Kernelized Locality Sensitive Hashing

Daniel Speyer and Geelon So
{dls2192,geelon}@columbia.edu

December 14, 2017

Abstract

1 LSH

Locality sensitive hashing is a general solution for searches in which the desired key is near or nearest to the query, rather than equal. It is not as efficient as tree-based solutions for $\mathbb{R}^n$ with small n, but unlike those solutions, it maintains its performance in high dimensions or in arbitrary metric spaces.

Most near search problems can be reduced to cr-near search. Suppose we have a database of points $x_1, x_2, ..., x_n$ drawn from some arbitrary metric space $\mathcal{X}$ and some query q also from $\mathcal{X}$, plus constants $c, r \in \mathbb{R}$. A cr-near search seeks either an x within distance cr of q or a statement that there are none within r. It is generally best to think of r as part of the query and c as the acceptable level of error.

To use LSH for problems like this we will need a family of hashing functions $h_1, h_2, ...$ that map $\mathcal{X}$ into some small finite space, all following two key inequalities:

$$\begin{aligned} ||x_1 - x_2|| < r &\quad\rightarrow\quad Pr[h(x_1) = h(x_2)] \geq p_1 \\ ||x_1 - x_2|| > cr &\quad\rightarrow\quad Pr[h(x_1) = h(x_2)] \leq p_2 \end{aligned}$$

Note that we may require an arbitrary number of hashing functions, which must be generated randomly. It is this randomness that allows the use of probabilities in the preceding inequalities. (The distribution of hashing functions is often written $\mathcal{H}$, but we'll be reserving that symbol for Hilbert spaces later.)

In the simplest example, $\mathcal{X}$ is a high dimensional Hamming space and the hash functions sample random columns. This may provide useful intuition for thinking about hash functions.

Once we have a source of functions and a gap between p_1 and p_2, we can widen the gap by taking n functions and concatenating their outputs (getting p_1^n and p_2^n). We can then deal with too low a p_1^n by taking m repetitions and searching all of them (getting mp_1^n and mp_2^n). This solves the cr-near problem with arbitrarily high probability, which can in turn solve other, similar problems.

2 Kernels

One of the most natural ways to encode structure into an abstract set $\mathcal{X}$ is to provide a mapping from that set to the reals, $\mathcal{X} \longrightarrow \mathbb{R}$. Or more generally, a mapping to some structure-rich mathematical object $\mathcal{O}$.

Indeed, at a high level, we can consider different mathematical objects as the 'canonical' idealization of certain types of structure; then, certain functions from $\mathcal{X}$ to these objects allow us to 'pull back' or instantiate that structure within our specific $\mathcal{X}$.

Of course, the structure we care about with respect to LSH is *similarity*, one idealization of which is the *(real) inner product*. Let us take a moment to reexamine the inner product on finite real vector spaces as a notion of similarity, before generalizing that notion using *kernels* to infinite dimensional vector spaces and arbitrary sets.

2.1 Finite-Dimensional Inner Product Spaces

Definition 1. Let V be a (possibly infinite) vector space over $\mathbb{R}$. An *inner product* on V is a bilinear map $K : V \times V \longrightarrow \mathbb{R}$ that is *symmetric* and *positive-definite*. We call the pair (V, K) an *inner product space*.

Symmetry and positive-definiteness are two of the most obvious requirements for any self-respecting function that calls itself a measure of similarity:

1. the similarity between two points x and y better not depend on the order we specify them, so we require $K(x, y) = K(y, x)$,
2. a nonzero object should be similar to itself, so if $x \neq 0$, then $K(x, x) > 0$.

These two simple intuitions of a 'similarity measure' then precisely define the conditions of an inner product on V.[1] Allow us to be (perhaps overly) rigorous in the following discussion, for it is better to bore in the finite case than to confuse in the general case.

For concreteness, let $V = \mathbb{R}^n$ with a fixed basis. With any $x, y \in V$, denote by $\langle x, y\rangle$ the *formal dot product* of x and y as an algebraic construction.[2] Then, as K is a bilinear form, there exists a matrix $\mathbf{K}$ where

$$K(x, y) = \langle x, \mathbf{K}y\rangle. \tag{1}$$

The symmetry condition on the inner product K forces the matrix $\mathbf{K}$ to be symmetric. The spectral theorem on symmetric matrices tell us that V has an eigenbasis with respect to $\mathbf{K}$. Consider an eigenvalue $\mathbf{K}v = \lambda v$. As we require K to be positive-definite, we know that

$$K(v, v) = \langle v, \mathbf{K}v\rangle = \lambda\langle v, v\rangle > 0.$$

Since the dot product of a vector with itself in any basis is nonnegative, this implies that the eigenvalue λ is positive; so in its eigenbasis, $\mathbf{K}$ is diagonal with positive terms on its diagonal—it follows that $\mathbf{K}^{1/2}$ exists, and this allows us to write:

$$K(x, y) = \langle \mathbf{K}^{1/2}x, \mathbf{K}^{1/2}y\rangle.$$

We deduce that there exists a linear map $\Phi : V \longrightarrow \mathbb{R}^n$ (spoiler: we call Φ a *feature map*) where the inner product on V directly

[1] For now, we can just view bilinearity as a necessary condition to ensure that K respects the real numbers as an algebraic object. So, in some sense, bilinearity is not intrinsic to K; rather, it is a condition *induced* by the field $\mathbb{R}$. Later on, when we generalize, we won't have a notion of (bi)linearity on arbitrary sets $\mathcal{X}$; however, there is a natural 'completion' of $\mathcal{X}$ as a vector space $\mathcal{H}$. Here, $\mathbb{R}$ induces the linear structure.

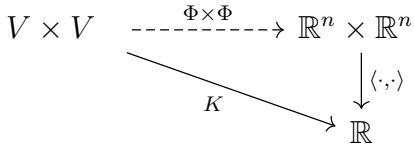
[2] One point of confusion may be that $\langle\cdot,\cdot\rangle$ often denotes an inner product—usually unproblematic as the dot product is in fact the inner product *induced by the choice of basis* on V. However, an abstract vector space has no canonical basis; thus, it has no canonical inner product. In particular, here, the inner product corresponding to the dot product is in general unrelated to the inner product K. We will reconcile these two views at Equation 2, which shows that every inner product corresponds to the dot product in an appropriate basis. Conversely, as every dot product is an inner product, if V is an abstract finite-dimensional real vector space, *specifying an inner product on V is equivalent to specifying a basis on V*.

corresponds to the usual dot product on $\mathbb{R}^n$. That is,

$$K(x, y) = \langle \Phi(x), \Phi(y) \rangle. \tag{2}$$

In other (ridiculously abstract) words, Equation 2 precisely says:

Proposition 2. *Let V be an n-dimensional real vector space. If $K : V \times V \longrightarrow \mathbb{R}$ is an inner product, then there exists a map $\Phi : V \longrightarrow \mathbb{R}^n$ such that the following map commutes:*



where $\langle \cdot, \cdot \rangle$ is the standard dot product on $\mathbb{R}^n$.

This is the main payoff to all this formality: we may view the pair $(\mathbb{R}^n, \langle \cdot, \cdot \rangle)$ as the canonical n-dimensional real inner product space. So, no matter which inner product space (V, K) we want to study, we're in fact guaranteed the existence of an isomorphism Φ to the familiar inner product space $\mathbb{R}^n$, and we may freely interchange the two objects in our minds.

Furthermore, we are now justified in dropping the distinction between the formal dot product and inner product; let us denote both by $\langle \cdot, \cdot \rangle$.

2.2 Generalization: Kernels and Hilbert Spaces

First, we will need to extend the 'similarity measure' from vector spaces V to abstract sets $\mathcal{X}$. In the former, notice that given a basis $\{v_1, \ldots, v_n\}$, the inner product K is fully determined by the collection of values, $K(v_i, v_j)$ where i, j range between 1 and n. This follows from bilinearity of K; this is equivalent to saying that we can represent the bilinear form K by a matrix of $n \times n$ values $\mathbf{K}$. In some sense, this means that there are only 'n ways' or 'n directions' in which objects of V may be (dis)similar.

This suggests that we can generalize Definition 1 because we didn't need all of V to start with! Suppose someone secretly had an n-dimensional inner product space (V, K), but just gave us n linearly independent vectors, say $[n] := \{v_1, \ldots, v_n\}$, along with the corresponding restriction of the similarity measure $K : [n] \times [n] \longrightarrow \mathbb{R}$.

Our view of $[n]$ would just be a collection of n abstract objects, and K just an $n \times n$ matrix over $\mathbb{R}$ (this is often called the *Gram matrix*). Still, we would have produced the same feature map $\Phi_{[n]} : [n] \hookrightarrow \mathbb{R}^n$, albeit restricted to $[n] \subset V$. But in the previous analysis, since V and $\mathbb{R}^n$ are isomorphic via Φ, this implies that $\Phi_{[n]}$ actually *recovers* V by its identification with $\mathbb{R}^n$; it is as though we've discovered that $[n]$ actually lived inside the space V.

Very naturally, the generalization of K from $[n]$ to $\mathcal{X}$ takes the view that K is a similarity 'matrix' on $\mathcal{X}$ (though we call it a *kernel*):

Definition 3. Let $\mathcal{X}$ be a nonempty set. A *kernel*[3] is a map $K : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$. We say that K is *positive-definite* if for any finite subsets A of $\mathcal{X}$, the corresponding Gram matrix

[3]Unfortunately, the terminology *kernel* can be terribly confusing. Within functional analysis, kernels are often interchangably called *kernel maps*, *kernel functions*, and *positive-definite kernels*. Yet, each of these also take on other meanings depending on author and context. 'Kernels' and 'kernel maps' are almost always synonymous. Sometimes, 'kernel functions' denote kernels of the form $K(x, y) = k(x - y)$, in situations where subtraction is defined. And in many instances, when context is clear, the author cares only for positive-definite kernels, so omitting the specifier *positive-definite*.

Furthermore, *kernels* used in this sense are unrelated to the algebraic kernel of a linear map. The terminology comes from the theory of integral operators. In particular, infinite-dimensional function spaces often

of A is symmetric and positive-definite.

In the finite case, we produced an embedding $\Phi_{[n]} : [n] \longrightarrow \mathbb{R}^n$ of $[n]$ satisfying Equation 2 (K corresponds to inner product). As we had only n objects, it is clear that we needed an inner product space with at most n dimensions, hence $\mathbb{R}^n$.

However, in the general case, where $\mathcal{X}$ may be infinite, we may need infinite dimensions. Consider the real vector space $\mathbb{R}^{\mathcal{X}}$ (vector addition and scalar multiplication are defined pointwise), and the map $\Phi : \mathcal{X} \longrightarrow \mathbb{R}^{\mathcal{X}}$ defined by:

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

For short, let $k_x := \Phi(x)$. Then, Φ maps $\mathcal{X}$ into a subset of $\mathbb{R}^{\mathcal{X}}$; let $V = \text{span}(\Phi(\mathcal{X}))$ be a linear subspace of $\mathbb{R}^{\mathcal{X}}$, so that elements f of V are of the form:

$$f = \alpha_1 k_{x_1} + \cdots + \alpha_n k_{x_n},$$

with $\alpha_i \in \mathbb{R}$ and n ranging over $\mathbb{N}$. We claim that because K is a positive-definite kernel, there exists a well-defined bilinear map on V, corresponding to our desired inner product. Of course, we want the inner product to satisfy:

$$\langle k_x, k_y \rangle = K(x, y).$$

But in fact, once we specify this, linearity determines the inner product on general elements $f = \sum_{i=1}^{n_1} \alpha_i k_{x_i}$ and $g = \sum_{j=1}^{n_2} \beta_j k_{y_j}$:

$$\langle f, g \rangle = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \alpha_i \beta_j K(x_i, y_j).$$

This should feel quite familiar, recalling how the Gram matrix of n linearly independent vectors fully determined the inner product in the finite case. However, we must be a little more careful here, as we are not guaranteed that the collection of k_x's are linearly independent. In particular, suppose that f, as follows, is identically zero:

$$0 \equiv f = \sum_{i=1}^{n} \alpha_i k_{x_i}.$$

That is, $f(x) = 0$ for all $x \in \mathcal{X}$. Then, $\langle f, g \rangle$ better equal 0 for all $g \in V$. Notice that it is sufficient to show that $\langle f, k_x \rangle = \langle k_x, f \rangle = 0$ for each $x \in \mathcal{X}$. And indeed, by assumption

$$\langle k_x, f \rangle = \sum_{i=1}^{n} k_{x_i}(x) = f(x) = 0.$$

This proves that V admits an inner product that is compatible with K in the sense of Equation 2. We will in fact go beyond producing an inner product space V; we can complete the space with respect to the norm induced by the inner product, producing a *Hilbert space*.

That is, let $\mathcal{H}$ be the completion of V by taking equivalence classes of Cauchy sequences on V. We claim that $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ is a collection of functions in the sense that $f(x) < \infty$ for all $f \in \mathcal{H}$ and $x \in \mathcal{X}$.[4] But let us relegate the proof to references, say [1, Thm 3.16], for it is not particularly enlightening. However, the consequences are quite important.

First of all, we immediately attain the analog to Proposition 2:

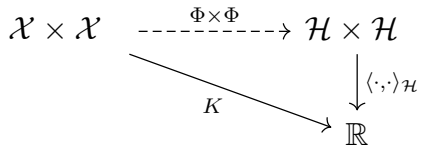
have inner products of the form:

$$\langle f, g \rangle = \int_{X \times X} K(x, y) f(x) g(y) \, \mathrm{d}x \mathrm{d}y,$$

where the K here performs the analogous role as $\langle x, \mathbf{K}y \rangle$ in Equation 1. Actually, it performs precisely the role we desire in the general case. These maps K were historically called kernels.

[4]Note that in general, it is not the case that the completion of a function space consists of functions. Elements of $L^2(\mathbb{R})$, for example, are equivalence classes of functions.

Proposition 4 (Moore). *Let $\mathcal{X}$ be a set, and $K : \mathcal{X} \times \mathcal{X} \longrightarrow \mathbb{R}$ a positive-definite kernel. Then, there exists a map $\Phi : \mathcal{X} \longrightarrow \mathcal{H}$ into a Hilbert space such that the following map commutes:*



where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the inner product on $\mathcal{H}$.

This is a powerful result because we may now view $\mathcal{X}$ not as an abstract set, but as vectors within a Hilbert space, which carries with it both algebraic and geometric properties which $\mathcal{X}$ now inherits. It is particularly amazing because the only thing we needed required was a positive-definite kernel K to obtain $\mathcal{H}$. And in fact:

Fact 5. *The Hilbert space $\mathcal{H}$ induced by K is unique. We call $\mathcal{H}$ the* reproducing kernel Hilbert space (RKHS) *on $\mathcal{X}$ with respect to K. (See [1, Prop 3.3]).*

Later, we'll see a converse: if $\mathcal{H}$ is an RKHS, so it is a Hilbert space composed of a collection of *functions* over $\mathcal{X}$ (in particular, $L^2(\mathbb{R})$ is not an RKHS), then $\mathcal{H}$ induces a unique kernel K on $\mathcal{X}$. So, analogous to the finite-dimensional case, we can think of $\Phi(\mathcal{X})$ in $\mathcal{H}$ as the 'canonical form' of $\mathcal{X}$ with respect to the similary measure K.

Despite this, the way we've presented the embedding of $\mathcal{X}$ into $\mathcal{H}$ is somewhat unnatural. While it makes sense to consider the collection of functions k_x, the identification of x with k_x may seem *ad hoc*. But if we venture a bit further into functional analysis, we may obtain a clearer understanding of RKHS's. Along the way, hopefully we can clarify why we might call Φ a *feature map*.

2.3 Feature Maps

Let us once again consider an abstract set $\mathcal{X}$, and a collection of functions over $\mathcal{X}$, say $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$. We may imagine each $f \in \mathcal{H}$ as an observable or a feature, where $f(x)$ gives the *measurement* value of the object x by feature f.

As a result, every $x \in \mathcal{X}$ is associated to the Cartesian product of measurements,

$$x \overset{\Phi}{\mapsto} \prod_{f \in \mathcal{H}} f(x). \tag{3}$$

The high-level punchline here will be that $\Phi(x)$ is actually a point in the *dual* of $\mathcal{H}$, following the canonical mapping of $\mathcal{X}$ into the double dual $\mathcal{X}^{**}$. But in the case where $\mathcal{H}$ is a Hilbert space and all the $\Phi(x) \in \mathcal{H}^*$ are bounded (i.e. continuous), then $\mathcal{H}$ is isomorphic to $\mathcal{H}^*$, justifying the initial definition $\Phi : x \mapsto k_x$.

Before elaborating on this, let us provide a visual example; perhaps a bit of concreteness will convince the reader why we should even care about the association in Equation 3 in the first place.

Imagine $\mathcal{X}$ is a set of abstract cats, while $\mathcal{H} = \{f_1, \ldots, f_n\}$ represents the pixels within the lens of a camera. Given a cat x, the value $f_i(x)$ gives the color the ith pixel on the camera. Then, the product $\Phi(x)$ in Equation 3 represents the *image* of the cat—the collection of colors all the camera pixels see as the camera takes the picture of cat x.

Here, it makes sense to call the mapping in Equation 3 the *feature map*, for it maps the cat to a collection of features, specifically its colors at different locations. In general, given any function $f : \mathcal{X} \longrightarrow \mathbb{R}$, we may dually view points of x as functionals, $\mathrm{ev}_x : \mathcal{H} \longrightarrow \mathbb{R}$, where

$$\mathrm{ev}_x(f) := f(x).$$

In functional analysis, we say that ev_x is the *evaluation at x*, and we view it as an element

of the *dual space* $\mathbb{R}^{\mathcal{H}}$ of $\mathcal{H}$, which we denote by $\mathcal{H}^*$.[5] To recapitulate, $\Phi(x) = \text{ev}_x$.

Returning back to kernels as measures of similarity, we can obtain a more subtle understanding of the function $k_x(y) = K(y, x)$. Here, k_x is an observable/feature, corresponding to *how similar is y to x*—in some sense, x is 'doing the measurement', while y is the object 'being measured'. So technically, we should think of $k_x(y)$ as:

$$k_x(y) \equiv \text{ev}_y(k_x).$$

That is, there are actually two objects: k_x and ev_y. [NEED TO CLEAN UP NOTATION] Given $K(\cdot, \cdot)$, it is perfectly natural to consider $\Phi : x \mapsto \text{ev}_x$ and $\Psi : x \mapsto k_x$. Now, assuming that there is a RKHS $\mathcal{H}$ containing $\Psi(\mathcal{X})$ such that $\Phi(\mathcal{X}) \subset \mathcal{H}^*$, then we would have

$$K(x, y) = \langle \Phi(x), \Psi(y) \rangle_{\mathcal{H}}.$$

It is the conditions of symmetry and positive-definiteness that allows, and actually forces, $\Phi \cong \Psi$ to be isomorphic via the canonical isomorphism between $\mathcal{H}$ and its dual.

3 Unusual Distance Metrics

4 KLSH

Now we are ready to put the pieces together into KLSH. Suppose we have a set of objects $\mathcal{X}$ on which the only defined function is a simularity metric $K : \mathcal{X} \times \mathcal{X} \to [0, 1]$ (with $K(x, x) = 1$). In the most common example, $\mathcal{X}$ is images and K is whatever the computer vision community recommends. And let us suppose we know K to be positive-definite.

Note that $\mathcal{X}$ is *not* a vectorspace. We cannot add two images, nor take their inner product. We cannot generate a random image drawn from a gaussian distribution. These things are not defined. Furthermore, we cannot find a function Φ that would map the image into a vectorspace. We know one exists, but we do not have it, nor do we know any interesting properties of the Hilbert space it maps into.

Nevertheless, we would like to apply LSH techniques that are defined in the kernel space to search problems in the object space.

4.1 Getting a Normal Distribution

Most LSH techniques – including hyperplane carving the unit sphere, which will be our first attempt here – require drawing points g from a normal distribution. As we observed, we cannot do this in $\mathcal{X}$ because it is undefined, and we cannot do this in $\mathcal{H}$ because we cannot describe any point in $\mathcal{H}$.

Nevertheless, we can draw points in $\mathcal{H}$ from a normal distribution, and compute their inner products with points of the form $\Phi(x)$.

The key is the central limit theorem. This theorem states that if you draw n points i.i.d. from any distribution and average them, the average will approximately be drawn from a normal distribution, whose mean is the same as that of the original, and whose covariance is that of the original times $\sqrt{n}$. The "approximately" goes away as n approaches infinity. In general, $n = 30$ is sufficient to make a good enough approximation, though the theoretical bound is poorly studied, and extremely multimodal distributions might require larger ns.

[5] More precisely, $\mathcal{H}^*$ is the *continuous dual* of $\mathcal{H}$, and not the *algebraic dual*. The former is the collection of continuous functionals on $\mathcal{H}$. The earlier point that $\mathcal{H}$ is a collection of functions over $\mathcal{X}$ is equivalent to saying that the evaluation functionals are continuous.

If we select a random point x from our database (equal chance of each point) and consider its $\Phi(x)$, this is a distribution in $\mathcal{H}$, albeit a strange and discreet one. That's good enough for the central limit theorem to apply. This lets us "draw" a point in $\mathcal{H}$ from $\mathcal{N}(\mu, \Sigma)$. Furthermore, we can convert this into a point from $\mathcal{N}(0, I)$ by subtracting μ and multiplying by $\Sigma^{-1/2}$. That is to say:

$$g = \Sigma^{-1/2}\left(\frac{1}{n}\sum_{i}^{n}\Phi(x_i)\right) - \mu \sim \mathcal{N}(0, I)$$

Granted, we still cannot compute this. By approximating Σ, however, we can compute per-point scalars w_i such that $g = \sum_i w_i \Phi(x_i)$. We can then use this form to compute inner products:

$$\begin{aligned}\langle \Phi(q), g\rangle &= \langle \Phi(q), \sum_i w_i \Phi(x_i)\rangle \\ &= \sum_i w_i \langle \Phi(q), \Phi(x_i)\rangle \\ &= \sum_i w_i K(q, x_i)\end{aligned}$$

4.2 Approximating the Covariance

4.2.1 Interpretation as Projection

4.2.2 Points to be Dropped

5 Tweaking Parameters

5.1 Numbers of Points

5.2 Number of Eigenvectors

6 Data-Dependent KLSH

6.1 Data-Dependent LSH

6.1.1 Approximate Evenness

6.2 Smaller Caps

6.3 Making the Calculations

References

[1] Paulsen, Vern I., and Mrinal Raghupathi. *An introduction to the theory of reproducing kernel Hilbert spaces*. Vol. 152. Cambridge University Press, 2016.