CS 6966: High-Dimensional Data Analysis **Due: Thursday 10.03 at 11:50pm**

Homework 2: Building Embeddings

Instructions: Your answers are due at 11:50pm submitted on canvas. You must turn in a pdf through canvas. I recommend using latex (`http://www.cs.utah.edu/~jeffp/teaching/latex/`, see also `http://overleaf.com`) for producing the assignment answers. If the answers are too hard to read you will lose points, entire questions may be given a 0 (e.g. **sloppy pictures with your phone's camera are not ok, but very careful ones are**)

Please make sure your name appears at the top of the page.

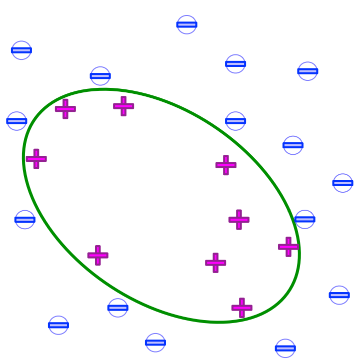
You may discuss the concepts with your classmates, but write up the answers entirely on your own. **Be sure to show all the work involved in deriving your answers! If you just give a final answer without explanation, you may not receive credit for that question.**

---

1. [**40 points**] Implement either variant of Random Fourier Features to approximate the Gaussian kernel $K(x,p) = \exp(-\|x-p\|^2)$ for $x, p \in \mathbb{R}^d$. You can use code you find online, but you better be sure it is correct for this specific choice of kernel. Describe your implementation, and in particular which of the two main variants (cos and offset, or $\cos - \sin$) you use.

   Create a plot to assess the accuracy of these maps. The x-axis should show the number of dimensions you create m. The y-axis should measure the accuracy $|K(x,p) - \langle \hat{\phi}(x), \hat{\phi}(p) \rangle|$. Since this is a random process, you should somehow ensure you get an accurate representation of the performance in spite of the noise from randomness. The x-axis should start from $m = 2$ and get large enough so that the error gets to about 0.01.

2. [**20 points**] Consider a setting in $\mathbb{R}^2$. I want to learn a classifier defined by an ellipsoid, so points inside are labelled $+$ and points outside are labelled $-$. If the task was to find a ball classifier, then I can do this with a lifting map to $D = 3$ dimensions as $\phi(x,y) = (x, y, x^2 + y^2)$, and then find a linear classifier of the lifted points. Describe the minimal dimensional lifting map $\phi : \mathbb{R}^2 \to \mathbb{R}^D$ (that is so D is as small as possible) so that in the D-dimensional space the halfspaces could correspond with an ellipsoid in $\mathbb{R}^2$.

3. [**40 points**] Download or import a pre-trained word embedding model, such as GloVE (`https://nlp.stanford.edu/projects/glove/`) or word2vec of dimension at least 50. There are various ways to download and play with these. Specify which model you use.

Measure the cosine distance between a **large set** of words, selecting some that you think are similar, and some very different. Determine and report a threshold τ in cosine distance where word pairs with cosine distance less than that value are "close" and ones greater than that value are "far." Justify your answer. Explain how confident you are in this assessment; less convincing explanations may be docked a few points.