ELL784: Problem Set 1

August 30, 2022

1. Jo has a test for a nasty disease. We denote Jo's state of health by the variable a and the test result by b.

   $a = 1$: Jo has the disease
   $a = 0$: Jo does not have the disease.

   The result of the test is either 'positive' ($b = 1$) or 'negative' ($b = 0$); the test is 95% reliable: in 95% of cases of people who really have the disease, a positive result is returned, and in 95% of cases of people who do not have the disease, a negative result is obtained. The final piece of background information is that 1% of people of Jo's age and background have the disease.

   OK—Jo has the test, and the result is positive. What is the probability that Jo has the disease? [Credit: David MacKay]

2. (a) Consider a standard 6-sided die, where it is known that the probability of getting an even number is the same as that for an odd number. Suppose the die is tossed 100 times, with the following results:

Outcome	**Occurrences**
1	12
2	20
3	16
4	24
5	12
6	16

   Obtain the maximum likelihood estimates for the probability of each of the 6 possible outcomes. (Hint: think about how many independent parameters the model has, and write down the likelihood in terms of them.) How can you make intuitive sense of these estimates?

   (b) Now supposing we discard the strict assumption of equal probability for odd and even numbers; but instead attempt to capture the same information in a prior distribution. Suggest a suitable form for this prior. (Hint: the odd/even distinction is analogous to the head/tail distinction; and your prior need not be a function of all the parameters.) Use your prior and the data above to get maximum a posteriori (MAP) estimates of the probabilities of the 6 outcomes. Comment on the difference between these and the estimates from part (a). What can you do to your prior distribution to drive the two sets of estimates closer to each other?

3. The file provided, P3.mat, contains data of the form $\{(x_1, t_1); (x_2, t_2); ...; (x_{20}, t_{20})\}$, stored in the vector variables x and t. Load this file into MATLAB, and attempt polynomial curve-fitting, varying the degree of the fitted polynomial (look at the documentation for the $polyfit()$ and $polyval()$ functions). Plot the residual sum-of-squares error as a function of the

degree, for a suitable range of choices. Does this enable you to estimate the most appropriate choice of degree? Now adopt a suitable regularisation approach, and repeat the plot with the regularised error. What is your best guess for the curve that generated the given data? Plot this curve on top of the data; what kind of noise do you think the data contains? Try plotting a histogram (MATLAB function *hist*()) of the residuals (differences between the data and the fitted curve) to test your hypothesis.