mathcal{V}$. While one normally thinks of these objects as finite dimensional vectors, they can also be infinite dimensional vectors or functions. A Hilbert Space is a complete vector space equipped with an inner product. An example of an inner product might be $\langle f, g \rangle = \int f(x)g(x)dx$.

2.1 Basic Properties

The inner product $\langle f, g\rangle$ in a Hilbert Space must respect the following properties.

1. Symmetry: $\langle f, g\rangle$,
2. Linearity: $\langle \alpha f_1 + \beta f_2, g\rangle = \langle \alpha f_1, g\rangle + \langle \beta f_2, g\rangle$,
3. Non-negativity: $\langle f, f\rangle \geq 0$.
4. Zero: If $\langle f, f\rangle = 0$, then $f = 0$.

2.2 Operators, Adjoints and Outer Products

We now define an operator. An operator C maps a function f in one Hilbert Space to another function g in the same or another Hilbert Space, i.e., $g = Cf$. This operator has the following property:

$$C(\alpha f + \beta g) = \alpha C f + \beta C g.$$

As an analogous intuition one can think of functions as vectors and operators as matrices. In linear algebraic terms a matrix typically transforms 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 $C^T : \mathcal{G} \rightarrow \mathcal{F}$ of an operator $C : \mathcal{F} \rightarrow \mathcal{G}$ is defined such that the following holds:

$$\langle g, Cf\rangle = \langle C^T g, f\rangle \qquad \forall f \in \mathcal{F}, g \in \mathcal{G}.$$

Likewise, the outer product $f \otimes g$ is defined such that $f \otimes g(h) = \langle g, h\rangle f$.

3 Reproducing Kernel Hilbert Space

We now introduce Reproducing Kernel Hilbert Spaces, specialized Hilbert Spaces. A Reproducing Kernel Hilbert space (RKHS) is a Hilbert space in infinite dimensions where each point of the space is a continuous linear function. An RKHS is constructed from a Mercer Kernel. A Mercer Kernel $K(x, y)$ is a function of two variables, such that

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

This is a generalization of the positive definite matrix.

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

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

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,

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

For example, using the Gaussian Kernel, the feature functions are unnormalized Gaussians. Here is an examples:

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

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

$$\langle \phi_x, \phi_y \rangle = \langle K(x,\cdot), K(y,\cdot) \rangle := K(x,y).$$

Intuitively, this quantity is the dot product between two feature vectors. By the symmetric property of kernels, $\phi_x(y) = \phi_y(x) = K(x,y)$.

Having defined feature functions, consider the space composed of all functions that are a linear combination of these feature functions. That is,

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

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: $\langle f, \phi_x \rangle = f(x)$, 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,

$$\begin{aligned}
\langle f, \phi_x \rangle &= \langle \sum_j \alpha_j \phi_{x_j}, \phi_x \rangle \\
&= \sum_j \alpha_j \langle \phi_{x_j}, \phi_x \rangle && \text{Linearity of inner product} \\
&= \sum_j \alpha_j K(x_j, x) && \text{Definition of kernel} \\
&= f(x).
\end{aligned}$$

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 explicit.

4 Embedding Distributions in Reproducing Kernel Hilbert Spaces

We now turn to the problem of embedding entire distributions in RKHS.

4.1 The Mean Map - Embedding Distributions of One Variable

We first show how to embed univariate in RKHS. Consider the mean map defined as:

$$\mu_X(\cdot) = \mathbb{E}_{X\sim D}[\phi_x] = \int p_D(x)\phi_X(\cdot)dx.$$

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. Now consider an RKHS mapping of the data into $\mathbb{R}_4$. Let the feature functions in this RKHS be

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

Given this mapping, the mean map is

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

This is the marginal probability vector in the discrete case. It is evident that the mean map (a 4-dimensional vector in this case) 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. That is, $\mathbb{E}_{X\sim D}[f(X)] = \langle f, \mu_X\rangle$. The proof is as follows,

$$\begin{aligned}\langle f, \mu_X\rangle &= \langle f, \mathbb{E}_{X\sim D}[\phi_X]\rangle && \text{Def of Mean Map}\\ &= \mathbb{E}_{X\sim D}[f(X)]. && \text{Reproducing property}\end{aligned}$$

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,

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

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

Now we show how C_{YX} is constructed. Suppose we have $\phi_X \in \mathcal{F}$ and $\phi_Y \in \mathcal{G}$, which are the feature functions of the two RKHS. For two random variables, the covariance of two centered variables is $\mathrm{Cov}(X,Y) = \mathbb{E}_{YX}[XY]$. In the infinite dimensional case, this translates to $C_{YX} = \mathbb{E}_{YX}[\phi_Y \otimes \phi_X]$, 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 $\mathcal{G}$. This leads to a formal characterization of the tensor product of the two Hilbert spaces,

$$\mathcal{H} = \{h : \exists f \in \mathcal{F}, \exists g \in \mathcal{G} \text{ s.t. } h = f \otimes g\}.$$

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,

$$\begin{aligned}\langle g, C_{YX} f\rangle &= \langle g, \mathbb{E}_{YX}[\phi_Y \otimes \phi_X] f\rangle \\ &= \mathbb{E}_{YX}[\langle g, [\phi_Y \otimes \phi_X] f\rangle] \\ &= \mathbb{E}_{YX}[\langle g, \langle \phi_X, f\rangle \phi_Y\rangle] \qquad \text{Def of outer product} \\ &= \mathbb{E}_{YX}[\langle g, \phi_Y\rangle \langle \phi_X, f\rangle] \\ &= \mathbb{E}_{YX}[g(Y) f(X)] \qquad \text{Reproducing property.}\end{aligned}$$

Taking the covariance operator with itself leads to the auto-covariance operator.

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 $\mathbb{P}(X,Y) = \mathbb{P}(Y|X)\mathbb{P}(X)$. In linear algebraic operations, the conditional distribution then emerges as $\mathbb{P}(Y|X) = \mathbb{P}(Y,X) \cdot \mathrm{Diag}(\mathbb{P}(X))^{-1}$. But we already know that the embedding of a joint distribution $\mathbb{P}(X,Y)$ is cross covariance operator C_{YX} and the embedding of a distribution $\mathbb{P}(X)$ 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

$$C_{Y|X} = C_{YX} C_{XX}^{-1}.$$

It can be shown that this operator has the following property, $\mathbb{E}_{Y|X}(\phi_Y|X) = C_{Y|X}\phi_X$.

5 Kernel Graphic Models

As we can embed marginal distribution, joint distribution and conditional distribution in RKHS space, we can use these embeddings to replace the conditional probability tables we used before in graphic model to build the kernel graphic model. We can also perform inference on kernel graphic model with the sum rule and chain rule in RKHS.

Specifically, the sum rule for densities and its corresponding rule in RKHS is

$$\mathbb{P}[X] = \int_Y \mathbb{P}[X,Y] = \int_Y \mathbb{P}[X|Y]\mathbb{Y} \iff \mu_X = C_{X|Y}\mu_Y.$$

Likewise for the chain rule,

$$\mathbb{P}[X,Y] = \mathbb{P}[X|Y]\mathbb{P}[Y] = \mathbb{P}[Y|X]\mathbb{Y} \iff C_{YX} = C_{Y|X}C_{XX} = C_{X|Y}C_{YY}.$$

Consider a simple graphic model shown in Figure 1. If the variables in this model are discrete, we can parameterize the model using probability vector $\mathbb{P}[A]$ and conditional probability matrix $\mathbb{P}[B|A]$, $\mathbb{P}[C|B]$ and $\mathbb{P}[D|C]$. Then, we can do inference based on these matrix. For example, let's see how we would compute $\mathbb{P}[A=a, D=d]$.

To do this, we need to compute the joint distribution matrix $\mathbb{P}[A,D]$. However, if we directly use matrix multiplication, $\mathbb{P}[A]$ would be integrated out so we can only get $\mathbb{P}[D]$. To solve this problem, we convert $\mathbb{P}[A]$ to a diagonal matrix $\mathbb{P}[\emptyset A]$. Then, $\mathbb{P}[A,D]$ could be calculated using

$$\mathbb{P}[A,D] = \mathbb{P}[\emptyset A]\mathbb{P}[B|A]^T\mathbb{P}[C|B]^T\mathbb{P}[D|C]^T.$$

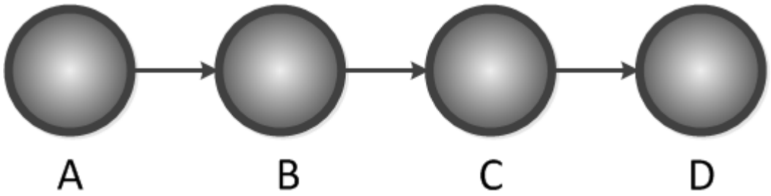

Figure 1: Example Graphical Model

To compute the probability $\mathbb{P}[A = a, D = d]$, we introduce the evidence vectors δ_a and δ_d where (for example) δ_a is the all-zero vector except for the element corresponding to element a.

If the variables were continuous, then we can use the cross-covariance operators we described earlier. For example,

$$C_{AD} = C_{AA} C^T_{B|A} C^T_{C|B} C^T_{D|C} \mathbb{P}[A = a, D = d] \propto \phi_a^T C_{AD} \phi_d.$$

These examples show that inference on kernel graphic model is similar to inference on regular graphic model. Therefore, we can apply the inference algorithm on regular graphic model, such like message passing algorithm to kernel graphic model by replacing the sum-product operations with tensor operations.