CPSC/AMTH 453 / 553 GENE/CBB 555 Final Exam
Date: 12-10-2020 Total points: 120 Points for 100%: 100 points

Multiple choice/selection (20 points) [no selection is allowed]:

1. (5 points) Which are the following are **distances** between probability distributions?
    a. KL-divergence (assymetric)
    b. Pointwise Euclidean distance between pdfs (this is not a GOOD distance)
    c. Earth mover's distance.
    d. Maximal mean discrepancy.
    e. Cross entropy (assymmetric)
2. (5 points) Which of the following is an essential difference between a GAN and a WGAN? **WGANS approximate the dual form of the earth mover's distance. WITNESS FUNCTION f. Kantorovich Rubenstein Duality.**
    a. A bottleneck in the discriminator.
    b. Weight clipping in the generator.
    c. A condition input given to the generator and discriminator.
    d. Weight clipping in the discriminator.
    e. Non-sigmoidal discriminator output activation.
3. (10 points) For each of the following say whether it is likely to be a **high** or **low** frequency noise on the data affinity graph (**High frequency = changes quickly over a data affinity graph, Low frequency= changes slowly, random noise is high frequency, systematic effect is low frequency)**
    a. The discrete batch signal for a mass cytometry dataset that was collected in two batches. **LOW**
    b. The transcript "dropout noise" that affects single cell RNA-sequencing. **HIGH**
    c. Blurriness in an image. **Low frequency**
    d. Salt-and-pepper noise on an image. **High frequency**
    e. A smudge on a camera lens that affects some pixels in image. (**Could be either depending on the size of the smudge).**

Short answer (30 points):

1. (10 points) How is mutual information computed? Why is it able to detect non-monotonic and non-linear relationships between variables?

    H(Y)-H(Y|X) how much knowing X reduces the entropy of Y.
    Entropy is the measure of uncertainty of a random variable.
    Uniform distributions have maximal entropy.

    As long as knowing X determines Y or narrows down the spread of Y, it will pick up the relationship.
2. (10 points) What is the best kind of neural network for each of these tasks and why?

    a. Generating new poetry similar to shakespeare. **(Transformer, LSTM/attention)**
    b. Removing noise from a set of blurry images **(Denoising autoencoder)**
    c. Generating specific images of specific types of (cats, dogs, birds..) as requested by a user **(Conditional GANs, Conditional VAE)**
    d. Classifying proteins based on their molecular structure graphs **(GNN)**
    e. Embedding EEG data of a set of patients **(Seq-to-seq, transformer)**
3. (10 points) What are the advantages and disadvantages of each of these dimensionality reduction methods
    a. PCA **Advantages: interpret new axes via the loadings, Disadvantage: linear dimensions, can't denoise well (in non-linear paths for example)**
    b. tSNE Advantage: **reduce dimensions to 2, and keep neighbors close, Disadvantage: doesn't keep overall structure of the data intact, medium/long range connections**
    c. Diffusion maps: **Advantage: it traverses the "native space" (manifold learning), good for clustering (spectral clustering), Disadvantage: Gives high dimensional representations that are not great for visualization, play with the bandwidth**
    d. PHATE: **Advantage: reduces to 2 dimensions, it maintains global and local structure, preserves manifold affinity Disadvantages: overplots points, have to zoom in to replot, parameters like bandwidth and "T"**
    e. MDS: **Advantage: keeps distances (so if you find a really unique distance this is good for preserving , MMDS reduce to 2-D (like tSNE/PHATE).**
4. (10 points) What is kernel PCA? What is the kernel trick? How is it an example of the kernel trick?

Any affinity matrix that is PSD (positive semidefinite) can be regarded as a dot product matrix of high dimensional features. We don't actually have to find the features, we can just treat it like the covariance matrix in PCA and use eigendecomposition to preserve the affinity in low dimensions.

Kernel PCA is actually the kernel trick applied to classic MDS. Instead of using euclidean distance between points (as in classic MDS) we can use any non-linear affinity (i.e. like the Gaussian Kernel).

Long answer (40 points):

1. (10 points) Write the pseudocode for computing a diffusion map. To which step does the "diffusion" refer? Beyond visualization, what are some other uses of the data diffusion?

Data matrix X with rows being observations, columns being features.
D=Pairwise_distance(X)
Affinity= apply_pointwise_kernel(D)

M= Markov_normalize(Affinity)
P=M^t (this step diffuses)
[U, V] = eigendecompose (P)

Clustering, denoising, “graph fourier transforms”

2. (10 points) When creating a diffusion map, what considerations are made with each of the following parameters: kernel type, fixed vs adaptive bandwidth, kernel size bandwidth (size for fixed and kNN for adaptive) and t. Explain how differing choices for each of these parameters would change your embedding as well as the eigenvectors of the diffusion operator.

Kernel type: Prior knowledge about data, and creating sparsity, 0-1 kernel leads to a lot of 0’s and you could use sparse matrix representations.

Fixed vs adaptive: Separating the geometry from density, remember diffusion paths walk towards density. Generally adaptive is better.

T: More about denoising and how many eigencomponents you want to consider for downstream applications.

3. (10 points) Suppose you have a list of people on a social media platform similar to twitter, the people they follow, and the texts in their tweets. How would you figure out if the people are politically polarized using graph signal processing ? How would you quantify that?

Create a graph. Vertices are people, edges are people following (you can leave it undirected). One side of the graph is party 1 and other side of the graph is party 2. Node features could be political word usage (how frequently does a person say one of 10 words, how frequently do they retweet particular political figures). Graph fourier transform these features. Low frequency = political polarization.

4. (10 points) Suppose you have a set of documents in a database. 1) Describe how you would figure out how many topics they cover in totality, and 2) Describe how many different types of documents occur in the database. You are free to use any techniques discussed in the class.

    Topic modlling

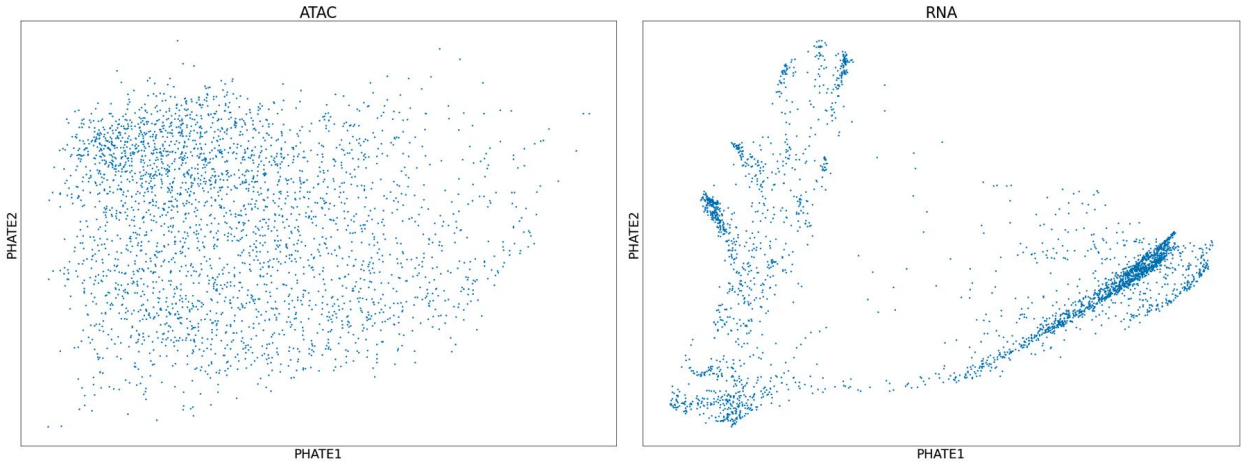
    How many topics?: Word2vec embed the key words in a document—> clustering, visualization of the words to find out how many topics

How many document types: Embed the documents on the basis of which topics they cover.

Need an answer that talks about term frequency and document frequency

Bonus question (20 points):
You have downloaded a publicly available single-cell dataset that profiles mouse skin cells in two different modalities: RNA expression and chromatin accessibility (ATAC). There is cell-to-cell correspondence across both datasets such that every cells has been completely measured across both modalities. You are interested in doing an integrated analysis with both datasets but are unsure how to do this.

Figure 1: PHATE Embeddings of both Modalities:



1. When you visualize each dataset you see the Figure 1. What could potentially cause the difference in the visualizations?
2. Undeterred, you decide that you want to integrate information from both datasets and create a common diffusion operator. How would you do this?
3. Name 2 cool things you could do with this common diffusion operator?