Homework 4 - due Wed. Feb. 26th
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

SVD and Least Squares. Study the materials presented in class on Singular Value Decomposition and its applications as Principal Component Analysis and the solution of Least Squares problems.

Exercises

Exercise 1 (Vandermonde Matrix, revisited). Back to the Vandermonde matrix discussed in the first homework set. Fix $\{x_i\}_{i=1}^m$ points in an interval I (e.g. $I = [-1, 1]$). Consider the map $\mathbb{R}^{d+1} \to \mathbb{R}^m$ defined by $c = (c_0, \ldots, c_d) \mapsto (p_c(x_1), \ldots, p_c(x_m))$, where $p_c = \sum_{l=0}^d c_l x^l$, i.e. from vectors of coefficients of polynomials of degree $\leq d$ to vectors $(p_c(x_i))_{i=1}^m$ of values of such polynomials. From the first homework set, we know that:

- this map is linear
- this map is represented, upon choosing the standard basis in $\mathbb{R}^{d+1}$ and $\mathbb{R}^m$, by the $m \times d$ Vandermonde matrix

$$A = \begin{bmatrix} 1 & x_1 & x_1^2 & \ldots & x_1^d \\ 1 & x_2 & x_2^2 & \ldots & x_2^d \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_m & x_m^2 & \ldots & x_m^d \end{bmatrix}.$$

  Therefore, if $c = [c_0; c_1; \ldots; c_d]$ is the column vector of coefficients of p, i.e. $p(x) = c_0 + c_1 x + c_2 x^2 \cdots + c_d x^d$, then Ac gives the the sampled polynomial values, i.e. $(Ac)_i = p(x_i)$, for $i = 1, \ldots, m$.

Suppose there is a (say, continuous) function $f : I \to \mathbb{R}$, and we have access to $(x_i, y_i)_{i=1,\ldots,n}$, where $y_i = f(x_i) + \sigma\eta_i$, where $\eta_i \sim \mathcal{N}(0, 1)$, and σ is a (fixed) positive number (standard deviation of the noise η_i).

We want to construct a polynomial $\hat{f}$ of degree $\leq d$ that approximates f in the least squares sense.

- Write down a least squares problem for $\hat{f}$ given the data $\{(x_i, y_i)\}_{i=1}^m$, as we did in class, in the form $Ac = y$ (what are c and y? which dimension do they have and what do they represent?)
- Write code implementing this computation, that solves the least squares problem using SVD, in the case of $d < m$.
- Let $f(x) = \sin(x)$, and the interval $I = [-1, 1]$. Fix $d = 10$, and let $x_1, \ldots, x_m$ be drawn uniformly at random on I. Fix $\sigma = 0.05$. Compute $\hat{f}$, and note that it is random, because of the randomness in the points x_i and in the noise η_i (to be sure, run the code a few times and visualize the corresponding $\hat{f}$'s obtained with each run). In fact, let's denote $\hat{f}_m := \hat{f}$, as the dependency on m, the number of points, will be in focus. Plot the mean and standard deviation of $||\hat{f}_m - f||_{L^2([-1,1])}$ as a function of m, and comment on their behavior as m grows (to compute these L^2-norms, use a function to compute the relevant integral already available in the software application you are using). [Note that at this point you are using the true f, but you have not used anything but the y_i's to produce $\hat{f}_m$.]