Homework 6 - due Wed. Mar. 11th
Mathematical and Computational Foundations of Data Science

Instructor: Mauro Maggioni
Office: 302D Whitehead Hall
Web page: https://mauromaggioni.duckdns.org
E-mail: mauro.maggioni *at* youknowwhat.edu

Homework Policies. As in the first homework set.

Assignment

Least Squares, K-nearest neighbors. Study the materials discussed in class on Least Squares classifiers and Nearest Neighbor classifiers, for example from Hastie-Tibshirani-Friedman's *The Elements of Statistical Learning*, that I roughly followed in class (sec. 2.3).

Exercises

Exercise 1 (50pts). Please download the code on Least Squares and K-Nearest Neighbor classifier on my website. It contains two scripts, the second of which is the example I showed in class on looking at the distribution of pairwise distances of points sampled uniformly from a cube and from a sphere, as well as the distribution of distances from a fixed point (the first one sampled).
1. Study the behavior of mean and standard deviation of the distribution of pairwise distances as D changes (in both cases, cube and sphere). Then normalize the distribution so that for every D it has mean 0 and standard deviation 1, and superimpose the histograms of a standard Gaussian density to the histograms you obtain, and compute the difference $1/N \sum_{i=1}^{N} |h_i - g_i|$ where N is the number of bins, h_i is the height of the i-th bin for the distribution of pairwise distances of points in the cube (resp. sphere), and g_i is the height of the i-th bin for the Gaussian distribution you fit to the histogram. Study this difference as D grows.
2. Change the code and look at points sampled from $\mathcal{N}(0, I_D)$. Study the distributions, in particular look at their means, medians, and std as D grows. What do you observe/conclude? Is this what you expected? Compute $\mathbb{E}[||X||^2]$ when $X \sim \mathcal{N}(0, I_D)$, as well as $\mathbb{E}[||X - Y||^2]$ when $X, Y \sim \mathcal{N}(0, I_D)$, with X, Y independent, and match your calculation to your simulations. Then discuss how it may be the case that the distribution of $\mathcal{N}(0, \frac{1}{D} I_D)$ is similar to the uniform distribution on the sphere in $\mathbb{R}^D$.

Exercise 2 (50pts). Please download the code on Least Squares and K-Nearest Neighbor classifier on my website (this is the same as in Ex. 1, but just in case you are reading this before doing Ex. 1). It contains two scripts, the first of which is the example I showed in class on comparing a Least Squares (LS) classifier with a K-Nearest Neighbor (KNN). Go through the code and make sure it is clear what the code is doing. Recall that we have two probability measures that we called $p^{(1)}, p^{(2)}$ in class, each of which is a mixture (with equal weights) of ten Gaussian distributions with variance $\sigma^2 I_D$ (σ^2 is called varmod in the code). Let's write $p^{(i)} = \frac{1}{10} \sum_{l=1}^{10} g_l^{(i)}$ where $g_l = \mathcal{N}(\mu_l^{(i)}, \sigma^2)$. The $\mu_l^{(i)}$ are themselves drawn randomly (once for every run of the code), from a $\mathcal{N}(0, I_D)$ for $i = 1$ and from $\mathcal{N}((1, 0, \ldots, 0), I_D)$ for $i = 2$.
Change the code so that the sets of means are drawn from $\mathcal{N}(0, \lambda^2 I_D)$ for $i = 1$ and $\mathcal{N}((1, 0, \ldots, 0), \lambda^2 I_D)$ for $i = 2$.
Explore what happens as the dimension D of the ambient space increases. Can you find settings for σ^2 or λ^2 for which KNN performs consistently better than LS? for which LS performs consistently better than KNN? What can you say about the difference in performance? What about good choices of the value of K (the number of nearest neighbors) in the KNN algorithm?