Review topics for final exam
AS.110.446, EN.550.416

Dr. Mauro Maggioni
Office: 405 Krieger Hall
Web page: www.math.jhu.edu/~mauro
E-mail: myfirstname.maggioni@jhu.edu

The final exam is, as per the Registrar's schedule, on May 11th, 9AM to 12PM. The final exam is designed so that 1.5 hrs should suffice.

Topics

1. Curse of dimensionality: discuss an instance of the curse of dimensionality, and why it makes certain estimation/classification problems hard.

2. Dimension reduction: think about the techniques for dimension reduction that we have discussed. These include principal components, random projections, diffusion maps, isomap. What are the motivations for these techniques? What does each of them achieve? What do you know about the performance of each technique, and under which geometric assumptions on the data do these guarantees hold? Which technique has the strongest geometric assumptions on the data? Which one the weakest? What can you say about algorithmic similarities and differences between these techniques? What technique would you use if the goal was clustering data? Which technique would you use if you have n data points in $\mathbb{R}^D$ dimensions, with $D \gg n$, but the data is not intrinsically low-dimensional?

3. Regression: what is a regression problem? What is an estimator for the function to be regressed? Describe one or more examples of estimators, and any guarantees that you know about such estimator. Discuss, if you can, how the performance depends on the dimension of the space where the data lies (if you do not remember an exact result, discuss your expectations on the behavior of the performance with the dimension).

4. Discuss spectral clustering: how is it performed? what are the basic ingredients in the algorithm? describe carefully the construction of random walks on a data set, and its connection with spectral clustering. Compare spectral clustering with K-means, and discuss an example where spectral clustering will perform well while K-means will perform poorly. Discuss an example where spectral clustering will perform poorly, even if one could argue there are two (say) well-separated clusters. Discuss the computational complexity of spectral clustering. Discuss how spectral clustering depends on the dimension of the ambient space, vs. the intrinsic dimension of the data.