MATH 2210 Name: _____

Workshop 4: Best Fit Curves ID: _____

Work in your group to complete the following exercises. You may print this handout, annotate the PDF or write your answer on paper. Make your grader's life easier by writing neatly and legibly!

Please include full explanations and write your answers using complete sentences (not just a bunch of mathematical symbols!). It is important to be able to explain your reasoning to someone else in writing.

Warmup

Question 1. Compute the line of best fit for the following points by using a least squares approximation:

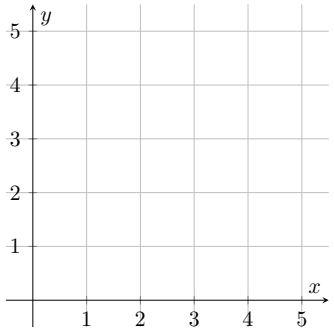
$$(1,0), (1,2), (2,3), (3,1), (5,4)$$

(a) We wish to find a line $y = r_0 + r_1 x$ which lies as close as possible to these points. To do this, we solve the normal equation:

$$\mathbf{A}^\intercal \mathbf{A} \begin{bmatrix} r_0 \\ r_1 \end{bmatrix} = \mathbf{A}^\intercal \mathbf{y}$$

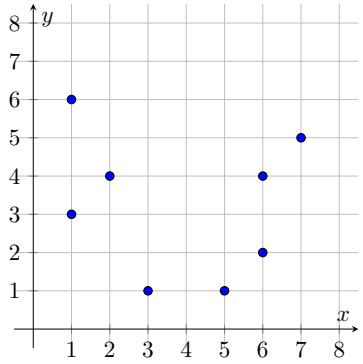
What should we take as our matrix $\mathbf{A}$? (**Hint:** The normal equations give us the best approximation to a solution to $\mathbf{A} \begin{bmatrix} r_0 \\ r_1 \end{bmatrix} = \mathbf{y}$, where each row of this system corresponds to one of our points.)

(b) Solve the normal equations to determine the equation of the best fit line. Graph this line, along with all of the points below.



Fitting Other Curves

In some cases, finding a best fit line provides us with a good understanding of what we are modeling, as there is an underlying linear relationship between the variables. However, this is not always the case. Consider the following data:



In this case, our data does not very close to any line. However, the data is higher for the small and large values of x and smaller for moderate values of x, so may be fit well by a parabola.

Question 2.

(a) We'd like to determine the coefficients of the parabola equation

$$y = r_0 + r_1 x + r_2 x^2$$

which best represents our data (in a least-squares sense). Present this in the form of a matrix equation $\mathbf{Ar} = \mathbf{y}$ for the data points given on the above axes.

(b) Use the normal equations to recast this best-approximation problem as solving a linear system. Solve this system to determine the coefficients $\mathbf{r}$ for the best-fit parabola. Round your coefficients to 3 decimal places.

The process that we used in the previous problem works in general to solve for the best-approximating coefficients of any (independent) linear combination of functions.

Theorem. Given a set of observations $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ and a set of basis function $f_1, \ldots, f_m$, the least squares approximating function

$$y = \sum_{i=1}^{m} r_i f_i(x)$$

has coefficients $r_1, \ldots, r_m$ which are solutions to the normal equation

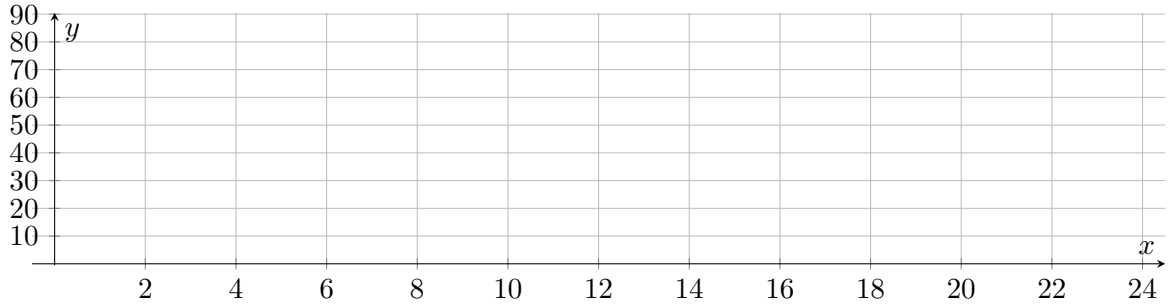
$$\mathbf{A}^\intercal \mathbf{A} \begin{bmatrix} r_1 \\ \vdots \\ r_m \end{bmatrix} = \mathbf{A}^\intercal \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix},$$

where $\mathbf{A}_{ij} = f_j(x_i)$ for each $1 \leq i \leq n$, $1 \leq j \leq m$.

Question 3. Suppose that we wish to find a model for the air temperature in a town. We are given the following average temperature readings for a few months over a two year period:

Month	Temp (°F)	Month	Temp (°F)	Month	Temp (°F)
March 2009	40	May 2009	52	August 2009	84
November 2009	45	January 2010	24	April 2010	34
June 2010	68	September 2010	79	November 2010	48

(a) In order to get an understanding of the shape of the data, plot it on the axes below. Let x represent the number of number of months since the start of 2009 (that is, March 2009 = 3, etc.).



(b) As is expected, the temperature appears to oscillate on a yearly basis, with warmer temperatures in the summer and cooler temperatures in the winter. Because of this, which basis functions would be a natural choice? (**Hint:** You should have a total of 3 basis functions, one of which is a constant function to handle the vertical translation. Also, be sure to account for the period of the oscillation (that is, the length of the repeating cycle) in your basis functions.)

(c) Use the Theorem from the previous page to compute the least-squares approximation for your choice of basis functions. Round your coefficients to 3 decimal places.

(d) Suppose that the data from later years shows a slow upward trend in the average monthly temperatures. Suggest another basis function that we can incorporate into our model to allow it to describe this warming behavior.

Measuring the Quality of an Approximation

By choosing different combinations of basis functions, we now have the ability to find many models that approximate the relationships in our sample data. How can we determine which of these models are good? One qualitative technique is visual inspection: does the function we computed in the model lie reasonably close to the data? A quantitative answer to this question comes from calculating R^2, the Coefficient of Determination.

Definition. The coefficient of determination, R^2 is a value (usually) in $[0, 1]$ which describes how much of the error in our approximation can be attributed to the variance of the sample data. When, $R^2 = 1$, our approximation perfectly fits the data, and when $R^2 = 0$, our chosen basis functions do not describe the data. We can calculate R^2 by the formula,

$$R^2 = 1 - \frac{(\mathbf{y} - \mathbf{Ar})^\intercal(\mathbf{y} - \mathbf{Ar})}{(\mathbf{y} - \overline{\mathbf{y}})^\intercal(\mathbf{y} - \overline{\mathbf{y}})},$$

where $(\mathbf{y} - \overline{\mathbf{y}})$ is the vector obtained by subtracting the sample y-mean, $\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i$, from each entry of $\mathbf{y}$.

Question 4.

(a) Compute the coefficient of determination for the linear model from Question 1(b). What does this suggest about the quality of the approximation?

(b) Compute the coefficient of determination for the quadratic model from Question 2(b). What does this suggest about the quality of the approximation?

Recap

Given a collection of n sample data points, we can calculate an function that approximates the data by following these steps.

1. Plot the data to observe its general shape (slope, asymptotes, periodic behavior, etc.). Use these observations to select a suitable set of basis functions $f_1, \dots f_m$ to model the data.
2. Use the basis functions to construct an $n \times m$ matrix $\mathbf{A}$.
3. Use the normal equations to solve for the coefficients of the approximation.
4. Evaluate the quality of the approximation by computing the coefficient of determination.
5. If necessary, update the basis functions to improve the quality of the model.