ASTR/PHYS 356

Problem Set 6

Posted: April 17, 2018 Due: April 28, 2017

Remember to give proper credit and properly reference other people's codes if you use them.

Problem 1: Regression and model selection (the frequentist way) (15 ptgs.)
Download file ps5_problem1.dat from the Canvas site and conduct a regression analysis with these data. Assume Gaussian errors with a 1-σ of 1.3.

a) Use different polynomials with degrees from 0 to 10 to fit the data and plot the fits to the data using three of the different models (similar to what is shown in Figure 8.13 of the AstroML book).
b) Plot the reduced chi-sq., adjusted coefficient of determination, AIC and BIC as function of the polynomial degree (similar to the lower panel in Figure 8.14 of the AstroML book).
c) Conduct a cross-validation analysis similar to that shown in Figure 8.14 of the AstroML book and decide which is the best model to fit the data from your analysis. Explain in detail your analysis and how you came to your conclusion.

Produce learning curves (similar to the ones in Figure 8.15 of the AstroML book) for three of the polynomial fits. Discuss what these plots reveal.

Problem 2: KDE and clustering (25 ptgs.)
In the following problem you will use the COMBO-17 (Classifying Objects by Medium-Band Observations in 17 filters) dataset from Wolf et al. 2004, A&A, 421, 913 (see: http://adsabs.harvard.edu/abs/2004A%26A...421..913W), which I have uploaded to the Canvas site. This is the first public catalogue of a large dataset (with 63,501 objects) with brightness measurements in 17 bands. Note that the Sloan Digital Sky Survey provides a much larger dataset, by 4 to 5 orders of magnitudes, but with measurements in "only" five bands. Hence, the COMBO-17 dataset has much less objects, but many more dimensions, than the SDSS dataset. For this problem we are interested in galaxies with redshift (z) less than 0.3, and the absolute magnitude values of these low-redshift galaxies in the Johnson B-band and the UV band at 280nm.

a) Make a color-magnitude (i.e., M280 - MB vs. MB) plot for these low-z galaxies.
b) Use KDE and nearest-neighbor density estimation to plot the data. Show these two plots and compare them with the original plot that shows the individual galaxies (from part a). Explain how you chose your kernel bandwidth (which should be based in oe of the methods we discussed in class) for the KDE process and value of K for the nearest-neighbor density estimations. Explain what you see in these plots in words. You should be able to see the so-called red sequence (in the upper left part of the plots) and the so-called blue sequence (in the lower right part of the plots) as two (more or less) distinct clusters of points.
c) Conduct a K-means cluster analysis of the color-magnitude points. Use one of the ways discussed in class to decide on the number of clusters (i.e., K number in the K-means cluster analysis). Show a new plot of the magnitude-color diagram where you identify the different cluster members with different symbols or colors.

d) Conduct a mean shift clustering analysis of the data. Explain how you selected the kernel width. Show a new plot of the magnitude-color diagram where you identify the different cluster members with different symbols or colors. Compare your findings with those from the K-means analysis of part c.
e) Conduct an agglomerative hierarchical clustering analysis on the color-magnitude plot. Show the dendogram and decide where you think it makes the most sense (scientifically) to "cut the tree". How many clusters do you get? Show a new plot of the magnitude-color diagram where you identify the different cluster members with different symbols or colors. Identify these clusters in the dendogram diagram. How does your answer compare to the other two clustering algorithms?
f) Conduct a Gaussian mixture model clustering analysis. Discuss, in detail, how you choose the number of Gaussians in your model. Compare the results with the other clustering algorithms.