Statistics, Data Analysis, and Machine Learning for Physicists
Timothy Brandt
Spring 2020

Homework 4 (optional, in lieu of final project)

Part 1: *Gaussian Process Regression to Fix Images*

In astronomy, we often deal with images, and occasionally these images have bad pixels. Spectra are, at their root, images that have been dispersed in one direction (which corresponds to wavelength or frequency) and collapsed along the other. We then try to analyze these images. Sometimes these images have bad pixels. Usually we know which pixels are bad, so that part is fine, but then we need to deal with them. The best approach would be to ignore them (for example, by setting their variance to infinity), but sometimes that isn't possible. Recall, for example, principal component analysis. So people often try to "fix" their bad data. The problem is that the fixes are often clumsy and might be unreliable. For example, one of the most common "fixes" is to replace a bad pixel with the median of its eight neighbors.

Try to come up with a better approach to fixing bad pixels using Gaussian process regression. You may use the images provided–they are the same ones from Homework 3 but cropped a bit. Your approach will be to predict the value of a given pixel without using that pixel, and then to compare your prediction with the pixel's actual value.

a. Start with the commonly-used median replacement. Choose a pixel at random and compare its value to the median of its eight neighbors. Repeat this with many random points over many images to make a histogram of the differences between the median-replaced value and the pixel value. Plot this histogram.

b. Next implement a Gaussian process regression around a random pixel. You don't want to do Gaussian process regression over the entire image ($\sim 100 \times 100$ pixels), since that would produce a $\sim 10^4 \times 10^4$ covariance matrix that you would need to invert. Instead, choose a 9×9 region centered on your random pixel (make sure it's at least four pixels away from the edge). Use Gaussian process regression on this small region (making sure *not* to use the central pixel) to estimate the value of the central pixel. Do this many times to make a histogram of the residuals. You'll want to assume a covariance for your data–you may use the identity matrix multiplied by a constant, say, 5^2.

c. Now tune your Gaussian process regression by varying the hyperparameters of your covariance function. You may stick with the squared exponential covariance function, or you may try something else (check

   `https://en.wikipedia.org/wiki/Gaussian_process#Usual_covariance_functions`

   for a set of common covariance functions). See how narrow you can get the histogram of residuals. Once you have optimized your choice of covariance function and parameters, use another set of random pixels to re-compute the histogram of residuals and verify that you reproduce your previous, optimized histogram (this is like using a set of test data). If you want, you can also try to tune your assumed covariance matrix for the data.

d. To be useful, a technique for fixing images should be reasonably fast. To speed up your routine, see how your distribution of residuals depends on the size of the patch you use around each pixel, and suggest the smallest patch that produces results indistinguishable from your initial, 9×9 patch (it need not be square).

e. Finally, compare your results with your tuned Gaussian process regression to median substitution. Can you do better, and if so, by how much? Note that you will never be able to achieve perfect agreement (don't bother trying), since each pixel value has noise associated with it.

Part 2: *A Mixture Model and Bootstrapping, Redux*

Download the data set supplied; these are the motions of stars in the plane of the sky. Right ascension and declination are two orthogonal directions; you can think of the measurements as $\dot{x}$ and $\dot{y}$. The motions in these two directions have been measured two different ways for each star, and the errors have been estimated. There are a couple of problems:

1. There is an overall offset between the measurements of the velocities;
2. The errors have been underestimated; and
3. Some of the stars are outliers that should be excluded.

Your task is to measure the offsets between the sets of measurements (and estimate their uncertainty), estimate a correction factor to the uncertainties, and decide which stars should be excluded as outliers.

If the errors were correct and Gaussian and there were no outliers, then you would be fitting for two parameters, $\Delta\dot{x}$ and $\Delta\dot{y}$, and could minimize

$$\chi^2_x = \sum_i \frac{(\dot{x}'_i - \dot{x}''_i - \Delta\dot{x})^2}{\sigma'^2_x + \sigma''^2_x} \tag{1}$$

and an equivalent equation for Δy. These would be the systematic offsets between $\dot{x}'$ and $\dot{x}''$, the velocities measured by the two different instruments.

To do this problem, use a mixture model to fit the data, where there is a probability g for each measurement that it is good and a probability $1-g$ that it is an outlier. Use a broad Gaussian distribution for the outliers (say, with $\sigma = 2$). Also, fit for two additional parameters: a constant factor a multiplying σ_x^2 and σ_y^2, and an additive variance, so that the true errors are given by $a^2\sigma_x^2 + b^2$ and $a^2\sigma_y^2 + b^2$.

a. Write down the likelihood for the mixture model. Don't forget the normalization of the Gaussian distribution with a and b in it, as the normalization will include the extra factors multiplying and adding to the errors (χ^2 is insufficient).

b. What are best-fit additional error parameters a and b, and what are the best-fit offsets $\Delta\dot{x}$ and $\Delta\dot{y}$?

c. Use bootstrap resampling of your data to obtain errors on all of these parameters.

Part 3: *Measure your Evidence, Watch your Assumptions*

Suppose that you would like to fit for the mass of an unseen black hole at the center of a star cluster. This paper has done so: https://arxiv.org/pdf/1702.02149.pdf Look at Figure 3, where they compare the likelihood of a model with and without an extra black hole (an extra free parameter). Assuming that the likelihood ratio is 10, compute the significance of the black hole in terms of the improvement in the Akaike Information Criterion and in the Bayesian Information Criterion (there were 19 pulsars fit, so assume that there are 19 measurements for the BIC). Now, determine the strength of the evidence in favor of the black hole model using $\exp[\Delta\text{BIC}]$ and $\exp[\Delta\text{AIC}]$. How strongly is the black hole model preferred, and how many Gaussian σ does this correspond to in each case?

Now look at Figure 4 from the same paper. The left plot is not a posterior of the mass, as claimed, but is actually a log-normal fitted to the peak of the two-dimensional distribution on the right. Given that this is a log-normal, what probability will the authors assign to the black hole having zero mass (i.e. not existing)? What should they have done instead using the distribution on the right to produce the plot on the left?

Being careful with your assumptions and statistics can make a very big difference. Perhaps unsurprisingly, another paper came out a few months after this one and showed no evidence in favor of a black hole in this cluster. I've picked on Nature twice in this course, but the journal has a well-deserved reputation for publishing sexy but shoddy research in astronomy.