Lab 6: Linear Models

CAS CS 132: *Geometric Algorithms*

Due December 4, 2025 by 8:00PM

In this lab, you'll be building models for the cost of homes in California based on 1990 census data using multiple regression. This is very similar to the example given in the section "Multiple Regression in Practice" from the textbook so please make sure to read the example therein if you want additional guidance. The objective of this lab is not to implement model fitting (as is the case for the example from the text). This will be implemented for you via a single call to `numpy.linalg.lstsq`. Rather, your primary objective is to build design matrices for linear, quadratic, and cubic functions. You can use any function in NumPy we've seen in this course, but it may be useful to look at the following functions before getting started:

- `np.ones`
- `np.column_stack`
- `np.linalg.norm`
- Recall that `a[:,i]` is the `i`th column of `a`.

This lab depends on `sklearn`. **Verify early that this is set up on your machine.** You'll be required to submit a write-up for this lab as a pdf on Gradescope. The details of what you're required to submit are given in the following section.

Lab Write-Up

The items in bold are what must be included in the write-up.

1. Implement the function `distance`, which calculates the distance between two vectors in $\mathbb{R}^n$ (represented as 1D NumPy arrays).

   Include the implementation in your write-up.

2. Implement the function `linear_design_matrix`, which builds the design matrix for simple multiple regression, i.e., for finding a model of the form

   $$f(x_1, \ldots, x_n) = \beta_0 + \sum_{1 \leq i \leq n} \beta_i x_i$$

   This matrix should be identical to the one in the text. The independent variables are given to you as a matrix, where each row represent a datapoint. The dataset has 8 independent variables, so the input `ind_vars` is a 2D NumPy array with 8 columns, and the design matrix you return on the input `ind_vars` should have 9 columns.

   Include the implementation in your write-up.

3. Implement the function `quadratic_design_matrix`, which builds the design matrix for multiple regression with a quadratic modeling function, i.e.,

$$f(x_1, \ldots, x_n) = \beta_0 + \sum_{1 \le i \le n} \beta_i x_i + \sum_{1 \le i \le j \le n} \beta_{ij} x_i x_j$$

This will require adding many new columns; the design matrix you return on the input `ind_vars` should have 45 columns.

Include the implementation in your write-up.

4. Implement the function `cubic_design_matrix`, which builds the design matrix for multiple regression with a cubic modeling function, i.e.,

$$f(x_1, \ldots, x_n) = \beta_0 + \sum_{1 \le i \le n} \beta_i x_i + \sum_{1 \le i \le j \le n} \beta_{ij} x_i x_j + \sum_{1 \le i \le j \le k \le n} \beta_{ijk} x_i x_j x_k$$

The design matrix you return on the input `ind_vars` should have 165 columns.

5. Run the file. You should see a small report on the data, the error, and three graphs. You should take a look at the code that constructs these graphs and describe what they demonstrate.

In your write-up, include the errors of each model, the 3 generated graphs, and a 2-3 sentence description of how to interpret the graphs.