ML2 dseb62 w2 - Nguyen Tuan Duy

Nguyen Tuan Duy

January 2023

1 Problem 1

For SNE:
Given point $x_1, x_2, ...x_N \in R^D$ we define the distribution P_{ij} which is the probability that point x_i chooses x_j as its neighbor:

$$P_{ij} = \frac{exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq i} exp(-||x_i - x_k||^2/2\sigma_i^2)} \tag{1}$$

We need to find good embedding $y_1, y_2, ...y_N \in R^d$ for $d < D$:

$$Q_{ij} = \frac{exp(-||y_i - y_j||^2)}{\sum_k \sum_{k \neq l} exp(-||y_l - y_k||^2)} = \frac{E}{Z} \tag{2}$$

let $||y_i - y_j|| = d_{ij}$
We need to optimize Q to be close to P, we do so by minimizing KL-divergence: to find the embedding $y_1, ...y_n \in R^d$:

$$KL(P||Q) = \sum_{ij} P_{ij} log(\frac{P_{ij}}{Q_{ij}}) \tag{3}$$

$$= \sum_{ij} P_{ij} log(P_{ij}) - P_{ij} log(Q_{ij}) \tag{4}$$

Where P_{ij} can be inferred from the data, we treat this as constant so (4) can be rewritten as:

$$KL(P||Q) = -\sum_{ij} P_{ij} log(Q_{ij}) + Constant \tag{5}$$

$$= -\sum_{ij} P_{ij} log(\frac{E}{Z}) + Constant \tag{6}$$

$$= -\sum_{ij} P_{ij} log(E) - P_{ij} log(Z) + Constant \tag{7}$$

$$\frac{\delta L}{\delta y_i} = (\frac{\delta L}{\delta d_{ij}} + \frac{\delta L}{\delta d_{ji}})\frac{\delta d_{ji}}{\delta y_i} = 2\frac{\delta L}{\delta d_{ij}}\frac{\delta d_{ij}}{\delta y_i} \tag{8}$$

$$\frac{\delta L}{\delta y_i} = -2\sum_{ij} P_{ij}\delta log(E) - P_{ij}\delta log(Z) \tag{9}$$

$$\delta E = -2(y_i - y_j)E \tag{10}$$

First we consider the first term:

$$\sum_{ij} P_{ij}\delta log(E) = \sum_{ij} P_{ij}(-2(y_i - y_j)E)\frac{1}{E} \tag{11}$$

$$= -2\sum_{ij} P_{ij}(y_i - y_j) \tag{12}$$

in the second term,$\sum_{k\neq l} P_{ij} = 1$, the derivative is non-zero when k = i or l = i:

$$P_{ij}\delta log(Z) = \sum_{i\neq j} \frac{1}{Z}\delta E \tag{13}$$

$$= 2\sum_{i\neq j} \frac{E}{Z}(y_i - y_j) \tag{14}$$

$$= 2\sum_{i\neq j} Q_{ij}(y_i - y_j) \tag{15}$$

Plug (12), (15) into (9) we have:

$$\frac{\delta L}{\delta y_i} = -2\sum_{ij} -2P_{ij}(y_i - y_j) + 2Q_{ij}(y_i - y_j) \tag{16}$$

$$= 2\sum_{ij} 2P_{ij}(y_i - y_j) - 2Q_{ij}(y_i - y_j) \tag{17}$$

$$= 4\sum_{ij} (P_{ij} - Q_{ij})(y_i - y_j) \tag{18}$$

For T-SNE, Q_{ij} is defined as:

$$Q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_k \sum_{k\neq l}(1 + ||y_k - y_l||^2)^{-1}} = \frac{E^{-1}}{Z^{-1}} \tag{19}$$

We use the same P_{ij} and our loss function is:

$$KL(P||Q) = -\sum_{ij} P_{ij}log(E^{-1}) - P_{ij}log(Z^{-1}) + Constant \tag{20}$$

$$\tag{21}$$

$$\frac{\delta L}{\delta y_i} = -2\sum_{ij} P_{ij}\delta log(E^{-1}) - P_{ij}\delta log(Z^{-1}) \tag{22}$$

$$\delta E^{-1} = -(y_i - y_j)E^{-2} \tag{23}$$

We consider the first term:

$$\sum_{ij} P_{ij}\delta log(E^{-1}) = -\sum_{ij} P_{ij} 2(y_i - y_j)\frac{E^{-2}}{E^{-1}} \tag{24}$$

$$= -2\sum_{ij} P_{ij}(y_i - y_j)E^{-1} \tag{25}$$

We consider the second term:

$$\sum_{ij} P_{ij}\delta log(Z^{-1}) = -2\sum_{ij}(y_i - y_j)\frac{1}{Z^{-1}}E^{-2} \tag{26}$$

$$= -2\sum_{ij}(y_i - y_j)\frac{E^{-1}}{Z^{-1}}E^{-1} \quad = -2\sum_{ij}(y_i - y_j)Q_{ij}E^{-1} \tag{27}$$

Plug (23), (25) into (20) we have:

$$\frac{\delta L}{\delta y_i} = 4\sum_{ij} P_{ij}(y_i - y_j)E^{-1} - (y_i - y_j)Q_{ij}E^{-1} \tag{28}$$

$$= 4\sum_{ij}(P_{ij} - Q_{ij})(y_i - y_j)E^{-1} \tag{29}$$

$$= 4\sum_{ij}(P_{ij} - Q_{ij})(y_i - y_j)(1 + ||y_i - y_j|)^{-1} \tag{30}$$

2 Problem 2

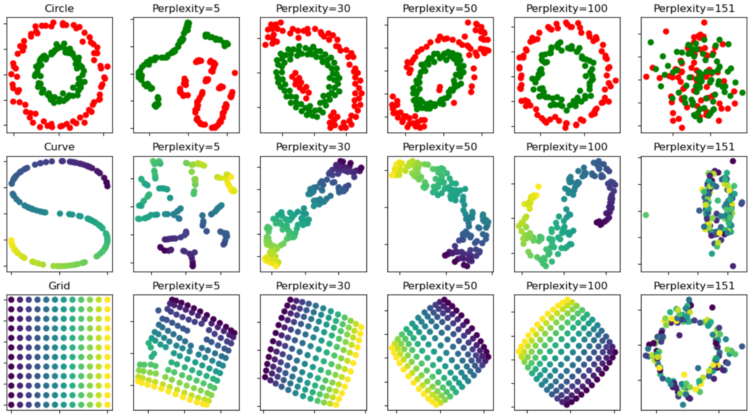

Figure 1: Result of fitting t-sne

Perplexity is defined as

$$perp(P_{j|i}) = 2^{H(P_{j|i})}$$

with $H(P) = -\sum_i P_i log(P_i)$ is the entropy. If perplexity is uniform over k elements - perplexity is k. Low perplexity means low σ whereas high perplexity means high σ, and σ sets the size of the neighborhood:
- When Perplexity = 5, σ is very low, all probability is in the nearest neighborhood. On the graph, all points closest to each other belongs to the same neighborhood
- When Perplexity = 30 and 50, sigma is higher, a larger neighborhood's local structure is preserve and the shape resembles original data.
- When Perplexity = 100, for all the circle data, the shape is similar to the original data, but distance between points and circle's size vary. On the s-curve, the shapes diverge from the original.
- When perplexity = 151, the perplexity is too high, which leads to uniform weights, ambiguous neighborhood for data points.

3 Problem 3

	Paris	curry	duck	goose
0	Parisian	curries	ducks	geese
1	Hopital_Europeen_Georges_Pompidou	chicken_curry	Joshua_Linhares	bird
2	Spyker_D##_Peking	vindaloo	Aflac_dumps	pheasant
3	France	balti	firefights_erupt_outside	turkey
4	Pantheon_Sorbonne	lamb_curry	Peking_roast	Canada_geese
5	Aeroports_De	pilau_rice	drake_mallard	waterfowl
6	Grigny_south	tikka_masala	goose	moose
7	Place_Denfert_Rochereau	naan_bread	black_bellied_whistling	mallard
8	guest_Olivier_Dolige	tandoori	bird	squirrel
9	Lazard_Freres_Banque	vindaloos	bluebill	Geese

	steve	dishwasher	fair	return
0	dave	dishwashers	Fair	returning
1	jeff	washing_machine	Kurylowicz_trial	returns
2	jason	dish_washer	P###_INX_futures	returned
3	robert	washer	equitable	rejoin
4	george	kitchen	Ranee_Gaynor	Return
5	kevin	dishwashing	MATAGORDA_Trout	leave
6	todd	refrigerator	spokeswoman_Brienna_Schuette	depart
7	jeremy	Bosch_dishwasher	fairness	Returning
8	ryan	clothes_washer	reasonable	reclaim
9	greg	dishwashing_machine	fairer	retun

	man	Python
0	woman	Jython
1	boy	Perl_Python
2	teenager	IronPython
3	teenage_girl	scripting_languages
4	girl	PHP_Perl
5	suspected_purse_snatcher	Java_Python
6	robber	PHP
7	Robbery_suspect	Python_Ruby
8	teen_ager	Visual_Basic
9	men	Perl

a) Closely embedded words are words in different forms (Goose and geese, duck and ducks, return and returned or returns), words that contains the original word (dishwasher and Bosch_dishwasher, man and woman, curry and chicken curry), different writing (duck and Duck, return and Return) or words that are close in meaning (fair and reasonable or equitable). Moreover, words that are closely embedded can be related character or events (Steve Jason is a singer, Paris is the capital of France, or the event of Joshua Linhares being put on trial for killing ducks)

b)

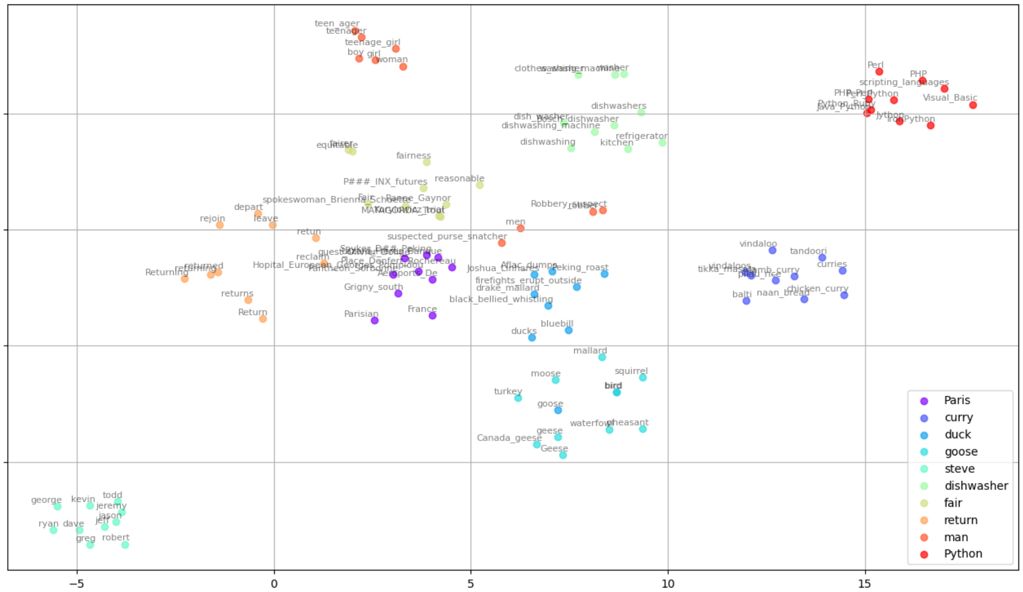

Figure 2: Result of fitting t-sne

The words associated with original words are placed close to each other as they have a similar context. The words associated with "Steve" and "Python" are placed far from each other because they have different context.

4 Problem 4

Compare T-sne and PCA:
Similar: They are both unsupervised dimensional reduction and data visualization technique for very high dimensional data

index	PCA	T-sne
1	linear	non-linear
2	preserve global structure	preserve local structure
3	Does not involves hyper parameters	Involves hyper parameters
4	Affected by outliers	Can handle outliers
5	deterministic	randomised
6	rotating vectors for preserving variance	Minimising the distance between the point
7	preserve using eigenvalues	using hyper parameters

explanation:
(1) PCA works well when there is linear relation between features while T-sne does a decent job even in non-linear dataset.
(3) T-sne hyper parameters include perplexity, learning rate, iterations.

(2), (4) PCA can lead to local inconsistencies, far away point can become nearest neighbor. For t-sne, low dimensional neighborhood should be the same as original neighborhood.
(5) PCA produces the same output each time, T-sne's intuition is based on random walk between data points, therefore may produce different result on the same data. Moreover, in Problem 2, difference in hyper parameters like perplexity can produce different result.
(6) Optimization problem for PCA is maximising the variance of the projected data. For T-sne, we minimize KL-divergence so that Q (distribution for projected data points) is close to P (distribution for given data)
(7) PCA decides how much variance to preserve using eigenvalues. T-sne decides the distance to preserve using Perplexity

5 Problem 5