ASSIGNMENT 1

(100 POINTS) DUE BY MONDAY, DECEMBER 21, 2020, 11:30 AM

Assignment Policy:

- No late submission is allowed.
- Collaboration is okay, however, students must submit assignment in their own words. Plagiarism in any form will not be accepted.
- All questions carry equal points.

Question 1 (Low Dimensional Reconstruction with PCA)**.** Here we will study the effects of PCA low-dimension reconstruction using MNIST Training Dataset[1] In this problem we will use MNIST train dataset only for reconstruction.

- Use PCA without input data normalization to reduce dimension to 1,10, 20, 50, 100 and 200, and project back to 784. Report the MSE for reconstruction, and plot the original and reconstructed digits.
- Use PCA with input data normalization to reduce dimension to 1,10, 20, 50, 100 and200 , and project back to 784. Report the MSE for reconstruction, and plot the original and reconstructed digits.

Question 2 (Comparing Dimensionality Reduction Methods)**.** In this problem we will compare some of the different dimensionality reduction methods which we have studied so far in the course.

- Use MNIST Training dataset (as in Question 1), and visualize and compare their 2-D projection with the following methods (you can use built-in packages in sklearn): (a) PCA (b) K-PCA (kernal=rbf) (c) Isomap (c) MDS (d) t-SNE (perplexity = 10, 20, 30, 40, 50). Please report the running times and intuitively compare and reason about the projection qualities.
- Generate the S-curve using the built-in routine in sklearn. Visualize and compare their 2-D projection with the following methods (you can use built-in packages in sklearn): (a) PCA (b) K-PCA (kernal=rbf) (c) Isomap (c) MDS (d) t-SNE (perplexity = 10, 20, 30, 40, 50). Please report the running times and intuitively compare and reason about the projection qualities.

Question 3 (Convergence of Histogram Density Estimator)**.** Suppose that $x \in [0, 1]$ and the underlying density $f(x)$ is absolutely continuous and that $\int (f'(u))^2 du < \infty$. Let risk is denoted by $R(\hat{f}_n, f) = \int (\hat{f}_n(x) - f(x))^2 dx$, where n denotes the number of samples and $\hat{f}_n$ is the corresponding histogram estimate with the bandwidth h. Prove that:

$$R(\hat{f}_n, f) = \frac{h^2}{12} \int (f'(u))^2 du + \frac{1}{nh} + o(h^2) + o(\frac{1}{n}).$$

The optimal value h^* that minimizes the above expression (ignoring the higher order terms is:

$$h^* = \frac{1}{n^{\frac{1}{3}}} \Big(\frac{6}{\int (f'(u))^2 du} \Big)^{\frac{1}{3}}$$

With the above optimal choice of h^*,

$$R(\hat{f}_n, f) = \frac{C}{n^{\frac{2}{3}}},$$

where $C = (\frac{3}{4})^{\frac{2}{3}} \Big(\int (f'(u))^2 du \Big)^{\frac{1}{3}}$.

[1] http://yann.lecun.com/exdb/mnist/. Please note several deep learning platforms offer an inbuilt MNIST downloader which helps access MNIST dataset efficiently.

[**Hint**:*You have to decompose risk as bias and variance and use the results,* $\boldsymbol{E}(\hat{f}_n(x)) = \frac{p_j}{n}$ *and* $\boldsymbol{V}(\hat{f}_n(x)) = \frac{p_j(1-p_j)}{nh^2}$. *You may have to invoke the Taylor series expansion and Mean Value Theorem many times.*]

Question 4 (Convergence of Kernel Density Estimator)**.** Let $\hat{f}(x)$ denote the kernel density estimator with kernel K and bandwidth h, of the density $f(x)$ ($x \in [0,1]$), such that f'' is absolutely continuous and that $\int (f'''(x))^2 dx < \infty$. Let $R_x = \boldsymbol{E}(\hat{f}(x) - f(x))^2$ be the risk at a point x and let $R = \int R_x dx$ denote the integrated risk. The kernel K satisfies the usual assumptions, i.e., $\int K(x)dx = 1$, $\int xK(x)dx = 0$, and $\sigma_K^2 = \int x^2 K(x)dx > 0$. Then,

$$R = \frac{1}{4}\sigma_K^4 h^4 \int (f''(x))^2 dx + \frac{\int K^2(x)dx}{nh} = O(\frac{1}{n}) + O(h^6))$$

Prove further that the optimal bandwidth decreases as $n^{-\frac{1}{5}}$ and the optimal risk $R = O(n^{-\frac{4}{5}})$, better than that of histogram density estimator.

[**Hint**: *First try to prove an equivalent expression of* R_x *and then integrate to get the expression for* R. *Then use similar Taylor series expansion approach as for Question 3.*]