Lecture 7: The Curse of Dimensionality

Attendance code: EJZSDPUN

Iain Styles

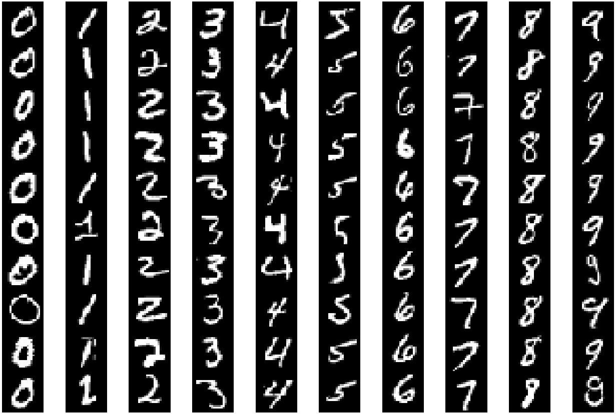
1 November 2018

Learning Outcomes

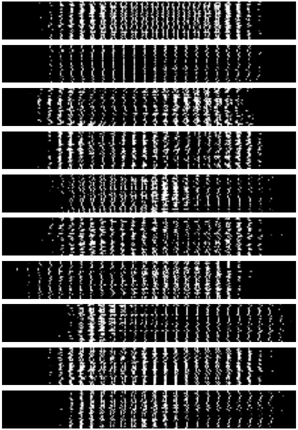
By the end of this lecture you should:

- Know how well naïve knn classsification performs on MNIST
- Know what effect reducing the dimensionality of the data has
- Understand and explain some of the properties of high dimensional spaces
- Be able to explain why they are important for learning

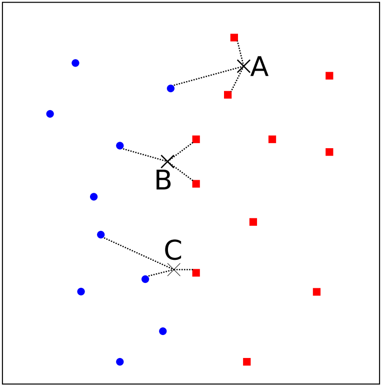
Recap: Classification



Vectorised MNIST



k nearest-neighbours Classification



$k$nn and MNIST

- Do we expect $k$nn to do well on MIST?

$k$nn and MNIST

- Do we expect $k$nn to do well on MIST?
- Vectorising the images loses much of their spatial information
- There is substantial variability between characters

$k$nn and MNIST

- Do we expect $k$nn to do well on MIST?
- Vectorising the images loses much of their spatial information
- There is substantial variability between characters
- No harm in trying. . .
- Need a measure of similarity: Euclidean distance
  For images vectors $\mathbf{x}$ and $\mathbf{y}$

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{((\mathbf{x} - \mathbf{y})^{\mathrm{T}} (\mathbf{x} - \mathbf{y}))} = \sqrt{\sum_i (x_i - y_i)^2}. \qquad (1)$$

- Smaller $\rightarrow$ more similar
- Use 10,000 training samples and 1000 test samples to save time

$k = 1$ nearest-neighbours

T \ P	0	1	2	3	4	5	6	7	8	9
0	83	1	1	0	0	0	5	0	10	0
1	0	100	0	0	0	0	0	0	0	0
2	1	11	53	2	1	0	3	4	25	0
3	0	11	2	48	0	1	4	3	28	3
4	2	9	0	0	42	0	2	3	16	26
5	2	7	0	4	0	36	2	0	43	6
6	3	6	0	0	0	1	80	0	10	0
7	0	11	0	1	0	0	1	75	4	8
8	2	13	0	6	1	3	3	4	65	3
9	0	5	1	1	4	0	0	4	2	83

$k = 1$ nearest-neighbours

T \ P	0	1	2	3	4	5	6	7	8	9
0	83	1	1	0	0	0	5	0	10	0
1	0	100	0	0	0	0	0	0	0	0
2	1	11	53	2	1	0	3	4	25	0
3	0	11	2	48	0	1	4	3	28	3
4	2	9	0	0	42	0	2	3	16	26
5	2	7	0	4	0	36	2	0	43	6
6	3	6	0	0	0	1	80	0	10	0
7	0	11	0	1	0	0	1	75	4	8
8	2	13	0	6	1	3	3	4	65	3
9	0	5	1	1	4	0	0	4	2	83

- Total accuracy: 67%

$k = 3$ nearest-neighbours

T \ P	0	1	2	3	4	5	6	7	8	9
0	95	1	0	0	0	0	1	0	3	0
1	0	100	0	0	0	0	0	0	0	0
2	4	14	68	0	0	0	1	2	11	0
3	2	13	4	64	0	1	3	2	8	3
4	2	13	1	0	51	0	4	2	3	24
5	5	13	0	10	1	39	2	0	24	6
6	2	7	0	0	1	1	88	0	1	0
7	0	18	2	1	1	1	0	68	3	6
8	3	18	0	3	1	3	3	4	65	0
9	1	7	0	1	1	0	0	2	2	86

- Total accuracy: 72%

$k = 5$ nearest-neighbours

T \ P	0	1	2	3	4	5	6	7	8	9
0	97	1	0	0	0	0	1	0	1	0
1	0	100	0	0	0	0	0	0	0	0
2	3	17	69	1	0	0	2	3	5	0
3	1	19	1	60	0	0	6	3	7	3
4	2	12	1	0	50	0	5	1	4	25
5	5	9	0	5	2	51	2	0	19	7
6	2	7	0	0	1	1	89	0	0	0
7	0	18	0	0	1	1	0	73	2	5
8	3	18	1	3	0	1	4	5	65	0
9	1	9	0	0	0	0	0	1	4	85

- Total accuracy: 74%

$k = 7$ nearest-neighbours

T \ P	0	1	2	3	4	5	6	7	8	9
0	95	1	0	0	0	0	3	0	1	0
1	0	100	0	0	0	0	0	0	0	0
2	1	17	70	0	0	0	2	4	6	0
3	1	20	0	61	0	1	6	2	5	4
4	3	9	0	0	55	0	4	1	2	26
5	5	9	1	5	1	51	3	0	17	8
6	2	7	0	0	0	1	90	0	0	0
7	1	17	1	0	1	0	0	75	1	4
8	3	16	1	2	0	1	4	5	66	2
9	1	7	0	0	0	0	0	2	2	88

- Total accuracy: 75%

Making it even better

- Consensus voting over k neighbours bring significant gains over single nearest neighbours

Making it even better

- Consensus voting over k neighbours bring significant gains over single nearest neighbours
- $k = 7$: diminishing returns?

Making it even better

- Consensus voting over k neighbours bring significant gains over single nearest neighbours
- $k = 7$: diminishing returns?
- 0, 1, 6, and 9 can be identified very accurately
- 3, 4, and 5 much more resistant

Making it even better

- Consensus voting over k neighbours bring significant gains over single nearest neighbours
- $k = 7$: diminishing returns?
- 0, 1, 6, and 9 can be identified very accurately
- 3, 4, and 5 much more resistant
- How can we improve this?
- Change the similarity metric

Making it even better

- Consensus voting over k neighbours bring significant gains over single nearest neighbours
- $k = 7$: diminishing returns?
- 0, 1, 6, and 9 can be identified very accurately
- 3, 4, and 5 much more resistant
- How can we improve this?
- Change the similarity metric – learn it from the data

Making it even better

- Consensus voting over k neighbours bring significant gains over single nearest neighbours
- $k = 7$: diminishing returns?
- 0, 1, 6, and 9 can be identified very accurately
- 3, 4, and 5 much more resistant
- How can we improve this?
- Change the similarity metric – learn it from the data
- Reduce the dimensionality

Making it even better

- Consensus voting over k neighbours bring significant gains over single nearest neighbours
- $k = 7$: diminishing returns?
- 0, 1, 6, and 9 can be identified very accurately
- 3, 4, and 5 much more resistant
- How can we improve this?
- Change the similarity metric – learn it from the data
- Reduce the dimensionality – high-dimensional vector space do not behave the same way as low-dimensionality spaces

Making it even better

- Consensus voting over k neighbours bring significant gains over single nearest neighbours
- $k = 7$: diminishing returns?
- 0, 1, 6, and 9 can be identified very accurately
- 3, 4, and 5 much more resistant
- How can we improve this?
- Change the similarity metric – learn it from the data
- Reduce the dimensionality – high-dimensional vector space do not behave the same way as low-dimensionality spaces

Making it even better

- Take each image vector (784-element column vector)
- Take scalar (dot) product with each of 40 random 784-element vectors.
- Replace each sample with the resulting 40-element vector

Making it even better

- Take each image vector (784-element column vector)
- Take scalar (dot) product with each of 40 random 784-element vectors.
- Replace each sample with the resulting 40-element vector

$$\begin{pmatrix} \mathbf{z}_1 & \mathbf{z}_2 & \dots & \mathbf{z}_N \end{pmatrix} = \begin{pmatrix} \mathbf{r}_1^{\mathrm{T}} \\ \mathbf{r}_2^{\mathrm{T}} \\ \dots \\ \mathbf{r}_M^{\mathrm{T}} \end{pmatrix} \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_N \end{pmatrix}. \tag{2}$$

- Form new training and test sets: 10000 and 1000 40-element vectors.
- Use k-nn to classify the training set.

$k = 7$ nearest-neighbours, 40 random projections

T \ P	0	1	2	3	4	5	6	7	8	9
0	98	0	0	0	0	1	0	0	1	0
1	0	100	0	0	0	0	0	0	0	0
2	3	4	79	1	0	1	4	3	5	0
3	0	4	2	84	0	1	1	3	2	3
4	0	1	0	0	85	0	1	2	2	9
5	0	2	0	3	1	86	4	2	1	1
6	1	0	0	0	1	6	91	1	0	0
7	0	3	1	1	1	0	0	91	0	3
8	1	0	5	13	3	4	1	2	70	1
9	0	1	0	0	6	0	0	2	2	89

- Total accuracy: 87%!!!

The Curse of Dimensionality

- Why did projecting the data onto 40 random vectors improve classification?

The Curse of Dimensionality

- Why did projecting the data onto 40 random vectors improve classification?
- High-dimensional spaces have weird properties.

The Curse of Dimensionality

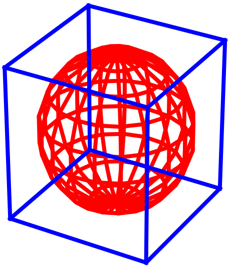
- Why did projecting the data onto 40 random vectors improve classification?
- High-dimensional spaces have weird properties.
- Some examples. . .

Hyperspheres inside hypercubes

- Hypercube: n-dimensional analogue of cube
- In each dimension, the cube has a side of length $2r$ such that the centre of each face is a distance r from the centre

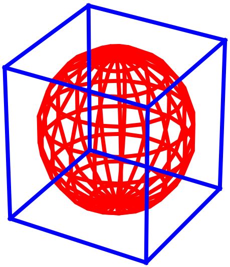
Hyperspheres inside hypercubes

- Hypercube: *n*-dimensional analogue of cube
- In each dimension, the cube has a side of length $2r$ such that the centre of each face is a distance r from the centre
- Hypercube encloses a *hypersphere* of radius r, defined as the set of points a distance r from its centre.
- Hypersphere intersects hypercube in the centre of its faces.



Hyperspheres inside hypercubes

- Hypercube: n-dimensional analogue of cube
- In each dimension, the cube has a side of length $2r$ such that the centre of each face is a distance r from the centre
- Hypercube encloses a *hypersphere* of radius r, defined as the set of points a distance r from its centre.
- Hypersphere intersects hypercube in the centre of its faces.



- How are the volume of the cube and the sphere related?

Where are the corners of the hypercube?

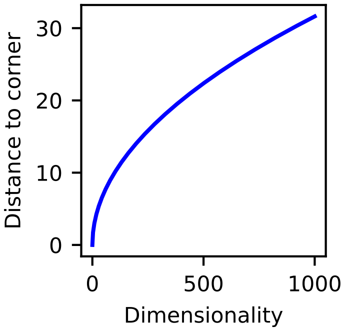
- In 2d, the corners of a square are $\sqrt{2}r \approx 1.41r$ from the centre.

Where are the corners of the hypercube?

- In 2d, the corners of a square are $\sqrt{2}r \approx 1.41r$ from the centre.
- In 3d, the corners are $\sqrt{3}r \approx 1.73r$ from the centre.

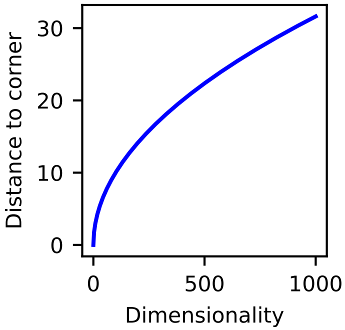
Where are the corners of the hypercube?

- In 2d, the corners of a square are $\sqrt{2}r \approx 1.41r$ from the centre.
- In 3d, the corners are $\sqrt{3}r \approx 1.73r$ from the centre.
- 4d: $2r$, 5d: $2.23r$ etc.
- In general, corners of a hypercube are $r\sqrt{n}$ from its centre



Where are the corners of the hypercube?

- In 2d, the corners of a square are $\sqrt{2}r \approx 1.41r$ from the centre.
- In 3d, the corners are $\sqrt{3}r \approx 1.73r$ from the centre.
- 4d: $2r$, 5d: $2.23r$ etc.
- In general, corners of a hypercube are $r\sqrt{n}$ from its centre



- In $d = 1000$, the corners of the hypercube are more than 30 times further out than the hypersphere it encloses.

How much volume does the sphere occupy?

- 2d: square: $4r^2$, circle: πr^2; ratio $\pi/4 \approx 0.785$

How much volume does the sphere occupy?

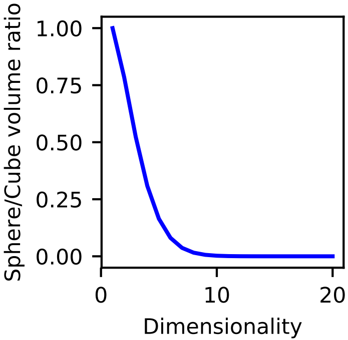
- 2d: square: $4r^2$, circle: πr^2; ratio $\pi/4 \approx 0.785$
- 3d: cube: $8r^3$, sphere: $4\pi r^3/3$; ratio $4\pi/24 \approx 0.52$
- Only half the cube is in the sphere!

How much volume does the sphere occupy?

- 2d: square: $4r^2$, circle: πr^2; ratio $\pi/4 \approx 0.785$
- 3d: cube: $8r^3$, sphere: $4\pi r^3/3$; ratio $4\pi/24 \approx 0.52$
- Only half the cube is in the sphere!
- n-d cube: $V = (2r)^n$; n-d sphere: $V = \frac{\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}+1\right)} R^n$, where $\Gamma(x)$ is the Gamma function.

How much volume does the sphere occupy?

- 2d: square: $4r^2$, circle: πr^2; ratio $\pi/4 \approx 0.785$
- 3d: cube: $8r^3$, sphere: $4\pi r^3/3$; ratio $4\pi/24 \approx 0.52$
- Only half the cube is in the sphere!
- n-d cube: $V = (2r)^n$; n-d sphere: $V = \frac{\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}+1\right)} R^n$, where $\Gamma(x)$ is the Gamma function.



Where is the volume in a hypersphere?

- Consider two hyperspheres with the same centre, one of radius r, the other of radius $r - \delta$
- Their volumes are $\alpha_n r^n$ and $\alpha_n (r - \delta)^n$ respectively
- The difference between them is a thing "shell" of thickness δ.

Where is the volume in a hypersphere?

- Consider two hyperspheres with the same centre, one of radius r, the other of radius $r - \delta$
- Their volumes are $\alpha_n r^n$ and $\alpha_n (r-\delta)^n$ respectively
- The difference between them is a thing "shell" of thickness δ.
- As a proportion of the larger shell the shell has volume

$$
\begin{aligned}
\frac{V_{\text{shell}}}{V_{\text{sphere}}} &= \frac{\alpha\left(r^n - (r-\delta)^n\right)}{\alpha r^n} && (3)\\
&= 1 - r^{-n}(r-\delta)^n && (4)\\
&= 1 - \left(r^{-1}(r-\delta)\right)^n && (5)\\
&= 1 - \left(1 - \frac{\delta}{r}\right)^n && (6)
\end{aligned}
$$

Where is the volume in a hypersphere?

- Consider two hyperspheres with the same centre, one of radius r, the other of radius $r - \delta$
- Their volumes are $\alpha_n r^n$ and $\alpha_n (r-\delta)^n$ respectively
- The difference between them is a thing "shell" of thickness δ.
- As a proportion of the larger shell the shell has volume

$$\frac{V_{\text{shell}}}{V_{\text{sphere}}} = \frac{\alpha\left(r^n - (r-\delta)^n\right)}{\alpha r^n} \tag{3}$$
$$= 1 - r^{-n}(r-\delta)^n \tag{4}$$
$$= 1 - \left(r^{-1}(r-\delta)\right)^n \tag{5}$$
$$= 1 - \left(1 - \frac{\delta}{r}\right)^n \tag{6}$$

- In the limit $n \to \infty$, this tends to 1
- The volume is concentrated in the shell.

Why is this relevant?

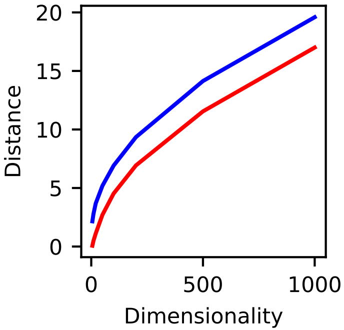
- The same phenomena affect pairwise distances
- Let's do an experiment

Why is this relevant?

- The same phenomena affect pairwise distances
- Let's do an experiment
- Generate 10^6 uniformly randomly distributed data points and compute the distances between all pairs of points.
- What are the min/max pairwise distances?

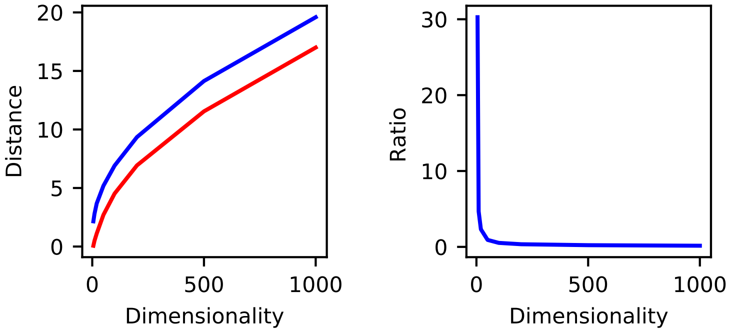
Why is this relevant?

- The same phenomena affect pairwise distances
- Let's do an experiment
- Generate 10^6 uniformly randomly distributed data points and compute the distances between all pairs of points.
- What are the min/max pairwise distances?



Why is this relevant?

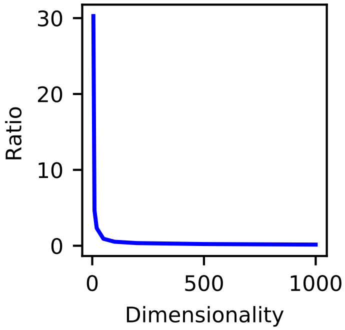
- The same phenomena affect pairwise distances
- Let's do an experiment
- Generate 10^6 uniformly randomly distributed data points and compute the distances between all pairs of points.
- What are the min/max pairwise distances?



A General Result

- Empirical verification of a well-known result

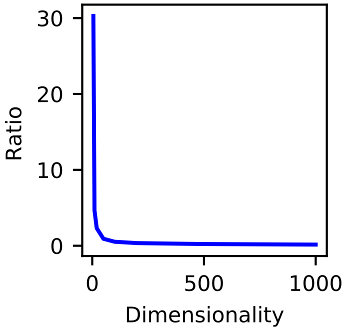
$$\lim_{n\to\infty} \mathbb{E}\left(\frac{d_{\max} - d_{\min}}{d_{\min}}\right) \to 0 \tag{7}$$



A General Result

- Empirical verification of a well-known result

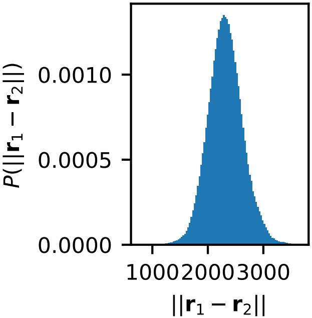
$$\lim_{n \to \infty} \mathbb{E}\left(\frac{d_{\text{max}} - d_{\text{min}}}{d_{\text{min}}}\right) \to 0 \tag{7}$$



- To what extent is it relevant to MNIST

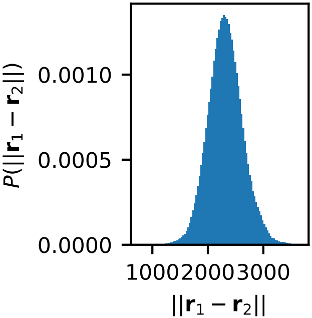
Distances in MNIST

- 1000 points from the test set and 1000 points from the training set



Distances in MNIST

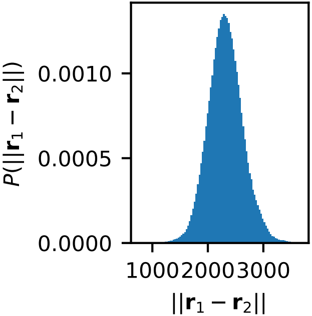
- 1000 points from the test set and 1000 points from the training set



- Mean/median of $\approx$ 2300 and a standard deviation of $\approx$ 300.
- 68% of pairwise distances lie between 2000 and 2600, and 95% between 1700 and 2900.

Distances in MNIST

- 1000 points from the test set and 1000 points from the training set



- Mean/median of ≈ 2300 and a standard deviation of ≈ 300.
- 68% of pairwise distances lie between 2000 and 2600, and 95% between 1700 and 2900.
- Not as "bad" as we might expect? Why?