

Wstęp do uczenia maszynowego

Redukcja wymiaru

Ewa Szczurek + BW (modyfikacje)

bartek@mimuw.edu.pl
Instytut Informatyki
Uniwersytet Warszawski



UNIWERSYTET
WARSZAWSKI

WYDZIAŁ
MATEMATYCZNY, INFORMATYKI I MECHANIKI
MM

Do czego służą składowe główne?

- **Uczenie bez nadzoru**

- Nie mamy żadnej zmiennej objaśnianej
- Wyjaśniamy zmienność cech X_1, \dots, X_p .

- **Wizualizacja:**

- Mamy n obserwacji używających p cech: X_1, \dots, X_p .
- Każda z tych obserwacji to wektor w p -wymiarowej przestrzeni.
- Sposób na efektywną reprezentację danych w mniejszym wymiarze - taką, która oddaje jak najwięcej zmienności w danych

- **Uczenie pod nadzorem:**

- Zastosowanie składowych głównych jako predyktorów w regresji składowych głównych
- Redukcja wymiaru modelu

- PCA

- t-sne

Analiza składowych głównych (ang. *principal components analysis*)

- procedura wyznaczania składowych głównych
- użycie składowych głównych do interpretacji danych

Dane USArrests: ilustracja analizy składowych głównych

- Dla każdego z 50 stanów USA mamy
 - cechy 'Assault', 'Murder', 'Rape': liczba aresztowań na 100 000 mieszkańców dla trzech rodzajów zbrodni: napad, morderstwo, gwałt.
 - cecha 'UrbanPop': procent mieszkańców mieszkających w miastach.
- Zatem mamy tu: $n = 50$ oraz $p = 4$.
- Każda z czterech cech została poddana procedurze standaryzacji, tak aby średnia wynosiła 0, a standardowe odchylenie 1.

Ładunki pierwszej składowej głównej dla USArests

	PC1	PC2
Murder	0.5358995	-0.4181809
Assault	0.5831836	-0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

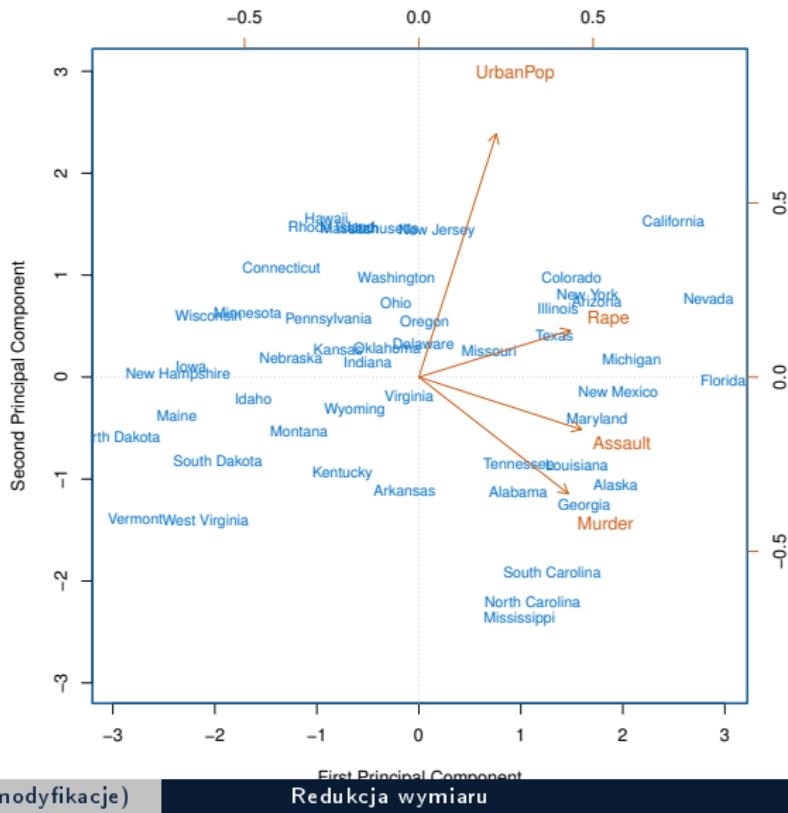
- Pierwsza główna składowa ustawia porównywalne ładunki dla poważnych zbrodni (Murder, Assault, Rape).
- Oznacza to, że wysoki poziom popełnienia przestępstwa w jednej z kategorii zbrodni zwykle oznacza też wysoki poziom w pozostałych dwóch kategoriach.
- Mniejszy ładunek jest przypisany zmiennej UrbanPop. Pierwsza składowa oddaje **poziom przestępcości**.

Ładunki drugiej składowej głównej dla USArrests

	PC1	PC2
Murder	0.5358995	-0.4181809
Assault	0.5831836	-0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

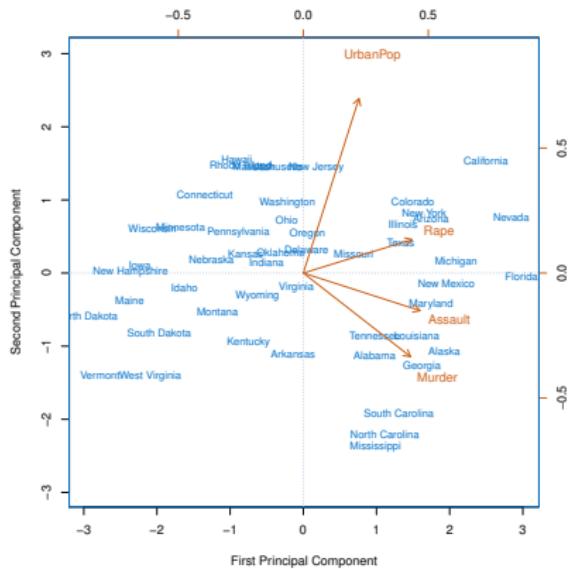
- Druga główna składowa największy ładunek przypisuje zmiennej UrbanPop.
- Zatem druga główna składowa oddaje **poziom urbanizacji stanu**.

Ilustracja graficzna dla pierwszych dwóch składowych (dwa wykresy w jednym rysunku)



Interpretacja rysunku - niebieskie oznaczenia

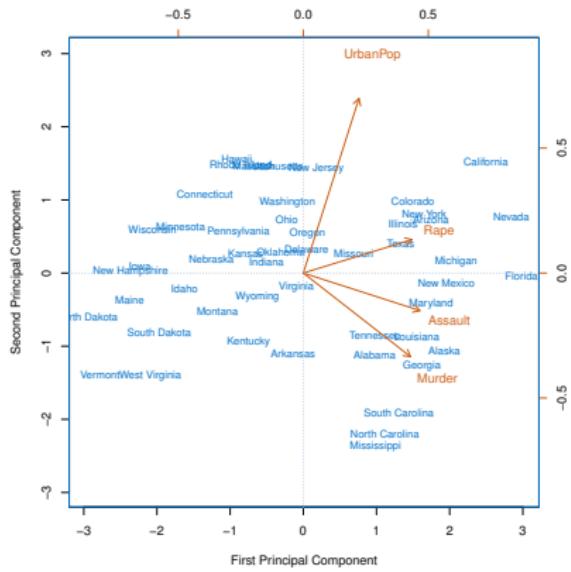
Na **niebiesko** zaznaczone są wyniki zapisów dla pierwszej i drugiej składowej dla każdego stanu.



- Duże dodatnie wartości pierwszej współrzędnej (California, Nevada, Florida) oznaczają wysoki poziom przestępcości w tych stanach.
- Takie stany jak North Dakota mają niski poziom przestępcości.
- Duże dodatnie wartości drugiej współrzędnej (Hawaje, California) oznaczają wysoki poziom urbanizacji.
- Stany położone w pobliżu początku układu (np. Indiana) mają poziomy przestępcości i urbanizacji zbliżone do średniej.

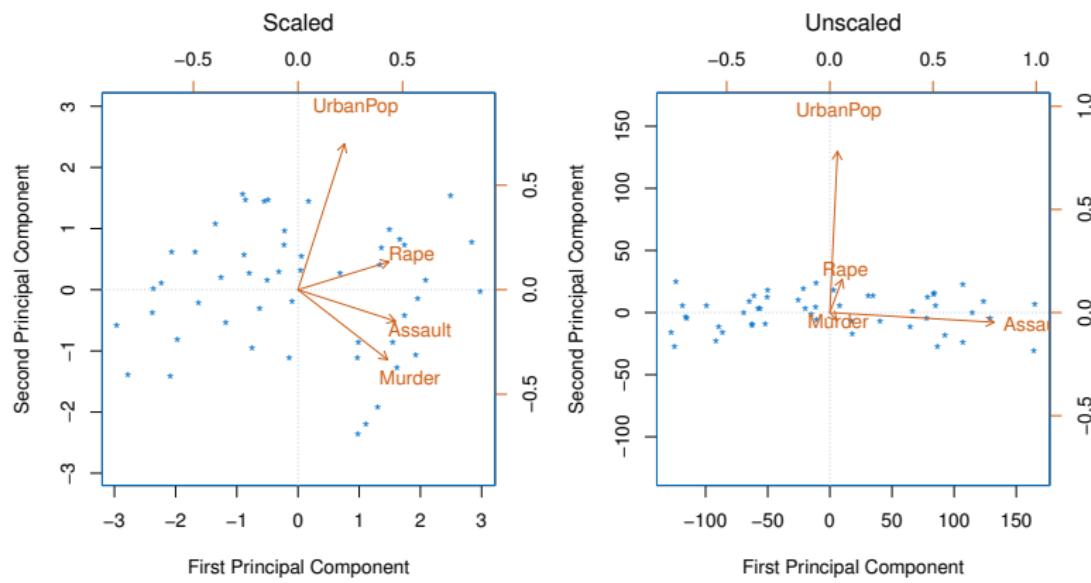
Interpretacja rysunku - pomarańczowe oznaczenia

Na pomarańczowo zaznaczone są pary ładunków pierwszych dwóch składowych głównych dla poszczególnych cech (kierunki).



- cechy odpowiadające rodzajom przestępstw są położone bliżej siebie, a UrbanPop jest położona dalej od pozostałych.

Wpływ skalowania na wynik PCA

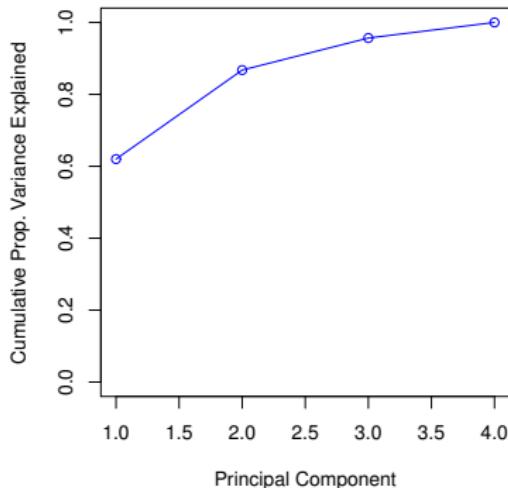
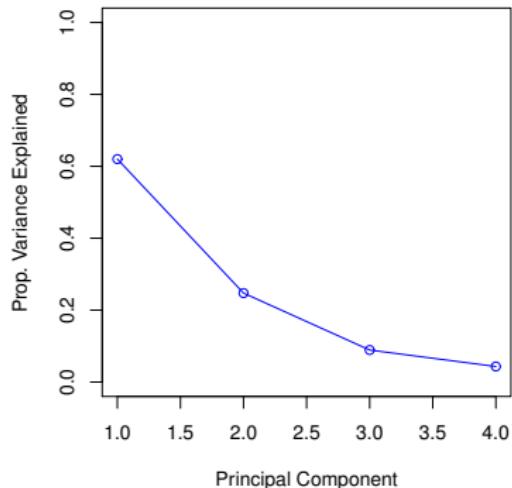


Proporcja wyjaśnionej wariancji (PVE) dla USArrests

- Dla danych USArrests pierwsza składowa główna wyjaśnia 62% wariancji, a druga 24.7%.
- Zatem łącznie pierwsze dwie składowe główne wyjaśniają prawie 87% wariancji.
- Pozostałe dwie składowe główne wyjaśniają tylko nieco ponad 13% wariancji w danych.
- Oznacza to, że ograniczenie się do tylko dwóch pierwszych składowych głównych ilustruje dane całkiem dobrze.

Wykres PVE dla danych USArrests

Takie wykresy zwykle są używane do zdecydowania ile wziąć pierwszych składowych głównych do reprezentowania danych.



Redukcja wymiaru poprzez PCA

- Często przed stosowaniem różnych technik statystycznych (regresja (PCR), klasyfikacja, klasteryzacja) najpierw redukuje się wymiar danych poprzez zastąpienie $n \times p$ macierzy danych X macierzą Z wymiaru $n \times M$, gdzie $M \ll p$ jest liczbą pierwszych M składowych głównych.
- Macierz $Z = [z_1, \dots, z_M]$ ma jako kolumny kolejne wektory zapisów dla pierwszych M składowych głównych.
- Często prowadzi to do zmniejszenia **poziomu szumu** w danych bo sygnał zawarty w danych jest zwykle skoncentrowany w pierwszych kilku składowych głównych.

t-distributed stochastic neighbor embedding (t-sne)

- metoda redukcji wymiaru do 2 lub 3 wymiarów na potrzeby wizualizacji
- z dużym prawdopodobieństwem dane podobne do siebie są redukowane do punktów bliskich sobie, a dane niepodobne do punktów odległych.

Kroki

- ❶ Wyznaczenie rozkładu prawdopodobieństwa na *parach* wielowymiarowych obserwacji w oryginalnym wymiarze, tak, że podobne obserwacje mają wysokie p-stwo, a różne mają mniejsze p-stwo.
- ❷ Zmapowanie punktów na podprzestrzeń o niskim wymiarze (np 2 lub 3)
- ❸ Wyznaczenie podobnego rozkładu poprzez minimalizację dywergencji Kullbacka–Leiblera pomiędzy tymi dwoma rozkładami względem położenia punktów na niskowymiarowej mapie.

Dywergencja Kullbacka-Leiblera

Niech p, q rozkłady przypisujące p-stwa, z jakimi zmienne losowe przyjmują kolejne wartości dla dla zmiennych dyskretnych, lub gęstości p-stwa tych samych dziedzinach.

Dywergencja Kullbacka-Leiblera dana jest wzorem:

$$d_{KL}(p, q) = \sum_i p(i) \log_2 \frac{p(i)}{q(i)}$$

dla rozkładów dyskretnych oraz

$$d_{KL}(p, q) = \int_{-\infty}^{\infty} p(x) \log_2 \frac{p(x)}{q(x)} dx$$

dla rozkładów ciągłych.

Dywergencja Kullbacka-Leiblera

Własności:

- $d_{KL}(p, q) \geq 0$
- $d_{KL}(p, q) = 0$ wtedy i tylko wtedy, gdy $p = q$.
- intuicja: im wyższe d_{KL} tym 'dalsze' od siebie rozkłady p i q

t-sne formalnie (1)

Zadane N wielowymiarowych punktów $\mathbf{x}_1, \dots, \mathbf{x}_N$. t-sne oblicza p-stwa p_{ij} które są proporcjonalne do podobieństwa punktów \mathbf{x}_i i \mathbf{x}_j :

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2 / 2\sigma_i^2)}$$

oraz

$$p_{i|i} = 0.$$

A zatem mamy $\sum_j p_{j|i} = 1$ dla każdego i .

Van der Maaten i Hinton, autorzy: *The similarity of datapoint x_j to datapoint x_i is the conditional probability, $p_{j|i}$, that x_i would pick x_j as its neighbor if neighbors were picked in proportion to their probability density under a Gaussian centered at x_i .*

t-sne formalnie (2)

Następnie definiujemy

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}.$$

Wzór wynika

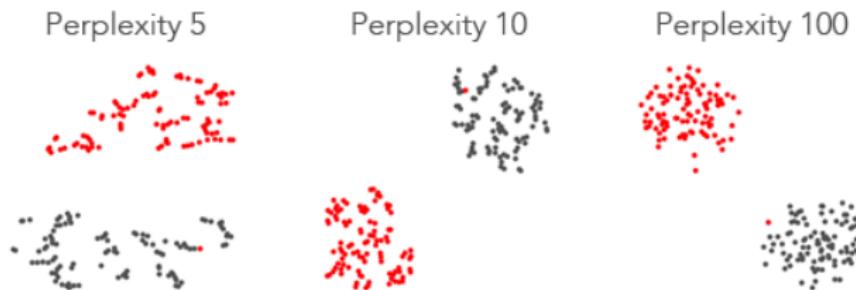
- ze wzoru Bayesa i faktu że zakładamy prior p_i i p_j dla N danych jako $1/N$, co daje p-stwo warunkowe $p_{i|j} = Np_{ij}$ i $p_{j|i} = Np_{ji}$ i
- z równości $p_{ij} = p_{ji}$.

Mamy $p_{ii} = 0$ $\sum_{i,j} p_{ij} = 1$.

Parametr σ_i jest zwykle dobierany w zależności od preferencji użytkownika.

t-sne formalnie (3) - perplexity

- Aby dobrać właściwe σ , wprowadza się zwykły parametr *perplexity* $\eta = 2^{H(p_{j|i}, \sigma)}$, gdzie H to entropia rozkładu $p_{j|i}$ przy założeniu odpowiedniej wartości σ
- wynik t-sne, będzie - w zależności od parametru η - przypisywał mniejszą lub większą wagę do zachowania lokalnych lub dalszych sąsiedztw



<http://blog.thegrandlocus.com/2019/12/a-tutorial-on-t-sne-2>

t-sne formalnie (4)

Szukane: punkty (nisko) d -wymiarowe $\mathbf{y}_1, \dots, \mathbf{y}_N$, $\mathbf{y}_i \in \mathbb{R}^d$, gdzie d zazwyczaj wynosi 2 lub 3, które w tej niskowymiarowej przestrzeni jak najlepiej oddają odległości p_{ij} .

Definiujemy zatem odległości w tej d wymiarowej przestrzeni jako

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_k \sum_{l \neq k} (1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}$$

i zadajemy $q_{ii} = 0$.

Ten wzór wynika z rozkładu t -Studenta. Rozkład t -Studenta, w porównaniu do rozkładu Gaussa, ma "cięższy ogon" - nadaje większe p-stwa skrajnym wartościom.

t-sne formalnie (5)

W ostatnim kroku poszukujemy takich położen punktów na niskowymiarowej mapie \mathbf{y}_i , że dywergencja Kullbacka-Leiblera p od q .

$$d_{\text{KL}}(p \parallel q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

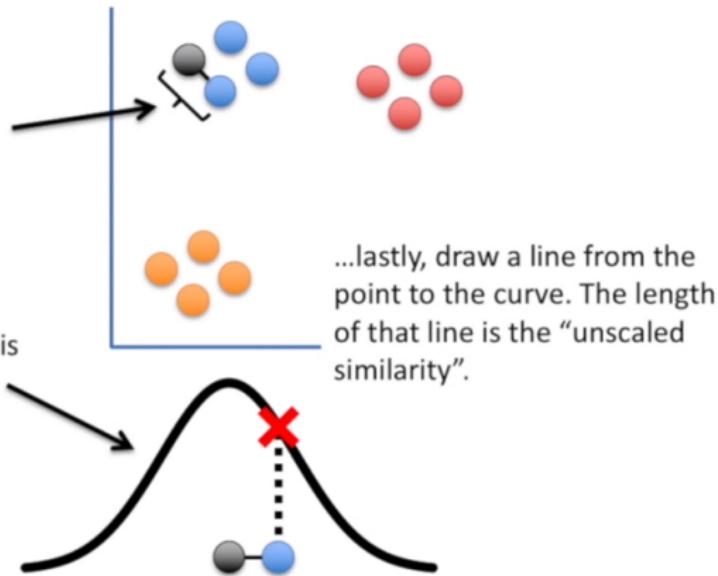
Minimalizacja odbywa się z wykorzystaniem algorytmu spadku gradientu.

t-sne toy example (1)

First, measure the distance between two points...

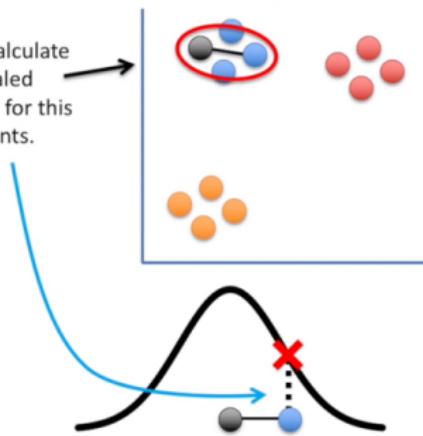
Then plot that distance on a normal curve that is centered on the point of interest...

...lastly, draw a line from the point to the curve. The length of that line is the “unscaled similarity”.

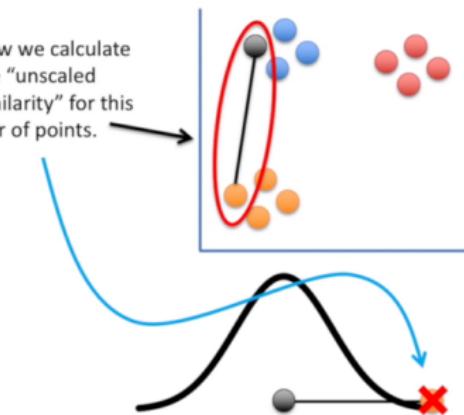


t-sne toy example (2)

Now we calculate
the “unscaled
similarity” for this
pair of points.

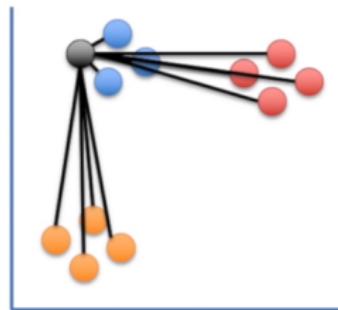


Now we calculate
the “unscaled
similarity” for this
pair of points.

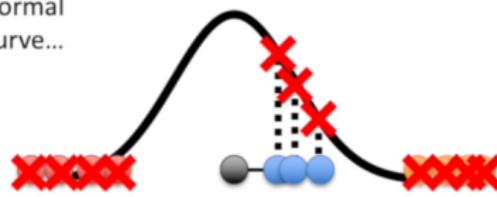


t-sne toy example (3)

Ultimately, we measure the distances between all of the points and the point of interest...

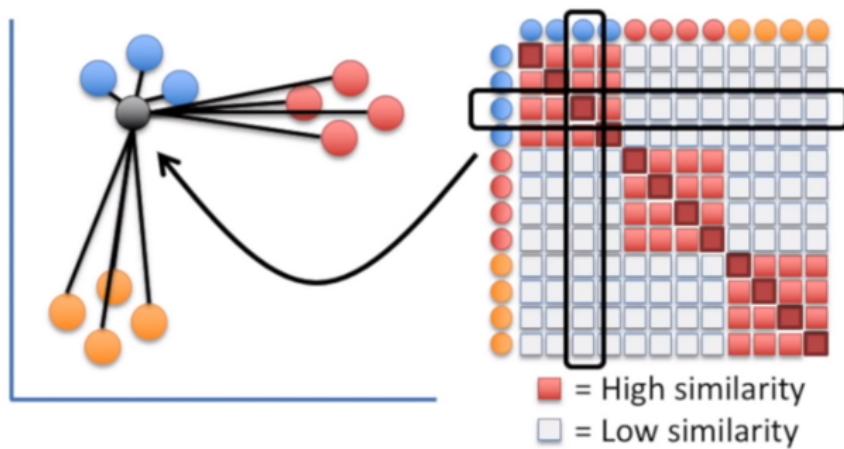


Plot them on the normal curve...



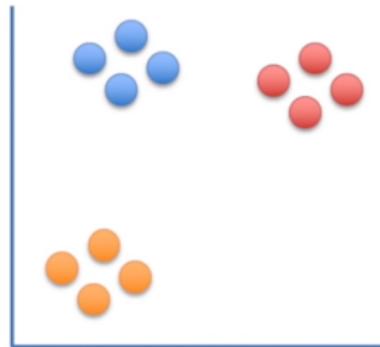
...and then measure the distances from the points to the curve to get the unscaled similarity scores with respect to the point of interest.

t-sne toy example (4)



t-sne toy example (5)

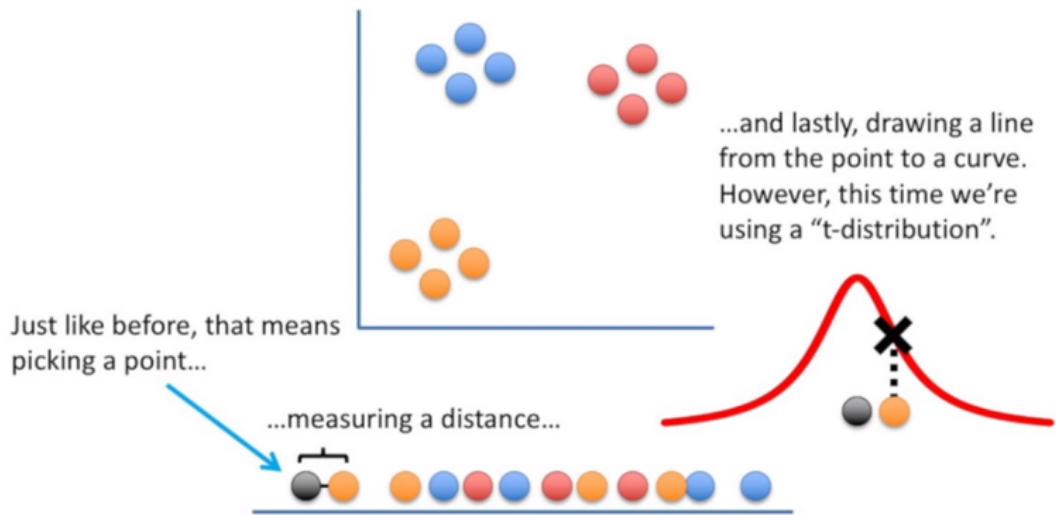
Now we randomly project
the data onto the number
line...



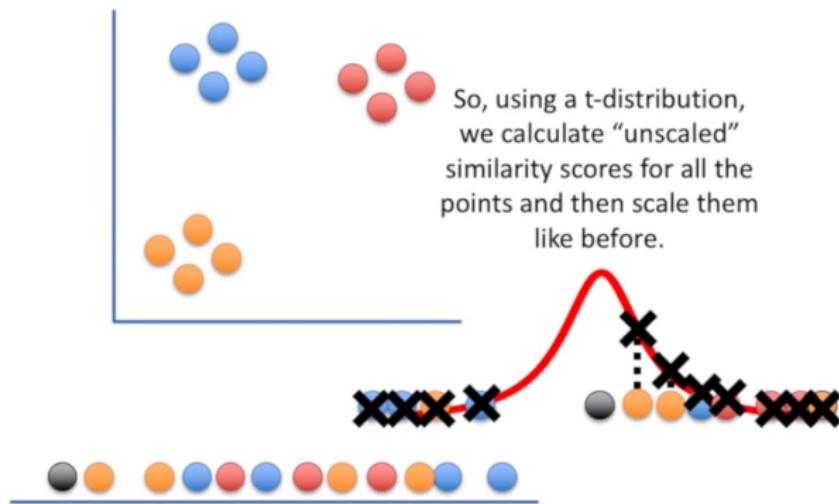
... and calculate
similarity scores for
the points on the
number line.



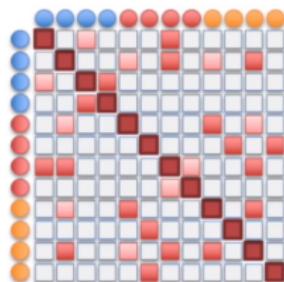
t-sne toy example (6)



t-sne toy example (7)

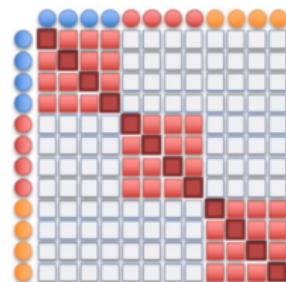
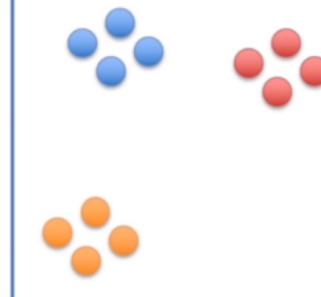


t-sne toy example (8)



■ = High similarity
□ = Low similarity

Like before, we end up with a matrix of similarity scores, but this matrix is a mess...



■ = High similarity
□ = Low similarity

...compared to the original matrix.



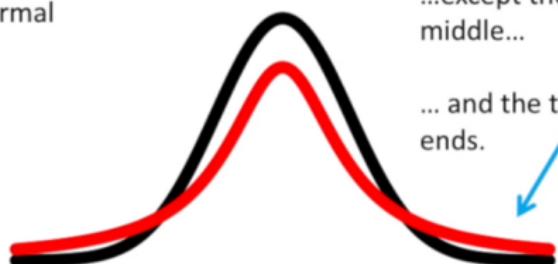
t-sne toy example (9)

A “t-distribution”...

...is a lot like a normal distribution...

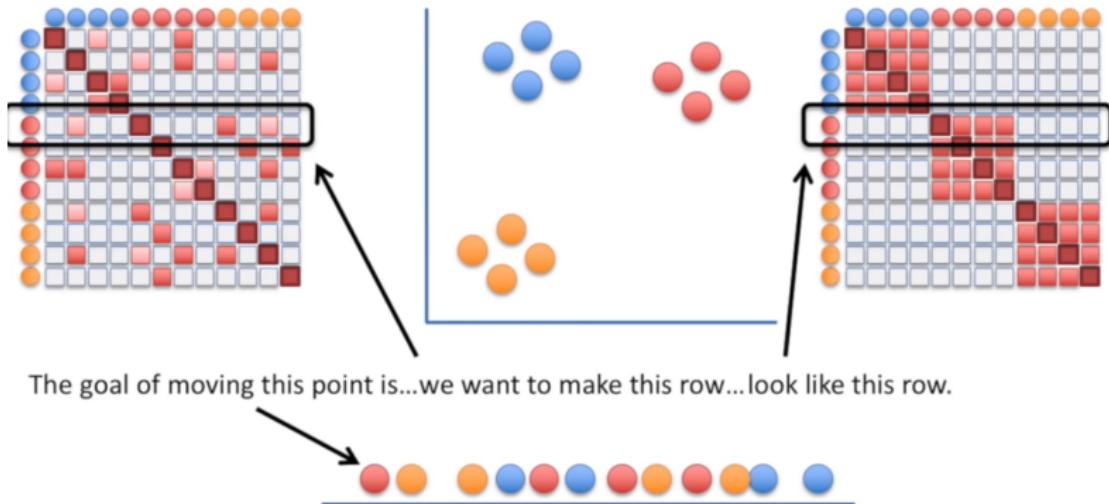
...except the “t” isn’t as tall in the middle...

... and the tails are taller on the ends.

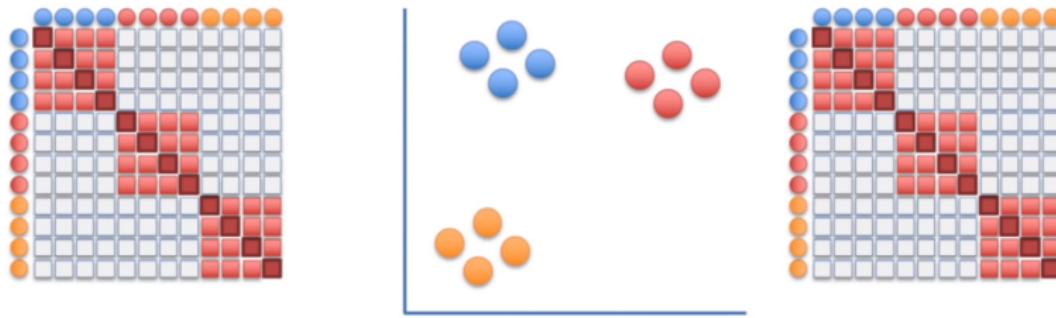


The “t-distribution” is the “t” in t-SNE.

t-sne toy example (10)



t-sne toy example (11)



t-SNE moves the points a little bit at a time, and each step it chooses a direction that makes the matrix on the left more like the matrix on the right.



It uses small steps, because it's a little bit like a chess game and can't be solved all at once. Instead, it goes one move at a time.

Zródła

- Gareth James, Daniela Witten, Trevor Hastie and Rob Tibshirani,
An Introduction to Statistical Learning
- Wikipedia https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding
- Josh Stramer, StatQuest <https://www.youtube.com/channel/UCtYLUTtgS3k1Fg4y5tAhLbw>