tSNE

t-distribution stochastic neighbor embedding

van der Maaten & Hinton 2008

based on Hinton & Roweis 2002 SNE

Non-Linear Dimensionality Reduction algo.

פופולרי במיוחד בשימוש של Data Science (ויזואליזציה)
יש ה חשוב ומעניין להציג אובייקטים ("מעניינים") ממימד
גבוה למימד נמוך (2D) כדי לראות את המרחקים
והקשר בין אובייקטים.

שני שלבים: (1) מיפוי הדמיון כהתפלגות בגבוה
(2) שיכון בממד נמוך ע"י עוות מינימלי.
D_{KL}

נתחיל בהסבר של SNE.

Data X: $X_i \in \mathbb{R}^p$ High.dim. data

Distances: $\forall_{ij}$ $d_{ij} = \|X_i - X_j\|$

נשתמש בקירוב גאוסיאני להפוך מרחק להסתברות
מותנית:

$$P_{j|i} = \frac{\exp(-d_{ij}^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-d_{ik}^2 / 2\sigma_i^2)}$$

פה, σ_i זו סטיית התקן של הגאוסיאן (קרנל)
ו- P_i היא ההתפלגות מעבר לשכני i.
(משתנה מקרי מולטינומיאלי)

איך נקבע את σ_i? הינטון ורואיס הציעו
להשתמש באנטרופיה של P_i כמדד
שמקרב את "מספר השכנים האפקטיבי"

<u>Perplexity</u> $= 2^{H(P_i)}$

כאשר האנטרופיה של שנון

$$H(P_i) = -\sum_j p_{j|i} \cdot \log_2 p_{j|i}$$

בהינתן מדד אי-וודאות/בלבול/Perplexity רצוי (ב[5-50])
נחפש לכל x_i את σ_i שאיתו $2^{H(P_i)}$ מושג.

small σ_i → "less neighbors" → more predictable transitions
lower entropy → lower perplexity

higher σ_i → ... → higher perplexity

<u>Embedding</u> find low-dimensional representation

$Y: \; Y_i \in \mathbb{R}^d \quad d \ll p$ ← מימד X

(עוד לא t-SNE) באופן דומה SNE

(Gauss. Kernel, $\sigma = \frac{1}{\sqrt{2}}$) נגדיר הסתברות q Y

$$q_{j|i} = \frac{\exp(d'^2_{ij})}{\sum_k \exp(d'^2_{ik})} \qquad d'_{ij} = \|y_i - y_j\|$$

while <u>minimizing</u> the (KL) divergence of P_i, Q_i

score (for SNE): $\quad C = \sum_i KL(P_i \| Q_i) = \sum_i \sum_j p_{j|i} \cdot \log_2 \frac{p_{j|i}}{q_{j|i}} \quad (*)$

איך נמזער את C ? Gradient Descent !

$$\frac{\partial C}{\partial y_i} = 2 \sum_j \underbrace{(p_{j|i} - q_{j|i} + p_{i|j} - q_{i|j})}_{\text{חוזק הכח : כגודל השגיאה}} \underbrace{(y_i - y_j)}_{\text{כיוון תנועה}}$$

Init. with random points (sampled from Gauss., $\sigma = \frac{1}{\sqrt{2}}$)

Use <u>momentum</u>

$$Y^{(t)} = Y^{(t-1)} + \underbrace{\eta}_{\text{learning rate}} \underbrace{\frac{\partial C}{\partial Y}}_{\text{gradient}} + \overbrace{\alpha(t)}^{\text{מומנטום}} \cdot \left[Y^{(t-1)} - Y^{(t-2)} \right]$$

Comment 1. Why KL-divergence?

Kullback-Leibler (Relative Entropy)

In Inf. theory: expected #bits (per sample) when compressing samples from P using code for Q (compared to a code for P)

In Bayesian Inference: Info. gained when moving from a prior dist Q to a prior P (for data from P).

$$D_{KL}(P \| Q) = -\sum_j P_j \log Q_j + \sum_j P_j \log P_j = \underbrace{H(P,Q)}_{\text{cross entropy}} - \underbrace{H(P)}_{\text{entropy}}$$

Symmetric version: Jensen-Shannon Divergence

$$D_{JS}(P \| Q) = D_{JS}(Q \| P) = \frac{1}{2} D(P \| M) + \frac{1}{2} D(Q \| M)$$

$$M = \frac{P+Q}{2}$$

אומר ממוצע הסטיות של הממוצע M ל- P ו- Q.

Comment 2 - So Why KL? (for SNE)

large $p_{j|i}$ are important. w/ small $q_{j|i} \to$ BIG penalty

small $p_{j|i}$, even if large $q_{j|i}$ yield small penalty

Moving to t-SNE

Solving two problems ① Outliers
② Crowding

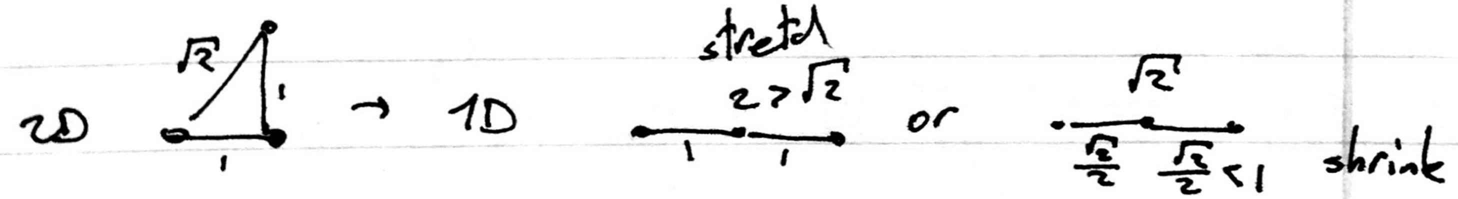
① X_i outlier → add noise to C (far from all)

Update: $C = KL(P \| Q) = \sum_{ij} p_{ij} \log p_{ij}/q_{ij}$

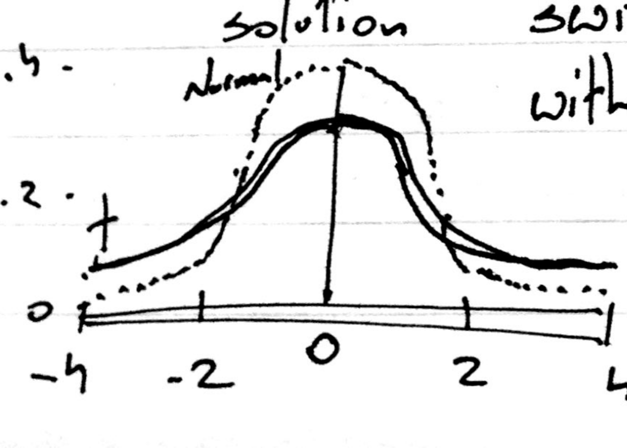
Hence: move from N dist. $p_{j|i}$ to one p_{ij} (pairwise)

$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2 \cdot N}$ → large p_{ij} (= close points) get stronger

② High-to-low dimen. embeddings expon. reduce the "space" for neighboring points. Causing under-est. for ~~smaller~~ close points, and over-est. of dist for distant



solution switch lower-dim kernel to t-distribution with heavier tail compared to normal



$X \to$ | High-D dists | $\xrightarrow{\text{Normal}}$ pair probs

$\updownarrow$ KL

$Y \to$ | Low-D dists | $\xrightarrow{\text{t-dist}}$ pair probs

Students t-distribution

PDF: $$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \cdot \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$, w/ param ν

$$\Gamma(x) = (x-1)! \quad , \quad \Gamma(1) = 1 \quad , \quad \Gamma\left(\tfrac{1}{2}\right) = \sqrt{\pi}.$$

w/ $\nu=1$ $$f(t) = \frac{\overset{=1}{\Gamma(1)}}{\sqrt{\pi}\cdot \underset{=\sqrt{\pi}}{\Gamma\left(\frac{1}{2}\right)}} (1+t^2)^{-1} = \frac{(1+t^2)^{-1}}{\pi}$$

in t-SNE Q is defined with a t-dist kernel

$$q_{ij} = \frac{\left(1+\|y_i - y_j\|^2\right)^{-1}}{\sum_{k\neq l}\left(1+\|y_k - y_l\|^2\right)^{-1}}$$

ושוב נשתמש באופטימיזציה למזעור Gradient Descend

$$\frac{\partial C}{\partial y_i} = 4\sum_j \underbrace{(p_{ij}-q_{ij})}_{\substack{\text{חוזק הכח} \\ \text{HiD-LowD}}} \underbrace{(y_i - y_j)}_{\text{כיוון כח}} \underbrace{\left(1+\|y_i - y_j\|^2\right)^{-1}}_{\text{״מרחקי צלעי״}}$$

ריווח קרובים,
ריכוז רחוקים.

[for speed-up, use Barnes-Hut simulation
where neighboring points are grouped to compute forces]