Problem 1

Non-linear dimension reduction

There is some data on BruinLearn in `homework5/problem1.tsv` you need to load for this problem.

You don't need to know the generating process for the data, but in case it is helpful, here it is. Rows 1 to 100 correspond to the first 'cluster' and rows 101 to 200 correspond the second 'cluster'. The data is generated by:

$$x_i \sim \text{Normal}_2(\mu_{k(i)}, I_2),$$

where $k(i) = 1$ for $i = \{1, 2, \ldots, 100\}$ and $k(i) = 2$ for $i = \{101, 101, \ldots, 200\}$. $\mu_1 = (0, 0)$ and $\mu_2 = (10, 10)$.

You are going to implement some components of t-SNE to get some intuition about how different components of the algorithm work.

Finally, everything you need to know is in the non-linear dimension reduction slides.

(a) (Page 33) Implement the $p_{j|i}$ matrix. Do yourself a favor and make it a function because you're going to use it quite a bit. You can make your function take a single shared $\sigma_i^2 = \sigma^2$. Sanity check: each row should sum to 1. Side note: if you decide to do the extra credit (see (k)), you should allow your algorithm to utilize different σ_i^2 otherwise you're gonna have a bad time. For completeness, the equation,

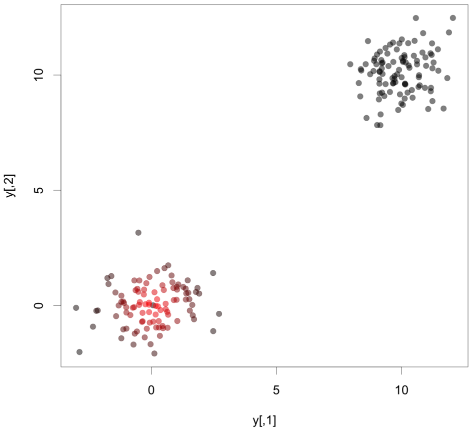
$$p_{j|i} = \frac{\exp(-||x_i - x_j||^2/2\sigma_i^2)}{\sum_{k \neq i} \exp(-||x_i - x_k||^2/2\sigma_i^2)}.$$

(b) (Page 33) Implement the p_{ij} matrix. Sanity check: the entire matrix should sum to 1. Again, the equation,

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

where N is the number of samples (200 in this case).

(c) Using $\sigma^2 = 1$, plot the entire dataset and color the points based on their probability relative to the first data point. To be rigorous: p_{1j} is the vector of probabilities of "j picking 1 as its neighbor". A reasonable color scale might be: $w_j \propto p_{1j}/\max_k(p_{1k})$. Your plot should show a change of color away from the first data point. Do the same for $\sigma^2 = \{0.1, 10, 100\}$. Each plot might look something like this:



(d) (Page 35) Implement the q_{ij} matrix. Sanity check: the entire matrix should sum to 1. The equation:

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

(e) Using $y_i = x_i$ (the original data is projected using the identity), plot the entire dataset and color the points based on their q_{1j} probability relative to the first data point. How is it different than the p_{1j} plot from (c) when $\sigma_i^2 = 1$?

(f) (Page 25) Implement the KL-divergence. Note, the contribution of $\{ij\}$ is zero when $p_{ij} = 0$.

(g) Using the real data as the low-dimensional projection, compute the KL-divergence when:

i. $\sigma^2 = 0.1$.

ii. $\sigma^2 = 1$.

iii. $\sigma^2 = 100$.

Any thoughts on what might be happening?

(h) Summarize your thoughts on how these hyper-parameters matter.

(i) Extra credit (1 point): Using $\sigma^2 = 1$, can you find a projection that reduces the KL-divergence? Note, there are plenty linear or non-linear ones. The easiest might be to do might be 'move' one cluster. Plot the projection and report the KL-divergence.

(j) Extra credit (2 points): implement the Perplexity and recompute the KL-divergence from the previous projections you made using the σ_i^2 you get for Perplexity $\{5, 25, 50, 100\}$. To implement the Perplexity, find the value of σ_i^2 that approximately satisfies the equation Perplexity$(P_i) = 2^{H(P_i)}$ where P_i is the i-th row in the $p_{j|i}$ matrix and $H(P_i)$ is the Shannon entropy. Plot a histogram of your σ_i^2 values and how did your results change?