MATH 2210

Project 2 - Curve Fitting

Introduction

The goals of this project are for you to use Least-Squares Approximation (recall Workshop 4) to create some models for data that interests you, and to evaluate the quality of the models that you create. Recall the following.

Least-Squares Approximation:

Given a set of observations $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n), \}$, and a set of basis function $f_1, f_m : \mathbb{R}^k \to \mathbb{R}$, 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 $1 \leq i \leq n$, $1 \leq j \leq m$.

Coefficient of Determination:

The coefficient of determination, R^2 is given by the formula,

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

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 $\mathbf{y}$.

R^2 usually falls between 0 and 1, and gives us a quantitative description of how well our model fits the data. The closer that R^2 is to 1, the better our model accounts for the variation in the sample data.

Instructions

You will need to submit a report of your work of at most two pages in length. You can typeset or hand write, but your report must be clear and easy to follow. If you are interested in learning to typeset math, talk to someone on the teaching team. You can also use a mix of text editor for words and handwriting for math. In your report, you should include the following:

- Obtain some sample data from a source that interests you. Some good types of data include population data, geographical data, weather data, nutritional data, etc., but feel free to select something else if you'd prefer. Be sure to provide a citation for your data source.
- Select two quantitative (numerical) variables from your data set that you'd like to study. Identify 1 as your independent (x) variable and the other as your dependent (y) variable. Choose between 8 and 12 data points from which you will construct your models.
- Calculate the best-fit line for your data sample. You need not show all steps of your calculations, but include your matrix $\mathbf{A}$ in your report.
- Graph your data points to see what sort of relationship they suggest. (One way to do this is to use `desmos.com` and take a screenshot.) Choose, with explanation, a second basis (of 3-5 functions) which you think may better represent your data. If you think that a linear model is sufficient, explain why and choose any second basis that you'd like.
- Calculation the best least-squares approximation with respect to your chosen basis. Again, you need not show all steps of your calculations, but include your matrix $\mathbf{A}$ in your report.
- Calculate the coefficients of determination for both of your models. Use these to evaluate the quality of each of the models.

Rubric (Out of 10 Points)

2 Points: The work is clearly presented in at most 2 pages.

1 Points: A data set is chosen (with citation), from which two study variables and 8-12 data points are identified.

1 Points: The shape of the data is used to select appropriate basis functions.

2 Points: Both least squares approximations are calculated with sufficient work shown.

4 Points: The coefficients of determination are calculated and used to analyze the quality of the approximations.