A Random Introduction To Random Projections

Aaditya Ramdas
adidas@cmu.edu

Abstract

The idea of random projections is now pervasive in machine learning literature, useful in both improving our theoretical understanding of high-dimensional problems and providing practical algorithms that have changed the field. In this survey, we hope to give the reader an introduction into this maginificent world of random projections, some of its curiosities and wonders, and its relations to other very important concepts in learning theory.

1 Introduction

The kinds of results that we will talk about can be found in the fantastic survey [B05] and other references are pointed out when relevant. We shall summarise many of these ideas in words before we get into the mathematical statements of the results or proof ideas.

The first result known as the Johnson-Lindenstrauss Lemma (lovingly referred to as JL-Lemma), is the cornerstone of the entire field. It states that if we are given a set of points in a high-dimensional space, then we can project these points down into a much lower-dimensional random subspace that is independent of the original dimension, and with high probability preserve a lot of its structure in terms of its inter-point distances and angles.This seems like a fairly simple and yet quite radical statement - how can the projected dimension be independent of the original dimension and what are the implications of being able to carry out such a random projection?

One such implication is that a high-dimensional classification problem with a large margin (to be defined later) is easy to learn, because one can always project down to a lower dimension without affecting separability and is hence an intrinsically low-dimensional problem. Another puzzling result is that if we have data that is separated by some small margin, then a random linear separator is, with sizeable probability, a weak learner (one whose error is strictly less than $1/2$).

One can think of a kernel function as an implicit map to a higher dimensional space in order to perform certain operations there without paying a high price computationally. We can now ask the crazy question, that if on using a kernel, the data is now linearly separable in the high-dimensional space, then can one randomly project back down to lower dimensions and preserve this separability?

Interestingly, the answer to the above question is in the affirmative, and we will show how to construct a combined map, first using the kernel and then using a random projection. This implies that any kernel, under which the data is linearly separable, can be thought of as a mapping to an implicitly *low*-dimensional subspace.

These are the kinds of questions we will explore in more detail in this survey. While it is impossible to do justice to such a rich and mysterious subject, we intend to tickle the reader's curiosity into exploring these ideas further.

2 Definitions, Notation

Formal Setting Let us assume that examples are shown to us according to some probability distribution P_X over an instance space of examples X and labeled by some existing unknown target function $f : X \to \{-1, 1\} =: Y$. Let P_{XY} be the joint distribution generated by (P_X, f).

Learning Model We work in the PAC-style framework, where given a set S of labeled training examples, each drawn independently from P_X and labeled by f, our objective is to come up with a hypothesis h with low true error $Pr_{x \sim P_X}(h(x) \neq f(x))$. If the hypothesis h has a true error very close to, but just less than, a half, then it said that h weak-learns P_{XY}.

Margin Separability We say that a set S of labeled examples is linearly separable by margin γ if there exists a unit vector w such that $\forall (x, y) \in S, \frac{<w,x>}{\|x\|} y \geq \gamma$. This means that the linear hyperplane separator $w^\top x \geq 0$ gets all the examples correct, and that the cosine of the angle between w and all examples x is at least γ.

Approx. Separability We say that P_{XY} is linearly separable by margin γ if there exists a unit vector w such that $Pr_{(x,y) \sim P_{XY}}[\frac{<w,x>}{\|x\|} y \geq \gamma] = 1$. We say that P_{XY} is linearly separable with error α by margin γ if there exists a unit vector w such that $Pr_{(x,y) \sim P_{XY}}[\frac{<w,x>}{\|x\|} y \geq \gamma] \geq 1 - \alpha$.

Kernel Function A kernel is a bivariate real-valued function $K : X \times X \to \mathcal{R}$. A legal kernel is one such that $\forall x, y \, K(x, y) =< \phi(x), \phi(y) >$. It can be thought of as a similarity function, or as the dot product of the two points in the ϕ-space. We normally project points into a higher-dimensional space in the hope that they are linearly separable in that space.

3 Johnson-Lindenstrauss Lemma

The JL lemma states that any set of S points in $\mathcal{R}^n$ can be orthogonally projected down into a d-dimensional subspace so that with high probability all pairwise distances and angles are preserved up to a $1 \pm \gamma$ factor if $d = O(\frac{\log |S|}{\gamma^2})$.

Conceptually, one can think of applying a random rotation to $\mathcal{R}^n$ and then reading off the first d coordinates. One can also multiply the $|S| \times n$ matrix of points by an $d \times n$ projection matrix A, where A's rows are d random orthogonal unit vectors. The following lemma is not often not stated in this form, but is equivalent to taking a set of S points and projecting them randomly into $log|S|$ dimensions while maintaining pairwise distance and angles.

JL Lemma Let $u, v \in \mathcal{R}^n$. Let A be an $d\times$ matrix of random orthogonal unit vectors. Let $u' = \frac{1}{\sqrt{d}} Au$ and $v' = \frac{1}{\sqrt{d}} Av$. Then, calculated over the randomness of A, $Pr[(1 - \gamma)\|u - v\|^2 \leq \|u' - v'\|^2 \leq (1 + \gamma)\|u - v\|^2] \geq 1 - 2e^{-(\gamma^2 - \gamma^3)\frac{d}{4}}$.

An outline of a simple proof by Dasgupta and Gupta is as follows. One can think of projecting a fixed vector onto a random subspace as projecting a random n-dimensional vector onto a fixed subspace (say the subspace formed by the first d coordinates of the vector). We use Hoeffding's inequality to show that the length of this vector is tightly concentrated around its mean d/n with high probability. We then use a union bound over all differences between pairs of points.

Initially, it was thought that the projection needed to be an orthogonal, random and spherically symmetric. Indyk and Motwani showed that approximately orthogonal was enough (random gaussians). Achlioptas proved that even spherical symmetry wasn't needed (random rademacher). He also noticed that these matrices could be two-thirds zero ($\pm 1, 0$ with probability $\frac{1}{6}, \frac{1}{6}, \frac{2}{3}$). Magen used the fact that it preserved pairwise angles between vectors, that it also preserves volumes of small sets.

However, Ailon and Chazelle argued that the matrix cannot get too sparse because it will typically distort a sparse vector, causing the resulting projection to be poorly concentrated. They get around this problem by showing that you can use a randomized FFT to densify a sparse vector without sparsifying the dense vectors, hence distributing the mass of all vectors over all their components, allowing you to take a sparse projection.

The construction of the Fast JL Transform (FJLT) is done by multiplying the points by AHD where H is the normalized Hadamard (real fourier) matrix, and D is a random diagonal matrix of rademacher variables, and A is a sparse projection matrix. For a unit vector $p \in \mathcal{R}^n$, we have $(HDp)_i = \sum_j H_{ij} D_{jj} p_j = O(1/\sqrt{n})$ with high probability over the randomness of D (using Hoeffding's and $|H_{ij}| = 1/sqrtn$). Because H and D are isometries and $(HDP)_i$ are iid, all the entries are of the same order, and hence is not sparse with its energy spread out, and can be projected safely by sparse A.

There are extensions of the JL Lemma from euclidean spaces to affine subspaces (Sarlos) or even other normed spaces (Johnson and Naor). One can move away from linear subspaces to attain the additive embedding of any set of points with bounded doubling dimension (Indyk and Naor).

4 Margins and Kernels

Large margins are known to make learning easy, even in high dimensions in the sense that if a learning problem is linearly separable by a large margin γ, then one needs only a number of examples that is polynomial in $1/\gamma$ to achieve good generalisation, with no dependence on the dimension of the ambient space that the examples lie in.

For example, the classic Perceptron Convergence Theorem states that the Perceptron algorithm makes at most $1/\gamma^2$ mistakes on any sequence of examples separable by margin γ and hence if the Perceptron algorithm is run on a sample of size $1/(\epsilon\gamma^2)$, the expected error rate of its hypothesis at a random point in time is at most ϵ.

As mentioned in [B05], one can view a kernel function as implicitly mapping points into a higher-dimensional ϕ-space where one does not pay a high price for high-dimensional computations, and standard margin bounds imply that if the learning problem has a large margin linear separator in that space, then one can avoid paying a high price in terms of sample size as well.

We can combine kernel functions with the JL-Lemma to note that if a learning problem indeed has a large margin γ under the kernel $K(x, y) = \phi(x).\phi(y)$, then a random linear projection of the ϕ-space down to a low-dimensional space approximately preserves linear separability. Hence, all such kernels K can, in principle, be thought of as mapping the input space X into a small $O(1/\gamma^2)$-dimensional space.

Often, using the so-called *kernel trick*, we can perform tractable computations in the ϕ-space without representating the features in that space. Suppose, however, that given a kernel K under which the data is separable, we want to list out an explicit set of transformed features (perhaps to run non-kernelizable algorithms). We will now define a feature mapping using kernel K and random projections that will give us a low-dimensional representations of the points with the good separability properties of kernel K.

Specifically, given black box access to kernel K, access to new unlabeled examples from distribution D, and parameters γ, ϵ, δ, [BBV04] construct a mapping $F : X \to \mathcal{R}^d$ in polytime, such that if the target concept does have margin γ with error α in the ϕ-space, then with probability at least $1 - \delta$ (over the random draws from D) the induced distribution in $\mathcal{R}^d$ is separable with margin $\Omega(\gamma)$ with error $\leq \alpha + \epsilon$ when $d = O\left(\frac{1}{\gamma^2} \log \frac{1}{\epsilon\delta}\right)$. We shall start with a key lemma.

Existence Lemma Consider any distribution over labeled examples in euclidean space such that there exists a vector w with margin γ. If we draw $d \geq O\left(\frac{1}{\epsilon}\left[\frac{1}{\gamma^2} + \log\frac{1}{\delta}\right]\right)$ examples $z_1, ..., z_d$ iid from this distribution, then with probability $\geq 1-\delta$ there exists a vector w' in $span(z_1, ..., z_d)$ that has true error at most ϵ at margin $\gamma/2$.

There are sample complexity bounds that state that $|S| = O(\frac{1}{\epsilon\gamma^2}\log^2(\frac{1}{\epsilon\gamma}) + \frac{1}{\epsilon}\log\frac{1}{\delta})$ points are sufficient so that with high probability, *any* linear separator of S with empirical margin γ over S has low true error. So, to get w' in the in the above lemma, if we draw $|S|$ points from the distribution and the true w was projected into their span, it would maintain the value of $w.z_i$ while possibly shrinking w and increasing its margin over the observed data, thus proving the existence of such a w' with true error at most ϵ. However, we want w' to not only have low true error, but attain the true error with a large margin, and we also want to show an existential statemnt, and not a universal statement about all separators with large empirical margins, and hence can get tighter bounds.

The existence lemma implies that if P_{XY} is linearly separable with margin γ under K, and we draw d random unlabeled examples $x_1, ..., x_d$ from P_X, then with high probability, there is a separator w' in the ϕ-space that can be written as $\alpha_1\phi(x_1) + ... + \alpha_d\phi(x_d)$ with low true error. Since with high probability $w'.\phi(x) = \alpha_1 K(x, x_1) + ... + \alpha_d K(x, x_d)$ will match the sign of x's label, we can think of $K(x, x_i)$ as the i-th feature of x and $(\alpha_1, ..., \alpha_d)$ as the approximate linear separator, giving us:

Mapping F_1 For $d = O\left(\frac{1}{\epsilon}\left[\frac{1}{\gamma^2} + \log\frac{1}{\delta}\right]\right)$, let us draw a set S of d points $x_1, ..., x_d$ from P_X, and consider the mapping $F_1 : X \to \mathcal{R}^d$ defined by $F_1(x) = (K(x, x_1), ..., K(x, x_d))$. Then, if P_{XY} has a margin γ in the ϕ-space ($\phi(P_X)$ is separable by margin γ), with probability $\geq 1-\delta$, the induced distribution $F_1(P_{XY})$ is linearly separable with error $\leq \epsilon$.

Unfortunately, F_1 may not preserve margins because we don't have a good bound on the length of the vector $(\alpha_1, ..., \alpha_d)$ defining the separator in the new space, or the length of the examples $F_1(x)$. The problem is that if many of the $\phi(x_i)$ are very similar, their associated features $K(x, x_i)$ will be highly correlated and to preserve margin, we want to choose an orthonormal basis of the space spanned by the $\phi(x_i)$ and do an orthogonal projection of $\phi(x)$ into this space. Hence we come up with the following mapping.

Mapping F_2 Consider the kernel matrix of S, $\mathcal{K}_{ij} = (K(x_i, x_j))_{x_i, x_j \in S}$ and its cholesky decomposition $\mathcal{K} = U^\top U$ where U is upper-triangular. Define the mapping $F_2 : X \to \mathcal{R}^d$ by $F_2(x) = F_1(x)U^{-1}$. Then, if P_{XY} has a margin γ in the ϕ-space, with probability $\geq 1-\delta$, the induced distribution $F_2(P_{XY})$ is linearly separable with error $\leq \epsilon$ at margin $\gamma/2$.

Notice that we can think of the JL-projection matrix A as choosing d random (rademacher/gaussian) points $r_1, ..., r_d \in \mathcal{R}^n$ and defining the projection of $x \in \mathcal{R}^n$ to be $JL(x) = xA = (x.r_1, ..., x.r_d)$. The final mapping that we propose can be thought of as a two-stage process, which is a combination of two types of random projections - the first uses points chosen at random from P_{XY} and kernels like in F_2, and the second using points chosen uniformly at random in the kernel space to form a random projection matrix like in JL.

Mapping F_3 For $d' = O\left(\frac{1}{\gamma^2}\log\frac{1}{\epsilon\delta}\right)$, consider a random JL-projection A into d' dimensions. Define the mapping $F_3 : X \to \mathcal{R}^{d'}$ by $F_3(x) = F_2(x)A$. Then, if P_{XY} has a margin γ in the ϕ-space, with probability $\geq 1-\delta$, the induced distribution $F_3(P_{XY})$ is linearly separable with error $\leq \epsilon$ at margin $\gamma/4$.

If the distribution P_{XY} is only separable with error α at margin γ, then all the previous results apply to the $1-\alpha$ portion of the distribution that is correctly separated by margin γ and the true error of the resulting mapping is now replaced by $(1-\alpha)\epsilon + \alpha$. One can also extend these results to the case where the target separator does not pass through the origin.

5 Manifold Learning

Interestingly, it was shown in [BW06] that random projections can be used for manifold learning. When a set of points occupies a submanifold of the ambient space, we can randomly project them into a much smaller dimension and preserve many of the important properties of manifolds.

For concreteness, consider a K-dimensional manifold $\mathcal{M}$ in ambient space $\mathcal{R}^N$, that has a condition number $1/\tau$ and volume V. Let Φ be a random orthoprojector from R^N to R^M with $M \geq O(\frac{\log(1/\delta)}{\epsilon^2} K \log(NV\tau^{-1}))$. Then, with probability at least $1-\delta$ for every pair of points $x, y \in \mathcal{M}$ then, $(1-\epsilon)\sqrt{\frac{M}{N}} \leq \frac{d(\Phi x, \Phi y)}{d(x,y)} \leq (1+\epsilon)\sqrt{\frac{M}{N}}$ where d could refer to both the geodesic and the l_2 distance between x and y.

In [HWB08], they argue that not only does it preserve many intrinsic properties of the manifold, but also the results of several classic algorithms are stable under these random projections. For example, for a very similar value for M, they guarantee that the GP algorithm will find (upto a small multiplicative error) the same estimate for the correlation dimension of the space. Similarly, for similar M, they show that the Isomap algorithm will have (upto a small additive error) the same residual variances on generating a K-dimensional embedding of points from the original dataset.

Some proerties that are preserved by this kind of a random projection are ambient and geodesic distances between pairs of points, dimension of the manifold, topology, local neighbourhoods and local angles, length and curvature of paths on the manifold and the volume of the manifold. Hence, it could be an extremely powerful tool for learning high dimensional data, even when it lies on a complex non-linear manifold.

6 Conclusion

The idea of random projections has been around for a while in the theoretical computer science community and is certainly playing an increasingly important role in learning. This survey intended to give the reader a light introduction to the topic, and perhaps show how this simple concept is so intricately related to many other concepts in learning (and probably many more to come!).

Many other uses of the idea of random projections is available in Vempala's survey book [V04] on the same topic. The inquisitive reader is also referred to my other survey connecting compressed sensing, group testing and matrix completion, all of which are strongly connected with random projections and the JL Lemma.

Randomness is here to stay.
With randomness we will play.

7 References

[B05] Blum (2005) Random Projections, Margins, Kernels and Feature Selection *SLSFS 2005*

[BBV04] Balcan, Blum & Vempala (2004) Kernels as Features: On Kernels, Margins and Low-dimensional Mappings *ALT 2004*

[BW06] Baraniuk & Wakin (2006) Random Projections of Smooth Manifolds *FoCM 2006*

[HWB08] Hegde, Wakin & Baraniuk (2008) Random Projections for Manifold Learning *NIPS 2008*

[V04] Vempala (2004) The Random Projection Method *DIMACS Series in Discrete Math*

[GWL] Google, Wikipedia & Lecture-notes