Problem set 2
Stat 154, SP 14
Due date: Friday, 2/21, 3:45pm

Please turn HW in section or outside N. El Karoui's office, 311 Evans.

Problem 1

Use the voting data to get a 2 or 3-dimensional representation of the members of the House of representatives. Compare at least 3 methods and use at least 2 different measures of dissimilarities. (A natural measure of dissimilarity might be for instance $d_{i,j} = \frac{1}{N}\sum_{k=1}^{N}|v_{ik} - v_{jk}|$, where v_{ik} is the result of the k-th vote of legislator i, and N is the total number of votes you use in your analysis.)

Explain in detail your analysis and draw some conclusions.

The voting data is in the folder Data/Voting Data on bSpace.

Problem 2

Consider the following problem. We have a model such that $(x_i, y_i) \in \mathbb{R}^2$, with

$$y_i = f_0(x_i) + \epsilon_i \ .$$

Suppose ϵ_i is random, independent of x_i and such that $\mathbf{E}(\epsilon_i) = 0$.

Suppose first that $x_i \in (0, 2\pi)$ and $f_0(x) = \cos(10x) + 2$. Suppose $x_i = (i - 1/2)/N * 2\pi$, or $i = 1, \ldots, N$.

1. Start with $N = 10$ and $\epsilon_i \sim \mathcal{N}(0, \sigma^2 = .1)$. Draw y_i's correspondingly. Plot the k-nearest neighbors regression function, for $k \in \{1, 3, 10\}$.
   - Compute by simulation the EPE(π) and $\mathbf{E}(EPE(X))$ where X has the uniform distribution on $[0, 2\pi]$. Do this for the k-nearest neighbors rule for $k = 1, 3, 10$. Describe your methodology and use the same data to fit all three methods. Please also give a breakdown in terms of bias and variance.
   - Continue the previous experiment by fitting a constant function, a linear and a quadratic function to the dataset.
   - Of all 6 methods, which one gives you the best EPE? Comment on the result - trying to give an intuitive explanation.
2. Start with $N = 100$ and repeat the previous questions. How does your answer change? (Use $k = 1, 3, 10, 20, 50$ nearest-neighbor rules and fit polynomials of degree up to 5).

Repeat the previous experiments with $f_0(x) = \sin(x)$, $f_0(x) = .1 + .2x$.

Explain intuitively what you think will happen if σ^2 is increased. If you'd like you can run more simulations with σ^2 varying from .1 to .5 to 1.

Problem 3

Suppose now that we are in the same set-up as above by $x_i \in \mathbb{R}^p$. For instance, take $p = 5$. Suppose furthermore that $x_i \in [0, 2\pi]^p$.

Investigate the impact of dimensionality of the results in the previous problem. You can limit yourselves to studying one situation: e.g $N = 50$, x_i's picked uniformly at random in $[0, 2\pi]^5$, and $f_0(x) = \sum_{k=1}^{5}\sin(\sqrt{k}x_k) + \sum_{k=1}^{4}cos(x_k x_{k+1})$. Now ϵ_i are $\mathcal{N}(0, \sigma^2\mathrm{Id}_5)$, with $\sigma^2 = 1$. Compare a few nearest-neighbor methods to a linear fit in terms of EPE at one point of your choosing and $\mathbf{E}(EPE(X))$, where X is uniform on $[0, 2\pi]^5$.

Comment on your numerical results.

Problem 4 [Nearest-neighbors and high-dimension] a) Do the following simulation: draw 1,000 vectors in $\mathbb{R}^p$ where $p = 100$ with i.i.d $\mathcal{N}(0, 1)$ entries. Call the corresponding vectors X_i, $i = 1, \ldots, 1000$. Pick an i at random. Call it i_0. What can you say (analytically or numerically) about

$$\frac{1}{\sqrt{p}}\min_{j \neq i_0}\|X_{i_0} - X_j\| \text{ and } \frac{1}{\sqrt{p}}\max_{j \neq i_0}\|X_{i_0} - X_j\| \ ?$$

What can you say more generally about the distribution of $\|X_j - X_{i_0}\|$?

[Optional: Do you have a sense of how your results might change as $p \to \infty$, $n \to \infty$ and $p/n \to c$, where $c \in (0,1)$?]

b) [Though interesting, this question might be quite a bit harder than the other ones; MA students should attempt this question. Undergrads are not required to] Good partial answers to the following question will get full credit. Do not spend too much time on it. Using a similar simulation setup or analytic derivations, consider the following problem: draw $X_1, \ldots, X_n$ i.i.d at random in $\mathbb{R}^p$. Pick $k \in \mathbb{N}$. Call $x_0 = \mathbf{E}(X_1) = \ldots = \mathbf{E}(X_n)$ and $S_{x_0,k}$ the set of points with the same k nearest neighbors as x_0. What can you say $d_k(x_0) = \max_{x \in S_{x_0,k}} \|x - x_0\|_2$? *(Hint: try to characterize it geometrically. You might find it interesting (or not) to read about Loewner-John ellipsoids. Also, try to find a simple (not completely trivial) lower bound for $d_k(x_0)$. Can you find examples where $d_k(x_0) = \infty$?)*

Do this for $n = 1000$, $p \in \{1, 10, 100\}$ and $k \in \{1, 3, 10\}$. To make things concrete, you could take the entries of X_i's to be i.i.d either $\mathcal{N}(0,1)$ or Unif$[-1,1]$. *(Hint: even if this proves too difficult, explain at least how you would try to solve the problem; for instance, if you go the numerical route, what is the optimization problem you have to solve.)*

Explain why knowing this diameter will help you assess the likely performance of nearest-neighbor methods. Suppose you wanted this $d_k(x_0)$ to be less than a given ϵ. (Explain why this might be a natural constraint.) Give an estimate of the sample size n that you need as a function of p and k. Again you can limit yourselves to X_i's drawn at random with i.i.d entries as above. The estimate could be a function obtained by numerical simulations, or the result of analytic derivations (in which case asymptotics are OK). Summarize your findings in lay-man terms. *(Hint: getting a very coarse upper bound on $d_k(x_0) = \max_{x \in S_{x_0,k}}$ should help you. If p is small and n is very large, you can also do asymptotics as $n \to \infty$. Finally, you can also try to understand what happens if n grows like* $\exp(p)$*. A natural question that could help here is: how many points to you need to put on a unit sphere in $\mathbb{R}^p$ so that any point on that sphere is within ϵ of one of your points. Some people call this an ϵ-net.)*