Your Name
email@fas.harvard.edu
CS181-S18

Assignment #1, v1.3
Due: 11:59pm February 2, 2018

Collaborators: John Doe, Fred Doe

Homework 1 Solutions

Grading Instructions

In the solutions, you will see several highlighted checkpoints. These each have a label that corresponds to an entry in the Canvas quiz for this problem set. The highlighted statement should clearly indicate the criteria for being correct on that problem. If you satisfy the criteria for a problem being correct, mark "Yes" on the corresponding position on the Canvas quiz. Otherwise, mark "No". Your homework scores will be verified by course staff at a later date.

Introduction

This homework is on different forms of linear regression and focuses on loss functions, optimizers, and regularization. Linear regression will be one of the few models that we see that has an analytical solution. These problems focus on deriving these solutions and exploring their properties.

If you find that you are having trouble with the first couple problems, we recommend going over the fundamentals of linear algebra and matrix calculus. We also encourage you to first read the Bishop textbook, particularly: Section 2.3 (Properties of Gaussian Distributions), Section 3.1 (Linear Basis Regression), and Section 3.3 (Bayesian Linear Regression). (Note that our notation is slightly different but the underlying mathematics remains the same :).

Please type your solutions after the corresponding problems using this LaTeX template, and start each problem on a new page. You will submit your solution PDF and at most 1 `.py` file per problem (for those that involve a programming component) to Canvas.

Problem 1 (Priors and Regularization,15pts)

In this problem we consider a model of Bayesian linear regression. Define the prior on the parameters as,

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \mathbf{0}, \tau_w^{-1}\mathbf{I}),$$

where τ_w is as scalar precision hyperparameter that controls the variance of the Gaussian prior. Define the likelihood as,

$$p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) = \prod_{i=1}^{n} \mathcal{N}(y_i \mid \mathbf{w}^\mathsf{T}\mathbf{x}_i, \tau_n^{-1}),$$

where τ_n is another fixed scalar defining the variance.

(a) Using the fact that the posterior is the product of the prior and the likelihood (up to a normalization constant), i.e.,

$$\arg\max_{\mathbf{w}} \ln p(\mathbf{w} \mid \mathbf{y}, \mathbf{X}) = \arg\max_{\mathbf{w}} \ln p(\mathbf{w}) + \ln p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}).$$

Show that maximizing the log posterior is equivalent to minimizing a regularized loss function given by $\mathcal{L}(\mathbf{w}) + \lambda\mathcal{R}(\mathbf{w})$, where

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{w}^\mathsf{T}\mathbf{x}_i)^2$$

$$\mathcal{R}(\mathbf{w}) = \frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{w}$$

Do this by writing $\ln p(\mathbf{w} \mid \mathbf{y}, \mathbf{X})$ as a function of $\mathcal{L}(\mathbf{w})$ and $\mathcal{R}(\mathbf{w})$, dropping constant terms if necessary. Conclude that maximizing this posterior is equivalent to minimizing the regularized error term given by $\mathcal{L}(\mathbf{w}) + \lambda\mathcal{R}(\mathbf{w})$ for a λ expressed in terms of the problem's constants.

(b) Notice that the form of the posterior is the same as the form of the ridge regression loss

$$\mathcal{L}(\mathbf{w}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^\top(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda\mathbf{w}^\top\mathbf{w}.$$

Compute the gradient of the loss above with respect to $\mathbf{w}$. Simplify as much as you can for full credit. Make sure to give your answer in vector form.

(c) Suppose that $\lambda > 0$. Knowing that $\mathcal{L}$ is a convex function of its arguments, conclude that a global optimizer of $\mathcal{L}(\mathbf{w})$ is

$$\mathbf{w} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y} \tag{1}$$

For this part of the problem, assume that the data has been centered, that is, pre-processed such that $\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$.

(d) What might happen if the number of weights in $\mathbf{w}$ is greater than the number of data points N? How does the regularization help ensure that the inverse in the solution above can be computed?

Solution

(a) $p(\mathbf{w})$ is a multivariate normal distribution. Plug the mean $\mathbf{0}$ and covariance matrix $\tau_w^{-1}\mathbf{I}$ into the PDF of multivariate normal distribution:

$$p(\mathbf{w}) = \frac{1}{(|2\pi\tau_w^{-1}\mathbf{I}|)^{1/2}} \exp(-\frac{1}{2}\mathbf{w}^\top(\tau_w^{-1}\mathbf{I})^{-1}\mathbf{w})$$

$$\ln p(\mathbf{w}) = -\frac{1}{2}\mathbf{w}^\top\mathbf{w}\tau_w + constant = -\tau_w R(\mathbf{w}) + constant$$

Check 1a: You correctly took the log of $p(\mathbf{w})$. It is acceptable to write out the constants, or just write out "+ constant" as done in the solution.

Similarly,

$$p(\mathbf{y}\,|\,\mathbf{X},\mathbf{w}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\tau_n^{-1}\pi}} \exp(-\frac{(y_i - \mathbf{w}^\top\mathbf{x}_i)^2}{2\tau_n^{-1}})$$

$$\ln p(\mathbf{y}\,|\,\mathbf{X},\mathbf{w}) = -\frac{1}{2}\sum_{i=1}^{n}(y_i - \mathbf{w}^\top\mathbf{x}_i)^2\tau_n + constant = -\tau_n L(\mathbf{w}) + constant$$

Check 1b: You correctly take the log of $p(\mathbf{y}|\mathbf{X},\mathbf{w})$. It is acceptable to write out the constants, or just write out "+ constant" as done in the solution.

Therefore, maximizing $\ln p(\mathbf{w}) + \ln p(\mathbf{y}\,|\,\mathbf{X},\mathbf{w})$ is equivalent to maximizing $-\tau_n L(\mathbf{w}) - \tau_w R(\mathbf{w})$. Hence it is equal to minimizing $L(\mathbf{w}) + \lambda R(\mathbf{w})$, where $\lambda = \tau_w/\tau_n$. (Note that $\tau_n > 0$ because it is a variance.).

Check 1c: You correctly show how the sum of those two is equal to minimizing $L(\mathbf{w}) + \lambda R(\mathbf{w})$.

(b) Rewrite the original expression as such and compute the gradient of each term:

$$\underbrace{\mathbf{y}^\top\mathbf{y}}_{\text{gradient is 0}} \underbrace{-\mathbf{y}^\top\mathbf{X}\mathbf{w} - (\mathbf{X}\mathbf{w})^\top\mathbf{y}}_{-2\mathbf{X}^\top y} + \underbrace{\mathbf{w}^\top\mathbf{X}^\top\mathbf{X}\mathbf{w}}_{2\mathbf{X}^\top\mathbf{X}\mathbf{w}} + \underbrace{\lambda\mathbf{w}^\top\mathbf{w}}_{2\lambda w}$$

Check 1d: The following expression gives you full credit (up to reordering of the terms):

$$\frac{\partial\mathcal{L}(\mathbf{w})}{\partial\mathbf{w}} = -2\mathbf{X}^\top\mathbf{y} + 2\mathbf{X}^\top\mathbf{X}\mathbf{w} + 2\lambda\mathbf{w}$$

(c) Function $\mathcal{L}$ is convex in $\mathbf{w}$ and therefore any local minimum is also a global minimum. Furthermore, any point $\mathbf{w}^*$ where the gradient is 0 is a local minimum. Solving for

$$\frac{\partial\mathcal{L}(\mathbf{w})}{\partial\mathbf{w}} = 0 \Leftrightarrow -2\mathbf{X}^\top\mathbf{y} + 2\mathbf{X}^\top\mathbf{X}\mathbf{w} + 2\lambda\mathbf{w} = 0 \Leftrightarrow \mathbf{X}^\top\mathbf{y} - (\mathbf{X}^\top\mathbf{X} + \lambda I)\mathbf{w} = 0 \Leftrightarrow \mathbf{w} = (\mathbf{X}^\top\mathbf{X} + \lambda I)^{-1}\mathbf{X}^\top\mathbf{y}$$

yields the above solution.

- **Check 1e**: You state that the optimum is found by setting the gradient in 1.b equal to 0.
- **Check 1f**: You solve the matrix equation correctly.

(d) If the number of weights in $\mathbf{w}$ is greater than the number of data points N, then $\mathbf{X}$ has more columns than rows and therefore does not have full column rank (i.e. $\mathbf{X}$ does not have linearly independent columns). This means that $\mathbf{X}^\top\mathbf{X}$ is not necessarily positive definite [1] and therefore not necessarily invertible. Adding the $\lambda\mathbf{I}$ regularization term guarantees positive-definiteness, ensuring that $\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}$ is invertible, as long as $\lambda > 0$.

- **Check 1g**: You correctly justify that $(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})$ guarantees invertibility whereas $(\mathbf{X}^\top\mathbf{X})$ without the regularization term does not.

[1] Explanation

2. Modeling Changes in Congress

The objective of this problem is to learn about linear regression with basis functions by modeling the number of Republicans in the Senate. The file `data/year-sunspots-republicans.csv` contains the data you will use for this problem. It has three columns. The first one is an integer that indicates the year. The second is the number of sunspots. The third is the number of Republicans in the Senate. The data file looks like this:

```
Year,Sunspot_Count,Republican_Count
1960,112.3,36
1962,37.6,34
1964,10.2,32
1966,47.0,36
1968,105.9,43
1970,104.5,44
```

and you can see plots of the data in Figures 1 and 2.

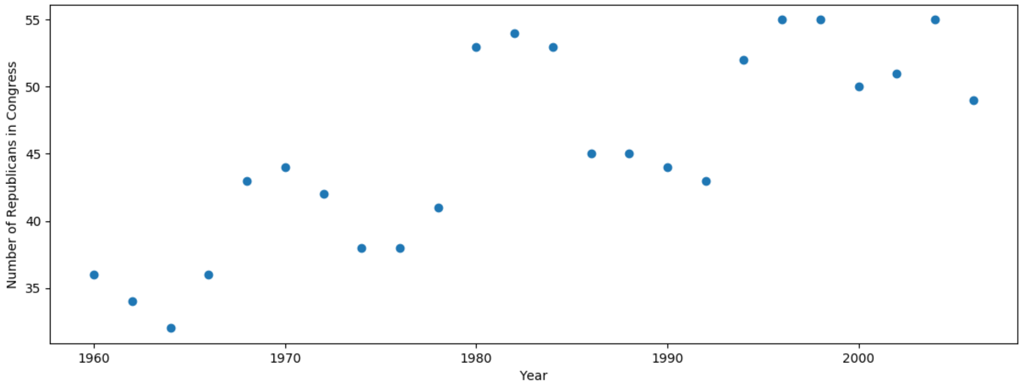


Figure 1: Number of Republicans in the Senate. The horizontal axis is the year, and the vertical axis is the number of Republicans.

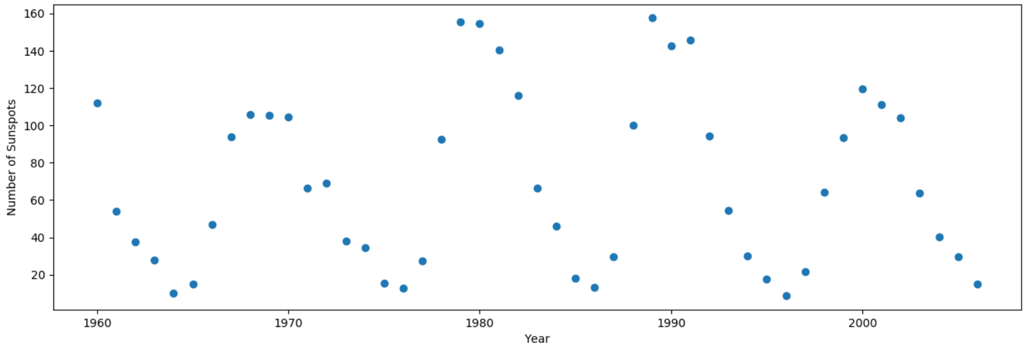


Figure 2: Number of sunspots by year. The horizontal axis is the year, and the vertical axis is the number of sunspots.

Data Source: http://www.realclimate.org/data/senators_sunspots.txt

Problem 2 (Modeling Changes in Republicans and Sunspots, 15pts)

Implement basis function regression with ordinary least squares for years vs. number of Republicans in the Senate. Some sample Python code is provided in `linreg.py`, which implements linear regression. Plot the data and regression lines for the simple linear case, and for each of the following sets of basis functions (only use (b) for years, skip for sunspots):

(a) $\phi_j(x) = x^j$ for $j = 1, \dots, 5$

(b) $\phi_j(x) = \exp \frac{-(x-\mu_j)^2}{25}$ for $\mu_j = 1960, 1965, 1970, 1975, \dots 2010$

(c) $\phi_j(x) = \cos(x/j)$ for $j = 1, \dots, 5$

(d) $\phi_j(x) = \cos(x/j)$ for $j = 1, \dots, 25$

In addition to the plots, provide one or two sentences for each with numerical support, explaining whether you think it is fitting well, overfitting or underfitting. If it does not fit well, provide a sentence explaining why. A good fit should capture the most important trends in the data.

Next, do the same for the number of sunspots vs. number of Republicans, using data only from before 1985. What bases provide the best fit? Given the quality of the fit, would you believe that the number of sunspots controls the number of Republicans in the senate?

Solution

- See linreg-soln.py
- **Check 2a**: Your graph should match the graph for (a) below.
- **Check 2b**: Your graph should match the graph for (b) below.
- **Check 2c**: Your graph should match the graph for (c) below.
- **Check 2d**: Your graph should match the graph for (d) below.

For years vs. number of Republicans in the Senate:

(a) underfits. It has a loss of 424.87, which is significantly worse than some of the other regressions. We can see graphically that the regression line fails to capture the sinusoidal behavior of the data, indicating an underfit.

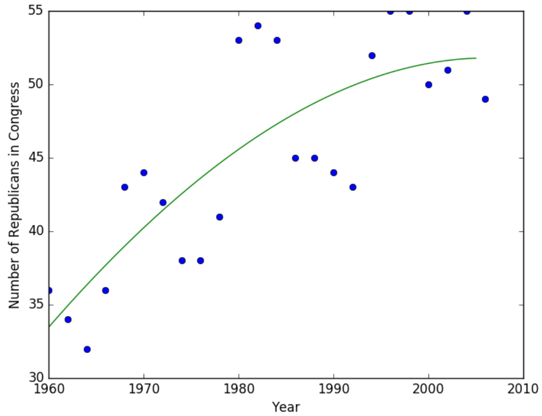
(b) fits well/slightly underfits (both answers are acceptable). It has a loss of 54.27, a significant improvement, and it captures the sinusoidal behavior of the data, although it doesn't capture this behavior quite as well for data points around the year 2000 and onward, suggesting a slight underfit.

(c) underfits. It has a loss of 1082.81, far worse than that of the other regressions. We can indeed see graphically that the regression line does a very poor job of capturing any trends in the data.

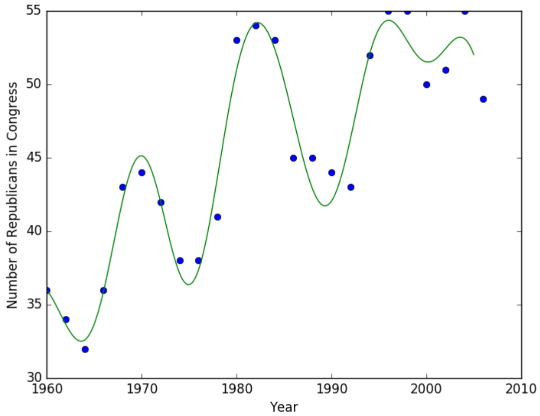
(d) fits well. It has a low loss of 38.83 (or similar–see Check 2.8), and it adequately captures the sinusoidal behavior of the data without capturing unnecessary noise in the data.

- **Check 2e**: Your explanation for (a) should be that it underfits. Must include a loss of 424.87 and a qualitative description.
- **Check 2f**: Your explanation for (b) should be that it is either a good fit, or that is slightly underfits. Must include a loss of 54.27 and a qualitative description.

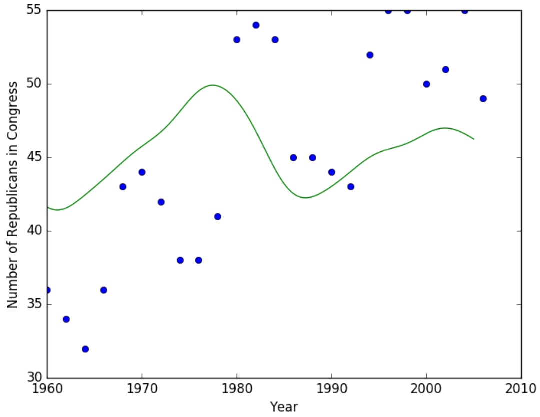
- **Check 2g**: Your explanation for (a) should be that it underfits. Must include a loss of 1082.81 and a qualitative description.
- **Check 2h**: Your explanation for (d) should be that it is a good fit, since it seems to match the data well without overly hugging the points. Must include a loss of 38.83 or something similar (results may be slightly different depending on environment) and a qualitative description.



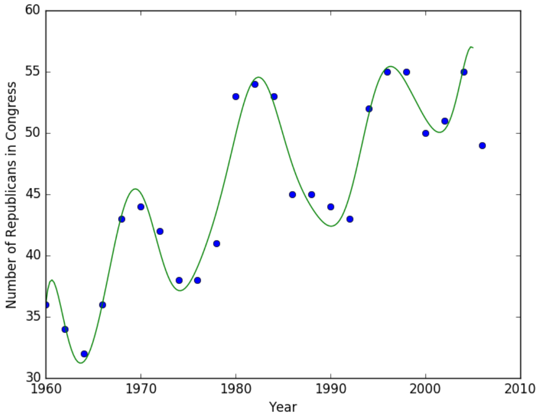
(a) for years



(b) for years



(c) for years

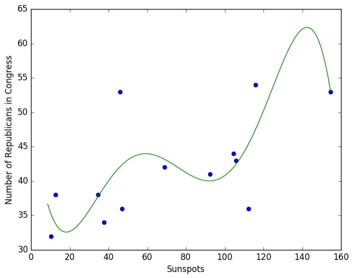


(d) for years

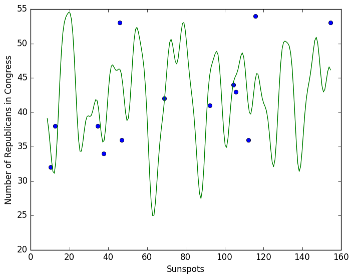
- **Check 2i**: Your graph should match the graph for (a) below.
- **Check 2j**: Your graph should match the graph for (c) below.
- **Check 2k**: Your graph should match the graph for (d) below.

For sunspots vs. number of Republicans in the Senate:

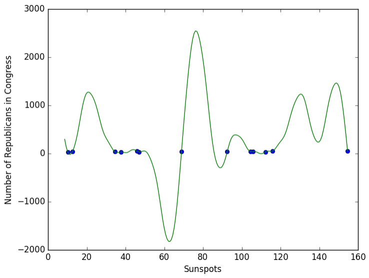
(a) underfits. It has a relatively high loss of 351.23, and we can see graphically that the regression line does a poor job of capturing trends in the data.



(a) for sunspots



(c) for sunspots



(d) for sunspots

(c) underfits/not a good fit. It has a relatively high loss of 375.11, and we can see graphically that the regression line does a poor job of capturing any trends in the data.

(d) overfits. It has an absurdly low loss of virtually zero, and while all the data points fall on the regression line, we can see that the peaks and troughs of the regression line are not found in the data, indicating an overfit.

Given that (a) seems to provide the best fit among the bases for sunspots but still underfits, we should not believe that the number of sunspots controls the number of Republicans in the senate.

- **Check 2l**: Your explanation for (a) should be that it underfits (simply saying it's not a good fit would also be acceptable for this one). Must include a loss of 351.23 and a qualitative description.

- **Check 2m**: Your explanation for (c) should be that it underfits, or is not a good fit. Must include a loss of 375.11 and a qualitative description.

- **Check 2n**: Your explanation for (d) should be that it overfits. Must include a loss of less than 10^{-20} and a qualitative description.

- **Check 2o**: You mention (a) as providing the best fit for sunspots, and given this, you explain that we should not believe that the number of sunspots controls the number of Republicans in the senate.

Problem 3 (BIC, 15pts)

Adapted from *Biophysics : Searching for Principles* by William Bialek.

Consider data $\mathcal{D} = \{(x_i, y_i)\}$ where we know that

$$y_i = f(x_i; \mathbf{a}) + \epsilon_i$$

where $f(x_i; \mathbf{a})$ is a function with parameters $\mathbf{a}$, and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is an additive noise term. We assume that f is a polynomial with coefficients $\mathbf{a}$. Consider the class of all polynomials of degree K. Each of these polynomials can be viewed as a generative model of our data, and we seek to choose a model that both explains our current data and is able to generalize to new data. This problem explores the use of the Bayesian Information Criterion (BIC) for model selection. Define the χ^2 (*chi-squared*) error term for each model of the class as:

$$\chi^2 = \frac{1}{\sigma} \sum_{i=1}^{N} \left(y_i - \sum_{j=0}^{K} a_j x_i^j \right)^2$$

Using this notation, a formulation of the BIC can be written as:

$$-\ln P(x_i, y_i | \text{model class}) \approx \frac{N}{2} \ln(2\pi\sigma^2) - \sum_{i=1}^{N} \ln p(x_i) + \frac{1}{2}\chi^2_{min} + \frac{K+1}{2} \ln N$$

where $\chi^2_{min}(K)$ denote the minimum value of the χ^2 over the set of polynomial models with K parameters. Finally, assume that $x_i \sim Unif(-5, 5)$ and that each $a_j \sim Unif(-1, 1)$. Let $K_{true} = 10$.

(a) Write code that generates N data points in the following way:

1. Generate a polynomial $f(x) = \sum_{j=0}^{K_{true}} a_j x^j$
2. Sample N points x_i
3. Compute $y_i = f(x_i) + \epsilon_i$ where ϵ is sampled from $\mathcal{N}(0, \sigma^2)$ with $\sigma = \frac{\max_i f(x_i) - \min_i f(x_i)}{10}$.

(b) For a set of y_i generated above and a given K, write a function that minimizes χ^2 for a polynomial of degree K by solving for $\mathbf{a}$ using numpy `polyfit`. Check for $N = 20$ that $\chi^2_{min}(K)$ is a decreasing function of K.

(c) For $N = 20$ samples, run 500 trials. This involves generating a new polynomial for each trial, then from that polynomial, 20 sample data points $\{(x_i, y_i)\}$. For each trial, we can calculate the optimal K by minimizing BIC. Compute the mean and variance of the optimal K over 500 trials.

(d) For N ranging from 3 to $3 \cdot 10^4$ on a log scale (you can use the function $3 * np.logspace(0, 4, 40)$ as your values of N), compute the mean and variance of the optimal K over 500 trials for each N. Plot your results, where the x-axis is the number of samples (N) on a log-scale, and the y-axis is the mean value of the optimal K with error bars indicating the variance over 500 trials. Verify that minimizing the BIC controls the complexity of the fit, selecting a nontrivial optimal K. You should observe that the optimal K is smaller than K_{true} for small data sets, and approaches K_{true} as you analyze larger data sets.

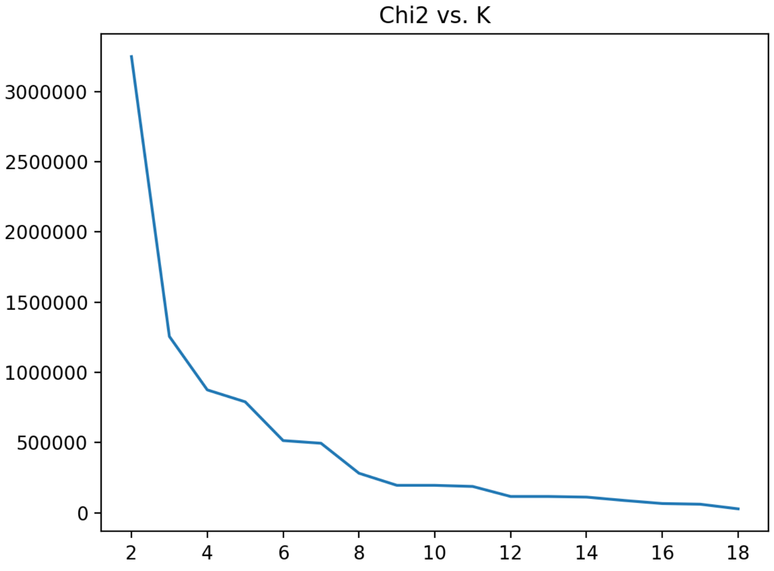
Solution

(a)

- **Check 3a**: You generated x_i and a_μ uniformly from the correct range.
- **Check 3b**: You have a y_i corresponding to each x_i that follows the equations given in the problem.

(b)

- **Check 3c**: You correctly find the minimum χ^2 term using polyfit.
- **Check 3d**: Your plot of the $\chi^2_{min}(K)$ as a function of K looks similar to the one below and is monotonically decreasing.

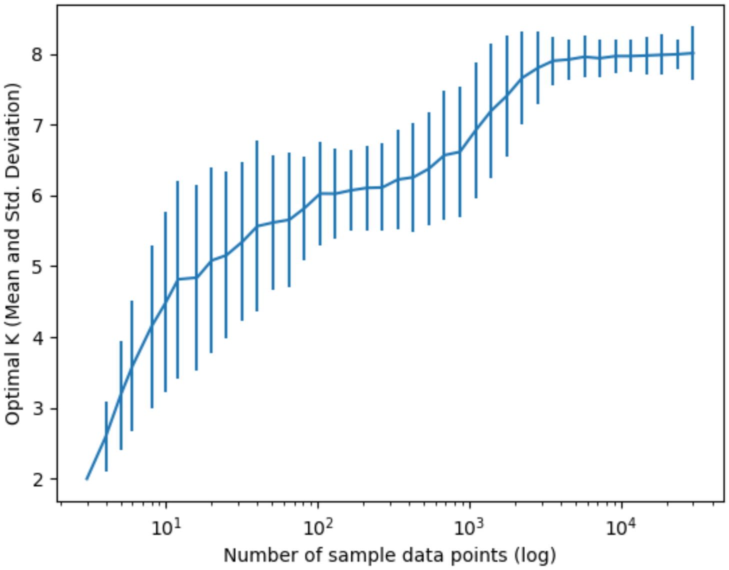


(c)

- **Check 3e**: You calculated the BIC correctly for a given trial and K.
- **Check 3f**: You reported your optimal K, and found that it is $\leq K_{true}$. In the limit of infinite data, optimal K should approach K_{true}, but we will accept any optimal $K \geq 7$.

(d)

- **Check 3g**: Your plot looks similar to the plot below.
- **Check 3h**: You verified that as N increases, the mean of the optimal K according to BIC across 500 trials converged toward K_{true} and the variance decreased.


Optimal K (Mean and Std. Deviation)
2
3
4
5
6
7
8
10^1
10^2
10^3
10^4
Number of sample data points (log)

Problem 4 (Calibration, 1pt)
Approximately how long did this homework take you to complete?

Answer: