COMP4702/COMP7703/DATA7703 - Machine Learning
Homework 2 - Parametric Models

Marcus Gallagher

Due: 12:00pm Tuesday 19th March

Core Questions

1. Use the following Data, which has been extracted from Marsland's book on page 8. Note this data is also available in the folder as `hw2q1.csv`.

x	t
0	0
0.5236	1.5
1.0472	-2.5981
1.5708	3
2.0944	-2.5981
2.618	1.5
3.1416	0

   A commonly-used measure of the goodness-of-fit for a regression model is the coefficient of determination (or R^2). Find the (unadjusted) R^2 value for each of the following functions. Please submit your answer correct to 4 decimal places.

   (a) $y = ax^3 + bx^2 + cx + d$

   (b) $y = ax^{10} + bx^9 + ... + jx + k$

   (c) $y = a * sin(5x)$

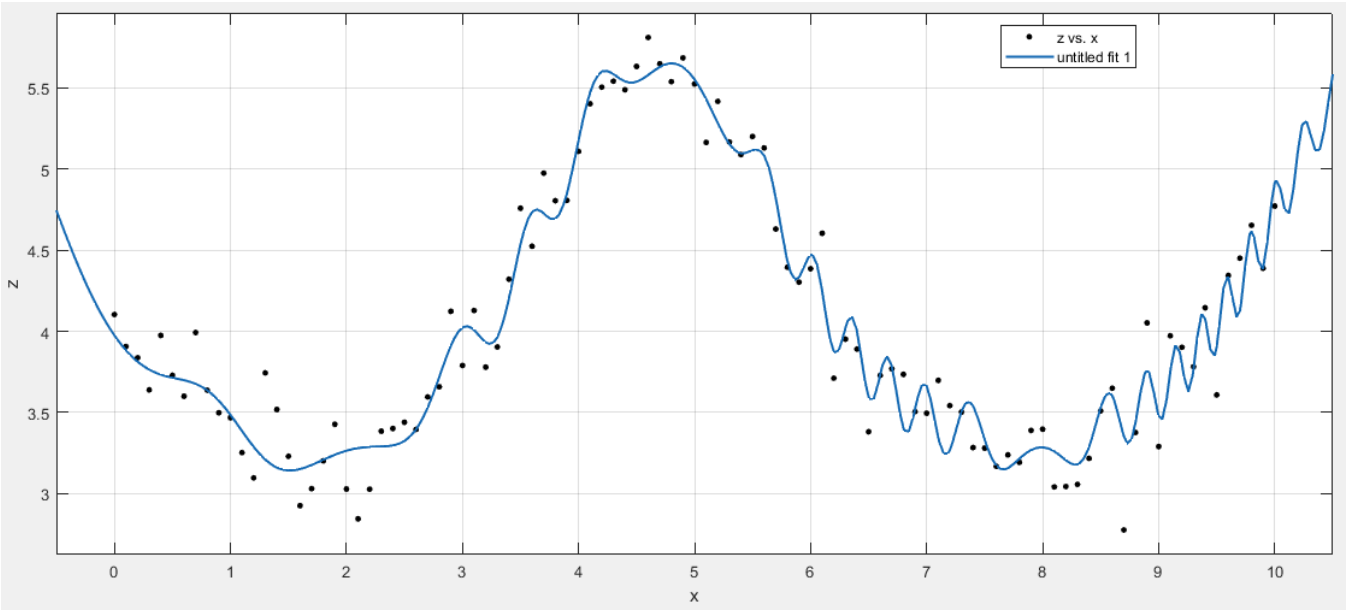
2. Using `hw2q2Training.csv` as the Training dataset and `hw2q2Validation.csv` as the Validation dataset, perform polynomial regression (e.g. using Matlab's `cftool()`) and answer the following:

   (a) Using the validation set for model selection, what polynomial degree order will be selected?

   (b) What is the sum of squared error (SSE) on the validation set, recorded at order 5? (correct to 2 decimal places)

   (c) The data you have used is generated using the same function plus noise as in Prac 2. In one sentence, explain why the best order here differs to the Prac question? (1 sentence)

3. Look at the following graph with an example model fit. Categorize the model as (a) over-fitted; (b) under-fitted; (c) well-fitted?



4. In the lectures and Prac 2 we have considered parametric probabilistic classification for a binary (2-class) problem with one-dimensional input data. This can be extended to the case where we have more classes. Write a function (e.g. in Matlab or python) that takes 3 inputs:

   - A $n \times 2$ set of (training) input data. Where the first column is a one-dimensional set of data, and the second column is the Class, ranging from 0 to k-1.
   - k, the number of classes.
   - x, an test input to classify

   Your function should determine which class has the highest posterior probability for the x-value and return the class number.

   Additionally, using the posterior graphs you produced in the practical can help to validate your answer.

Extension Question

5. Measuring model complexity is a tricky business. Two well-known (and related) measures of model complexity from statistics are the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Find and read a little about the definitions of AIC and BIC. For parametric models, it can be straightforward to calculate AIC and BIC. However for many machine learning models, we may not know:

   - the maximum value of the likelihood function for the model, $\hat{L}$
   - the number of parameters in the model, k

   (a) When does AIC = BIC?

   (b) Produce a 3D plot of AIC for suitable ranges of $\hat{L}$ and k.