Change in objective function

Summary

Testing the idea of a new penalty term for the objective function of tsne:

- The simplest approach is using L2 penalty, but we have to find the regularization constant manually.
- Another approach is to try to convert the distance between the fixed points and its old neighbors by introducing a gaussian or a student t distribution at the location of the fixed point. The intuition behind this idea is that, if a point y_j is neighbor of a point y_i, it should have large likelihood to be still neighbor of this point at the new position y'_i.
- This way we can minimize the KL divergence and maximize the above likelihoods at the same time, that forces the neighbors of the fixed points move closer to the new position of the fixed points to preserve the neighborhood relation.
- The params are found manually when experimenting on the small MNIST dataset.

1. Using Student t-distribution around the fixed points

> Updated 15/03/2018

- For each fixed point y'_i, find its `k` nearest neighbors y_j based on its old position y_i.
- Convert the distance between y'_i and y_j into the probability by using a student t distribution with a degree of freedom $\nu = 1$, a centre μ at the new position y'_i and a scale σ as a param:

$$p(y_j|y'_i) \propto \left(1 + \frac{||y_j - y'_i||^2}{\sigma^2}\right)^{-1}$$

Let $p(y_j|y'_i)$ be the probability that a single neighbor y_j of the old point y_i being still attracted by the new position y'_i.

> We omit the normalized constant in the formula of t distribution beacause when we take the log, it becomes an additional (unnecessary) term.

- The likelihood that all the `k` neighbors of y_i are still attracted by the new position y'_i, is a joint distribution $p(y_1, \dots, y_k|y'_i) = \prod_{j=1}^{k} p(y_j|y'_i)$.
- We wish to maximize this likelihood for each fixed point y'_i, that can be achieved by minimizing the negative log likelihood of the above joint distribution .
- We introduce a new term in the objective function of `tsne`:

$$C = KL(P||Q') + \sum_i^m \left(-\log \prod_j^k p(y_j|y'_i)\right)$$

$$C = KL(P||Q') + \sum_i^m \left(-\sum_j^k \log p(y_j|y'_i)\right)$$

$$C = KL(P||Q') + \sum_i^m \sum_j^k \log\left(1 + \frac{||y_j - y'_i||^2}{\sigma^2}\right)$$

> Q' is calculated based on the new position of `m` fixed points. Note that, we do not have a regularization constant, to avoid the domination of the log likelihood, we have to set the scale σ^2 very large.

- When calculate gradient for the neighbor points y_j, we add the following term:

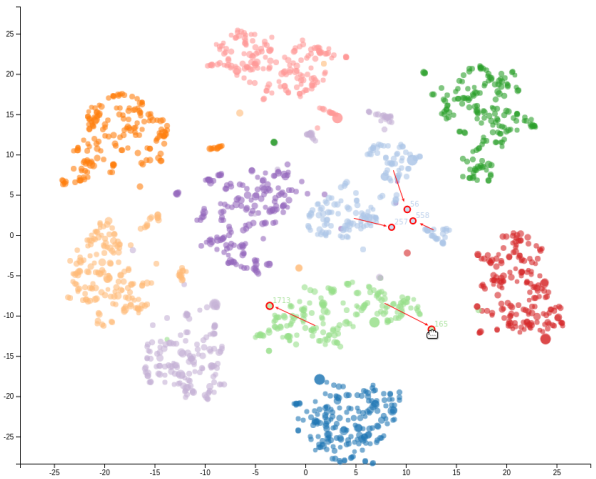
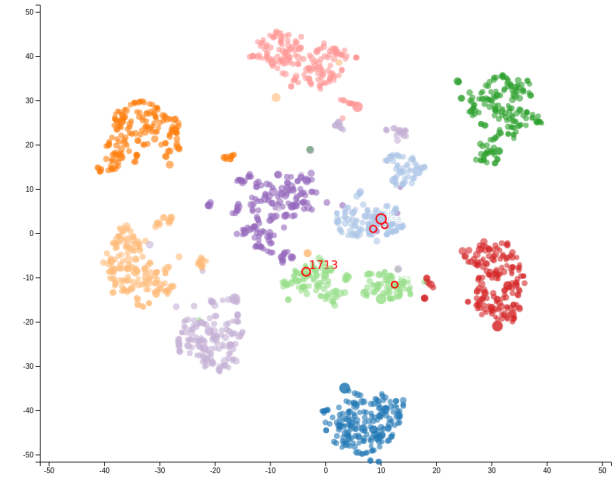
$$\frac{\partial}{\partial y_j}\left(-\log p(y_j|y'_i)\right)$$

$$= \frac{\partial}{\partial y_j}\log\left(1 + \frac{||y_j - y'_i||^2}{\sigma^2}\right)$$

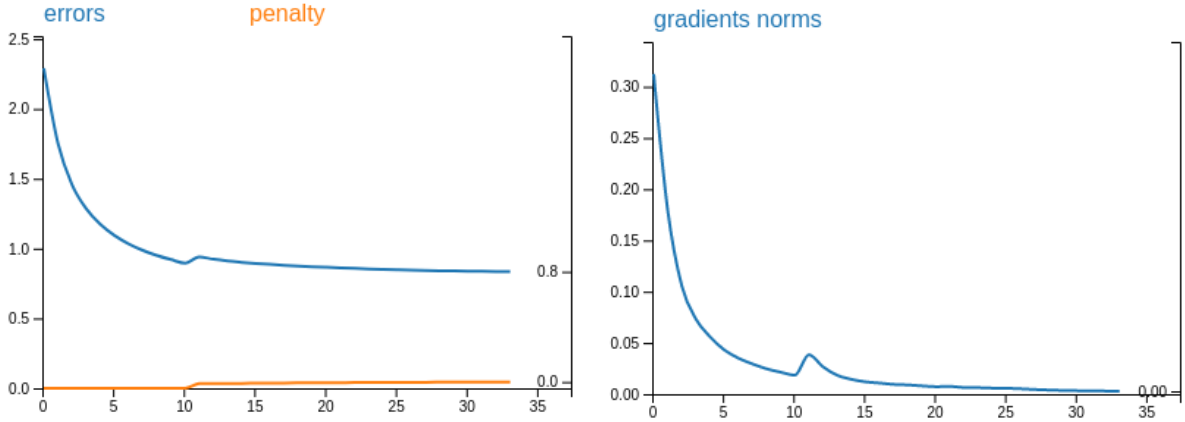
$$= \frac{\partial}{\partial y_j}\left(1 + \frac{||y_j - y'_i||^2}{\sigma^2}\right)\left(1 + \frac{||y_j - y'_i||^2}{\sigma^2}\right)^{-1}$$

$$= \frac{2}{\sigma^2}(y_j - y'_i)\left(1 + \frac{||y_j - y'_i||^2}{\sigma^2}\right)^{-1}$$

> The current value of σ^2 that gives us a clear result on the small MNIST dataset is around `1e5`

Before moving	After moving
	

> The value the likelihood term is ploted (and named `penalty`). We do not see the large change after moving 5 points.



2. Using Gaussian distribution around the fixed points

> Updated 14/03/2018

- Using the same setting as used with the t-distribution, we define the probability that a point y_j being still a neighbor of new fixed point y'_i as

$$p(y_j|y'_i) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(\frac{-||y_j - y'_i||^2}{2\sigma^2}\right)$$

$$\log p(y_j|y'_i) = \frac{-1}{2\sigma^2}||y_j - y'_i||^2 - constant$$

- The additional goal is to minimize the negative log likelihood of the joint distribution of the neighbors given the fixed points.

$$C = KL(P||Q') + \sum_i^m\left(-\log\prod_j^k p(y_j|y'_i)\right)$$

$$C = KL(P||Q') + \frac{1}{2\sigma^2}\sum_{i}^{m}\sum_{j}^{k}||y_j - y'_i||^2$$

- This way, we can replace the `regularization term` in the approach using L2 penalty by $\frac{1}{2\sigma^2}$.

> Some values of σ^2 in range `[1e3, 1e5]` give a clear result.

- The additional gradient for each neighbor point y_j is the partial derivative of the negative log likelihood w.r.t. y_j, that is exactly the same as using L2-penalty.

$$\frac{\partial}{\partial y_j}\left(-\log p(y_j|y'_i)\right)$$
$$= \frac{\partial}{\partial y_j}\left(\frac{1}{2\sigma^2}||y_j - y'_i||^2\right)$$
$$= \frac{1}{\sigma^2}(y_j - y'_i)$$

3. Using L2-regularization term

> Updated 09/03/2018

- Suppose that the new positions of the fixed points as y'.
- Update the new positions of y' into the current embedding coordinates to calculate new $\mathbf{Q}'$
- Add a new regularization term (L2-penalty) to the objective function

$$\sum_{i=1}^{m}\sum_{j=1}^{k}||y'_i - y_j||^2$$

  in which `m` is the number of fixed points and `k` is the number of neighbors around each fixed point. Testing with the value of `k` is `5%` total number of data points.
- The new objective function

$$C = KL(P||\mathbf{Q}') + \frac{\lambda}{mk}\sum_{i=1}^{m}\sum_{j=1}^{k}||y'_i - y_j||^2$$

  Test with `MNIST-small` dataset, $\lambda = 1e-3$ gives us a clear result.
- Calculate gradient for each points, that is the derivative of the new objective function with respect to each data points.
  - For the fixed points y'_i, we do not expect theirs positions will be changed anymore, so we fix their gradients to zero

$$\frac{\partial C}{\partial y'_i} = 0$$

  - Repeat `m` times: for each neighbor y_j of the fixed points y'_i

$$\frac{\partial C}{\partial y_j} = \frac{\partial KL}{\partial y_j} + \frac{-2\lambda}{k}(y'_i - y_j)$$

- Some minor changes with params of `tsne`:
  - `perplexity=30.0`
  - `early_exaggeration=12.0` (in `sklearn`: 12.0; in the original paper: 4.0)
  - `learning_rate=100.0` (default in `sklearn`: 200.0, in the original paper: 100.0)