Statistics Mini-course Problem Set 2

Due on Fri. Apr 22

We will do some exercises related to model selection, density estimation, and basis-function fitting. You should solve these exercises on a computer and the best way to hand in the problem set is as an `ipython notebook`. Rather than sending me the notebook, you can upload it to `GitHub`, which will automatically render the notebook. Rather than starting a repository for a single notebook, you can upload your notebook as a `gist`, which are version-controlled snippets of code.

If you want to upload your notebook as a gist from the command-line, you can use the package at this http URL and use it as follows. Log into your `GitHub` account:

```
gist --login
```

and then upload your notebook `statminicourse_2016_PS2_YOURNAME.ipynb` as

```
gist statminicourse_2016_PS2_YOURNAME.ipynb
```

If you want to make further changes, you can clone your gist in a separate directory and use it as you would any other git repository.

Problem 1: Download the data set `here`. This data set has a set of values (x, y, σ_y). Determine the best order of polynomial to fit these data using the AIC and BIC criteria and using cross validation. Go up to order 20.

Problem 2: Repeat Problem 1, but rather than fitting a polynomial, fit the data using a sum of sines and cosines. That is, fit $y(x) = a_0 + \sum_{k=1}^{K} a_k \sin(kx) + b_k \cos(kx)$, with a_k and b_k free parameters. Determine the best K using AIC, BIC, and cross validation, considering K up to 10. Compare the best-fit from this problem to that from Problem 1.

Problem 3: Estimate the density of the x values in the data set using a Kernel Density Estimate. Determine the best band-width using cross validation for both an Epanechnikov and a Gaussian kernel. Compare the resulting density distributions.