CS 6966: High-Dimensional Data Analysis **Due: Thursday 9.12 at 11:50pm**

Homework 1: Curse of Dimensionality

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. [**50 points**] Identify some phenomenon that encounters the curse of dimensionality. That is some property that seems to be or clearly is easy in low-dimensions, and as it is generalized to high-dimensions where the dimension d is a parameter, it becomes significantly harder as a function of d.
   You can use something we discussed in class; however, then you will be expected to go somewhat beyond what was done in the lecture. I encourage you to look for other phenomenon **beyond** what was covered in class – such new things will more easily achieve full credit. You can look online, but make sure you reproduce the parts below yourself.

   (a) Mathematically define or describe the phenomenon you have identified.

   (b) Show empirically some property getting harder as a function of d. This should be a plot or a table where you measure something for different values of d, and we can see how it gets harder as d increases. You only need to increase d large enough so that this effect becomes clear. Make sure the experimental set up is clear, and that all columns of a table or axis of a plot are clearly labeled and explained.

   (c) Explain mathematically why the phenomenon gets harder as the dimension d increases. This could be a formula describing it as a function of d (which should grow with d).

2. [**20 points**] What is the volume of the maximum size d-dimensional hypercube that can be placed entirely inside a unit radius ($r = 1$) d-dimensional ball? (Show your work.)

3. [**30 points**] Consider two point sets $A = \{a_1, a_2, \ldots, a_n\}$ and $B = \{b_1, b_2, \ldots, b_n\}$ both in $\mathbb{R}^d$. Assume they are linearly separable, which means that there exists halfspace h^* (all points on one side of a $(d-1)$-dimensional hyperplane) that contains all of A and none of B. In fact there may be many such halfspaces, and I want to be able to easily tell if another halfspace h also linearly separates A from B.

   (a) What is the minimal information I need to store about A and B so for a query halfspace h I can tell if they are linearly separated by h?

(b) How would this change, if I was ok with having a weaker guarantee, as follows: I am allowed to say TRUE if h separates at least a $(1-\varepsilon)$-fraction of the data in A and B (for $\varepsilon \in (0, 1/4)$) and I could sometimes be wrong, but I should be correct with probability at least $1 - \delta$.