Lab02 – Modelling and Model Error

CS4601/5601 Theory of Machine Learning

Learning Outcomes

- Build confidence and familiarity with model parameters, model input/output, model error, and the differences between all of them.
- Interpret and code models from their mathematical notation.
- Characterize a model's performance when compared to a given dataset.

Overview

At the start of our discussions it is important to appreciate the relationships between our model, the model's parameters, and the data sets we want to predict. In future labs and lectures we will discuss how we represent our models in "feature space" and "solve" our models so that the model output approximates what we see in our datasets.

For this lab you will be investigating two commonly used models and the data that these models attempt to replicate. We want you to build familiarity with the different parts of the model and appreciate shared themes and commonalities between models. When first using models, it can be helpful to identify their constituent parts. To help with this process this lab will have the following format:

1. **Data Familiarity:** Identify and visualize your given data, including the data's independent and dependent variables.
2. **Model Familiarity:** Identify the model that you will use to simulate, or predict, the data that you are given.
3. **Model Implementation:** Code your model such that it accepts inputs of independent variable(s) and tunable parameters so that it produces a prediction or dependent variable as output.
4. **Model Output and Visualization:** Visualize your model's output and plot the model's output against the given data.
5. **Error between Model and Data:** Calculate and summarize the difference, or error, between your model's output and the given data.

The two models you will use are a Gaussian distribution and a multi-dimensional linear model. Remember that models take independent variables as input and produce dependent variables as their output. By adjusting the model parameters, you can change the relationship between the independent variable (feature) input and the dependent variable (response) output.

Instructions

You have been provided with two data sets (gaussdist.csv and advertising.csv). Your job will be to use your jupyter notebook to explore the dataset and implement the models. You will then use these models to make predictions of the dependent variable. You will finish by quantifying the difference, or error, between the model predictions (output) and the dataset that you have.

Jupyter Notebook – Running experiments

Create a Jupyter notebook named <lastname>_lab02. The notebook should have a title, your name, and an introduction that describes what will be covered in the lab and why these topics are important for our class.

Each of the following steps should be performed in individual cells in your jupyter notebook.

Experiment 1: Gaussian Distribution
1 Independent Variable – 1 Dependent Variable – 2 Parameters

Part I – Given Data

1. Load the gaussdist.csv into your notebook.
2. Check the shape of the loaded data; note that there are many rows and 2 columns.
3. Create a scatter plot of this 2-D dataset.
4. Referring to your scatter plot, identify the columns associated with the response variable and the feature variable. Hint: Look at the expected output for a Gaussian – it is not monotonic. The relationship between independent and dependent values here is that each input value yields a unique output value..
5. It is good practice to store features and response variables in separate numpy arrays. Features are often called x and responses are often called y.
6. Now, update your plot code from above to plot the feature (x-axis) versus the response (y-axis) of the gaussdist.csv dataset. Make sure to label your plot.

Part II – Model Familiarity

1. In a new cell, use markdown to write out the mathematical form of the Gaussian Distribution PDF. (Hint: Jupyter markdown cells support MathJax – You can find a LaTeX cheat sheet here. For example, in a Markdown cell, typing $$ \alpha = 2 \sigma $$ will yield α = 2σ.) In this same cell, identify the following:
    a. Model parameters
    b. Independent Variables

c. Dependent Variables

Part III – Model Implementation

1. Implement the Gaussian distribution model. Your code should operate on an independent variable and "fixed" model parameters given as function inputs to produce the dependent variable as output. You may find it helpful to start by writing code that takes a single independent variable (scalar) and produces a single dependent variable (scalar), but ultimately you should support the use of vector input/output. You may want to use the np.power function, among others, and the np.pi constant.

Part IV – Model output and Visualization

1. Use your coded model to generate model predictions (output) using the dataset's independent variable values and model parameters (μ, σ) = (4, 0.5). On the same figure, you will plot the model predictions from the following set of model parameters and the given data (5 sets of data total). Make sure to use proper figure legends for this figure. Hint: You may find it useful to also plot each parameter set separately to help interpret what each parameter is doing. Another hint: There are many ways to put multiple plots on the same axis in Matplotlib; you can do it with a single function call. Here's an example where the same independent variables are used but there are several dependent variable sets, which is the case you have here: plt.plot(x, np.vstack((y_pred, y_pa, y_pb, y_pc, y_pd)).T, marker='+', linestyle='none') vstack treats 1-D array vectors as row vectors and then stacks them making multiple rows, but plot wants data in columns, hence the .T to take the transpose.

	μ	σ
Parameter Set a	1	0.75
Parameter Set b	1	1.25
Parameter Set c	5	1.25
Parameter Set d	6	1.25

Part V – Error between Model and Data:

Each plot that you produced in Part IV depicts the data that you were given and the data that your model produced for a given set of parameters. Our goal is to choose a set of parameters that minimizes the difference between the model output and the given data. The difference between these two sets of values is known as the model error and it is now your job to quantify this error.

One common metric used to quantify error is known as **Mean Absolute Error**. Use this metric to characterize the error between your model and the given data.

1. Use markdown to write out the mathematical form of Mean Absolute Error. In this same cell, identify the following:
    a. Model Predictions.
    b. Dependent Variable(s) from the dataset.
2. For each set of model predictions, calculate the mean absolute error between the model prediction and the given data.
3. Plot all of the model predictions and the given data in a new figure and add the calculated error values to the figure legend. Again, multiple approaches may work, but you may want to use the label parameter (which can be a list of strings when there are multiple series) to the plot function to provide the text for the figure legend.

Experiment 2: Multiple Linear Regression Model
4 Independent Variables – 1 Dependent Variable – 4 Parameters

For this experiment, you should repeat each of the 5 parts as you did for experiment 1. To help we have provided some additional information below.

1. You will use the advertising.csv dataset for the multiple linear regression linear model. The Sales variable will be the response variable that you want to predict (dependent variable). This means that there is a 4-1 relationship between independent and dependent variables.
2. When plotting the data and model output, you now have 4 independent variables and 1 dependent variable. For this lab we want you to only use 2-dimensional (axis) plots to show the relationships between independent variables and independent variables as well as independent variables and the dependent variable. Therefore, you will need multiple plots to visualize the dataset and model output. No - you may not plot multiple features on the same axis. Yes - we know that is a lot of plots. No - do it anyway.
3. When identifying the independent variables of the **multiple linear regression** it may be helpful to label the variables using the variable names from the dataset. While this is not necessary, it may help interpretability and avoid confusion.
4. You are tasked with picking the model parameters. Pick two sets of model parameters and plot the model predictions. Points will not be deducted for a poor choice of model parameters - unless your choices are not in the spirit of the lab – e.g. all 0's.
5. Don't forget to quantify the error between your model predictions and the given data (Sales).

Questions:

After you run all the experiments create a markdown cell at the beginning of your notebook to answer the following questions (it should appear after the introduction). Copy and paste each question into the cell prior to answering it. The following questions will be graded:

1. Describe the relationship between the number of independent variables, dependent variables, and model parameters. If there is no dependence between them, please state this and discuss why this might be or what implications it has.
2. For this lab you were given datasets that had paired independent and dependent variables (supervised data).
    a. What would your model do if you gave it an independent variable value not from this dataset?
    b. Do you think the resulting output would be correct?
    c. How can you be sure?
3. You used mean absolute error to quantify the difference between your given data and model predictions. Lookup (either online or from your textbook) another metric used to quantify error.
    a. Compare and contrast this new metric with mean absolute error.
    b. Discuss what you think the advantages/disadvantages might be between MAE and your other metric. Hint: What shape do different error functions have as you change model parameters to get predictions that are closer to your observed data?
4. We had you plot multiple figures for experiment 2.
    a. Can you think of a way to plot 2 independent variables and the dependent variable on the same plot?
    b. What about a way to visualize 3 independent variables and the dependent variable on the same plot? 4 and 1?
    c. What about a way to visualize a dataset of 100 features and 1 dependent variable on the same plot?
    d. With these plots in mind, describe how the error metric can help.
5. How well did the model parameter sets you chose for experiment 2 perform? With the methods and tools that you have at your disposal, can you think of a more structured way to find a good set of parameters rather than picking a set and checking? Hint: It is relatively inexpensive to evaluate these models. How could you use a loop, or set of loops, to find a set of model parameters that have low error?

Submission Instructions and Grading Criteria

Save the notebook as a PDF named lastname_lab02.pdf. Upload the PDF through Canvas and also upload your notebook file to Canvas if your professor is not having you submit it another way, such as via GitHub.

I will be looking for the following:

- A title, your name, an introduction (including your own summary of the lab), and your answers to the reflection questions at the top of the notebook in Markdown.

- Proper effort was put into each part of each experiment
- That you have the appropriate number of plots and they look reasonable. I will be checking for proper axis labels.
- Obvious effort went into answering the reflection questions.

Presentation	**10%**
Followed submission and formatting instructions	5%
Plots: axes are properly labeled, used correct axes for variables, points were colored as required, lines were coloring as required, used a legend, chose appropriate axes limits to make plot readable and do not cause misleading interpretations, etc.	5%
Model #1: Gaussian Distribution	**30%**
Part I – Given Data **Part II – Model Familiarity** **Part III – Model Implementation** **Part IV – Model output and Visualization** **Part V – Error between Model and Data**	5% 5% 10% 5% 5%
Model #2: Multivariate Linear Model	**30%**
Part I – Given Data **Part II – Model Familiarity** **Part III – Model Implementation** **Part IV – Model output and Visualization** **Part V – Error between Model and Data**	5% 5% 10% 5% 5%
Reflection Questions	**25%**
Problem 1	5%
Problem 2	5%
Problem 3	5%
Problem 4	5%
Problem 5	5%
Exceeds Expectations	**5%**