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

Marcus Gallagher

Core Questions

1. Use the following data:

x	t
0.0000	2.7600
0.5236	-2.3902
1.0472	1.3800
1.5708	0.0000
2.0944	-1.3800
2.6180	2.3902
3.1416	-2.7600

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 when the sum of squared residuals is minimised. 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) $y = a * \cos(5x)$

Answer: Using Matlab's `cftool` (for example), we get:

(a) 0.4226

(b) There are only 7 data points, so there is no unique solution for a degree 10 polynomial. One option is to fit a degree 6 polynomial, which can be viewed as a degree 10 polynomial with some coefficients equal to zero. In this case, the model interpolates the data exactly and $R^2 = 1$.

(c) The data was generated from a function of this form, so it can also interpolate perfectly (up to some approximation error). $R^2 = 1$. To get this in Matlab:

```
x = [0.0 0.5236 1.0472 1.5708 2.0944 2.6180 3.1416]
t = [2.76 -2.3902 1.38 0.0 -1.38 2.3902 -2.76]
beta0 = 0.5
modelfun = @(b,x)(b*cos(5*x));
mdl = fitnlm(x,t,modelfun,beta0)
```

Note: Matlab will be using an iterative optimisation technique to find the value of a. `beta0` is an initial value for the optimizer. Try a few values for this to check that you always find a good solution!

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

   (a) Using the validation set for model selection, what polynomial degree order will be selected?
   Answer: On the basis of validation set error, an order 11 model 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)
   Answer: 202.50

   (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)
   Answer: The x values cover a larger range.

3. In the book: Pattern Recognition and Machine Learning, by Chris Bishop (available via blackboard, under "Books and Primary References"), an example of underfitting and overfitting is given and discussed (pages 6-9). Study Figure 1.4, Table 1.1 and Figure 1.6.

   Given the model coefficient values shown in Table 1.1, what do you think the coefficient values would look like for the models shown in Figure 1.6? Answer in three sentences or less.

   Answer: The main point is that as more data is added, it has a "regularising" effect, reducing the tendency of the model to fit individual data points. The coefficients for $N = 15$ would be similar to the table, but for $N = 100$ the magnitude of the coefficients should be significantly decreased, perhaps be only over a couple of orders of magnitude rather than 7 or 8.

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 4 inputs:

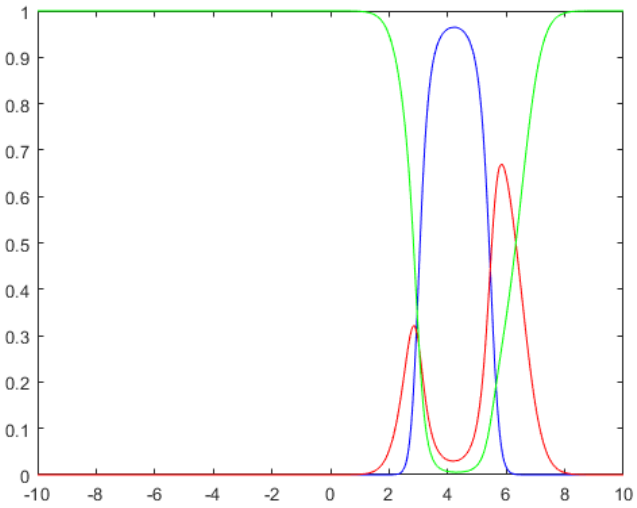
   - A $n \times 2$ set of (training) input data. Where the first column is a one-dimensional set of data, and the second column is the class, ranging from 0 to k-1,
   - k, the number of classes,
   - x, an test input to classify, and
   - $\mathbf{p}$, the k-dimensional class prior probability vector satisfying $\sum_{i=0}^{k-1} p_i = 1$.

   Your function should return the posterior probabilities for each class for the given x-value.

   Using your function, plot a labelled posterior graph similar to that produced in Prac 2, where the horizontal axis represents the value of x and the vertical axis represents the posterior probabilities for each of the k classes. Use the data provided in `iris.csv`, taking the first column as x and the last column as the class label. Assume an equal class prior probability.

   For this question, submit your plot and a listing of your code. You can also include command line output demonstrating usage of the code. Marking is primarily about the output of the code rather than the design.

   See the Matlab code in `hw2q4.m` to generate the solution.



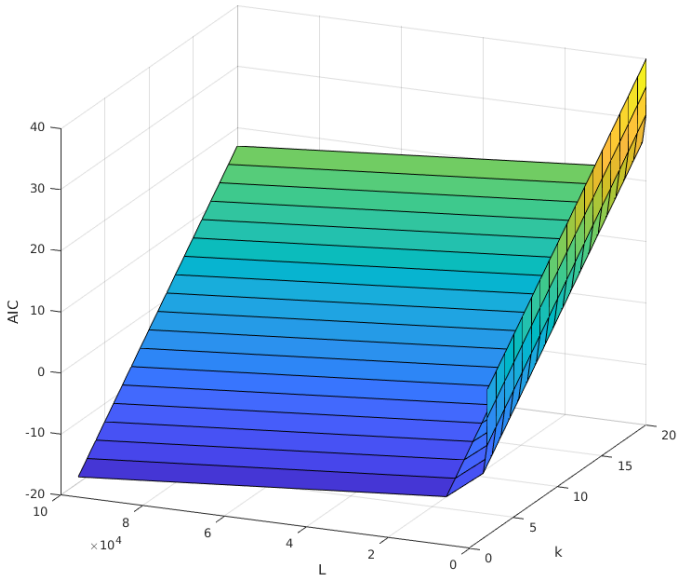
Extension Question

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.

   (a) Suppose you are reading a research paper that lists the AIC and BIC for a model trained on a dataset with 1000 training points. You would like to determine the maximum value of the likelihood. If AIC = 20003.2189 and BIC = 69080.7717, what is the maximum value of the likelihood to four decimal places, and how many parameters does the model have?
   Answer: likelihood = 0.2, $k = 10000$.

   (b) Produce a 3D plot of AIC for suitable ranges of $\hat{L}$ and k, where $\hat{L}$ is the maximum value of the likelihood function for the model and k is the number of parameters in the model.

   Answer: See Matlab code in `hw2q5.m`. The plot will look different for different ranges so we will be pretty flexible with the choice made here (either to catch the curve in the surface or motivated by the numbers in part (a) of the question.


AIC
40
30
20
10
0
-10
-20
10
8
6
4
2
0
×10^4
L
0
5
10
15
20
k