Lecture 8: Dimensionality Reduction

Iain Styles

5 November 2019

Reducing the Dimensionality of MNIST

One problem with the analysis we have done so far is that we have considered points that are distributed throughout a hypervolume. "Real" data does not behave in this way; it tends to lie on some low-dimensional subspace. Trivially, in MNIST, some pixels have the same value in all images, as shown in Figure 1, which shows the mean and standard deviation of 1000 MNIST images, together with a mask that shows which pixels vary, and which do not. 175 of the 784 pixels in the image, which correspond to dimensions in the underlying vector space, have zero variance. In this case they also have zero mean, but that is less important. This means that we could, trivially, reduce the dimensionality of the data to 609, by just retaining those pixels which vary and therefore contain useful information.

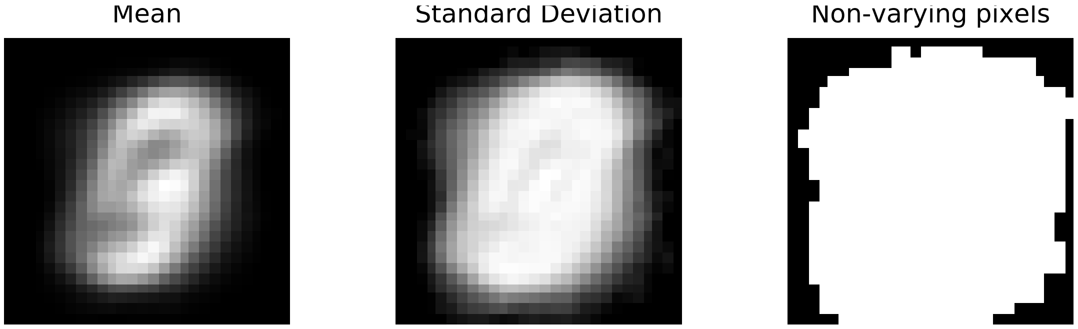

Figure 1: Image statistics from 1000 MNIST images. *Left*: The pixel-mean. *Centre*: the pixel standard deviation. *Right*: Black pixels are those with zero variance.

In Figure 2 we show the cumulative sum of the pixel variances, and see that nearly all of the variance is from 500 pixels, and around 75% of it is from only half of the pixels, suggesting that further reductions in dimensionality may be possible. In fact, by looking at each pixel in isolation, we are only scratching the surface of what is possible in terms of reducing the dimension of the data in a meaningful way. One easy-to-picture reason why we might be able to reduce the dimensionality further is that correlations between pixels mean that the value of one pixel can be predicted from the value of another, but more complex situations are possible. For instance, imagine that you have a picture or a poster that is rolled up. When unrolled, points on the poster can be described quite adequately in a 2d coordinate system. However, when the poster is rolled up, it becomes a three-dimensional object that requires a 3d coordinate system. If we could find a transformation that could "unroll" the poster, then we could work with it in 2d coordinates. This argument can be extended to higher dimensions, and the task of *dimensionality reduction* is to find the transformations that allows to represent data using their *intrinsic dimensionality*: the number of meaningful degree of freedom.

To understand what this means in terms of MNIST, let us consider what the underlying degrees of freedom are.

- We have ten digit class.
- Members of each class may vary by width and height (although this is somewhat mitigated by the alignment process.)
- Each character class will have some shape variation – for example, whether one's have a serif and/or a base (eg 1 vs 1).

An order-of-magnitude estimate suggests that there may be a few ten's of intrinsics degree of freedom in the data. The question now is, how do we find a transformation that extracts them?

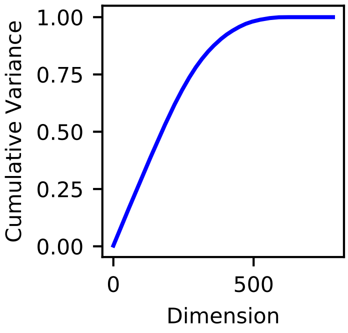

Figure 2: Cumulative variance as a function of dimensionality. The pixels were ordered by variance and the cumulative sum was calculated. This shows that around 75% of the variance in the data is from fewer than half of the pixels.

Dimensionality Reduction

We will limit our discussion to *linear* methods for dimensionality reduction; that is, methods that apply a single global linear transformation (rotation, scale, shear etc) to the data. To understand what this might do, imagine a sheet of paper embedded in 3d. If we can rotate and shift the sheet so that it aligns with the 3d coordinate system (x, y, z), then we can discard the coordinate that aligns with the thinnest dimension of the sheet of paper (its thickness), and just retain those coordinates that describe the main variations in the sheet: we can reduce the problem from 3d to 2d. We will be trying to do the same here: to shift and rotate the data to find a coordinate system that describes the main variation whilst preserving the internal structure of the data. Our hope is that by reducing the dimensionality of the data in a way that preserves its internal structure, we can mitigate some of the issues that we see in high dimensional spaces.

There are many ways in which the dimensionality of a dataset can be reduced; we will focus on the random*random projection*. This is an attractive method because it is computationally very simple and low-cost, but there is some elegant theory behind it that provides much insight into the problems that high dimensionality causes.

Random Projections

We have already seen random projections in action, and have seen, experimentally, how much benefit they can bring to a high-dimensional learning problem. The central idea is exceptionally simple. Given an original dataset containing N samples in M dimensional space, arranged as an $M \times N$ matrix $\mathbf{X}$ (one sample per column):

- Generate K random vectors with M components randomly sampled from $\mathcal{N}(0,1)$. Arrange these as a matrix $\mathbf{R}$ of size $M \times K$, with one vector per column.
- Normalise the columns of $\mathbf{R}$ so each has unit length.
- Compute $\mathbf{X}' = \mathbf{R}^{\mathrm{T}}\mathbf{X}$
- The columns of $\mathbf{X}'$ contain the samples projected (Figure 3) into the lower dimensional space define by the K random vectors.

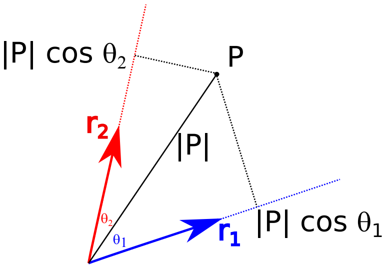

Figure 3: Projection of a point $\mathbf{P}$ onto two unit random vectors $\mathbf{r}_1$ and $\mathbf{r}_2$. Noting that $\mathbf{r}_1$ and $\mathbf{r}_2$ are unit vectors, the projection is given by the dot product $\mathbf{P} \cdot \mathbf{r} = |P| \cos\theta$.

We saw, in k-nn classification, that this works. But why does it work? The key idea is that whilst distances in high-dimensional spaces are of limited use, *if* we could could map the the points into some low-dimensional space (where distances are more useful) *in a way that preserves the local structure of the data*, then we can render distances useful again.

Is it possible to find such a map? It turns out that random projections, subject to a few constraints, can be be shown to generate just such a mapping. The key to this is a mathematical result known as the *Johnson-Lindenstrauss lemma*[1]. This is a statement that points in a high-dimensional space can be mapped onto a low-dimensional space in such a way that relative distances between points are preserved, whilst the absolute distances are reduced. The lemma states that:

Given a set X of N data points in $\mathbb{R}^M$ (M-dimensional real space), there is a linear map $f : \mathbb{R}^M \mapsto \mathbb{R}^K$ that maps X onto a *random* lower-dimensional space $\mathbb{R}^K$. The map obeys

$$(1-\varepsilon)\|x_1 - x_2\|^2 \leq \|f(x_1) - f(x_2)\|^2 \leq (1+\varepsilon)\|x_1 - x_2\|^2 \quad (1)$$

for all $x_1, x_2 \in X$ and for $0 < \varepsilon < 1$ and $K > 8\ln(N)/\varepsilon^2$. It has also been proven [2] that this holds when the function f is a random, orthonormal linear transformation. A proof of this result is far beyond the scope of this course. Let us try to unpick its implications.

Firstly, we need to be clear of the central result here: Johnson-Lindenstrauss and its extension by Frankl and Maehara together state that a "random, orthonormal linear transformation" can be used to project a set of points onto a lower-dimensional subspace whilst preserving the distances between them to within some bounds controlled by ϵ: distances between points are scaled by a factor between $1-\varepsilon$ and $1+\varepsilon$. Noting the condition that $K > 8\ln(N)/\varepsilon^2$, this means that smaller values of ε (better distance preservation) requires higher values of K.

One now comes across yet another remarkable property of high-dimensional spaces. The requirement that the the random vectors be orthormal – $\mathbf{r}_i \cdot \mathbf{r}_j = 1$ if i=j else 0 – seems like it could be rather onerous. There are well-established methods for doing this (such as Gramm-Schmidt orthogonalisation), but they carry a computational cost. Fortunately, high dimensionality saves us and it turns out that an explicit normalisation step is often not needed, because in very high-dimensional spaces, the probability that two random vectors

[1] Johnson, William B., and Joram Lindenstrauss. "Extensions of Lipschitz mappings into a Hilbert space." Contemporary mathematics 26.189-206 (1984): 1.

[2] Frankl, Peter, and Hiroshi Maehara. "The Johnson-Lindenstrauss lemma and the sphericity of some graphs." Journal of Combinatorial Theory, Series B 44.3 (1988): 355-362.

will be orthogonal tends towards 1! We demonstrate this numerically in Figure 4, noting that because $\mathbf{x} \cdot \mathbf{y} = |x||y| \cos\theta$, $\mathbf{x} \cdot \mathbf{y} = 0$ means that the $\cos\theta = 0$ corresponds to an angle of $90°$ which means that $\mathbf{x}$ and $\mathbf{y}$ are orthogonal. These results demonstrate that as the dimensionality increases, the probability that two random vectors are are *not* orthogonal tends to zero. This is again a consequence of the dimensionality, and it can be shown that the number of sets of nearly orthonal vectors is exponential in the dimensionality[3]. If the dimensionality of our problem is high enough, random vectors are self-orthogonalising!

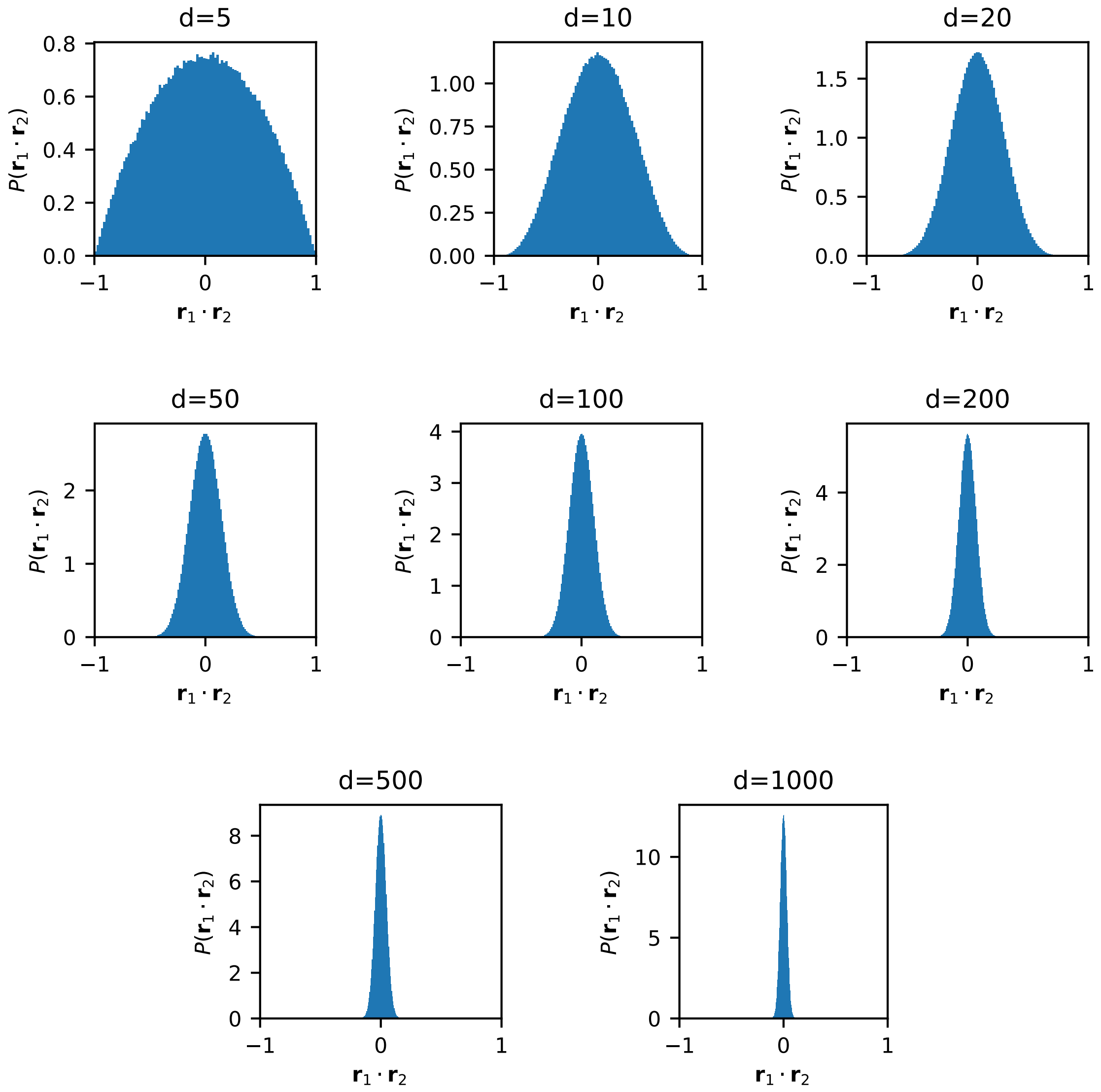

Figure 4: $P(\mathbf{r}_i \cdot \mathbf{r}_j | i \neq j)$ from 100,000 normalised random vectors in different dimensional spaces.

Finally, let us investigate the effect that random projection has on the pairwise distances. We compute all pairwise distances on the data projected on to forty random vectors, with distances computed in the lower-dimensional space. The results are shown in Figure 5. The mean/median of the distribution has been reduced to ≈ 580 from ≈ 2300, and the standard deviation to ≈ 100 from ≈ 300. This

[3] Kainen, Paul C.; Kůrková, Věra (1993), "Quasiorthogonal dimension of Euclidean spaces", Applied Mathematics Letters, 6 (3): 7âĂŞ10, MR

does not seems entirely consistent with the Johnson-Lindenstrauss lemma, which stated that distances would be preserved. Why is this?

Although the main implications of Johnson-Lindenstrauss seem clear, there is a sublety. One would expect dimensionality reduction to *reduce* the distances between points, not to preserve them, but J-L seems to say the opposite. However, it is perhaps more helpful to rewrite the theorem as

$$1 - \varepsilon \leq \frac{\|f(x_1) - f(x_2)\|^2}{\|x_1 - x_2\|} \leq 1 + \varepsilon \tag{2}$$

which allows us to see that J-L is in fact a theorem about the *relative* distances between points: it states that it is the relative distances that are preserved, whilst the absolute distances become reduced due to dimensional scaling. This is what makes it such a powerful result for machine learning: we can use random projections to reduce the dimensionality and overcome the "curse" at a very low computational cost, whilst preserving the structure (relative distances) within the data. This is why we see such a dramatic performance improvement in k-nn classification on MNIST digits.

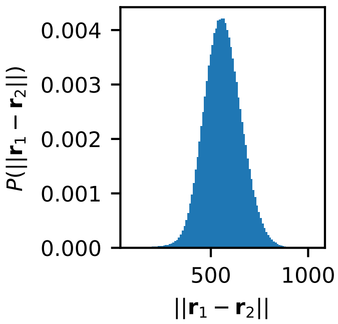

Figure 5: Distribution of pairwise distances following projection onto forty random vectors in pixel space between 1000 examples from the MNIST test set, and 1000 examples from the training set.