Phy657 – Module 1

Learning objectives:

1. Fundamental principles of Bayesian statistics:
    a. What is the difference between Bayesian and frequentist approach
    b. What does Bayes theorem state and possible applications
2. General idea of curve fitting:
    a. Explain why curve-fitting can be seen as a minimization problem: describe what is the quantity being minimized in the LogLikelihood fit.
3. Bayesian curve fitting
    a. Understand the interpretation of curve fitting as probability
    b. Understand the difference between frequentist and Bayesian approach to fitting

Reading assignment:

Bishop – Ch.1.1,1.2,1.3

Deadline for work submission:

1. Reports and notebooks must be submitted before Monday , January 26

Topic for discussion:

1. Difference between frequentist and Bayesian approach to fitting
2. Impact of model selection on the quality of the fit
3. Regularization and its impact on the quality of the fit
4. How do we assess the quality of the fit

Activity 1: simple regression problem, frequentist approach

Polynomial Curve Fitting – 1st method [essentially linear least square fitting]
You will generate a set of N data generated from the function $y = \sin(2\pi x)$, smeared with random noise according to a Gaussian distribution with $\sigma = 0.3$. You can start by generating 10 points and then progress to 100, 1000 to see the difference between different data sets. These represent our training sample $\hat{t}$. Our goal is to predict the values of this target variable for some new value of $\hat{x}$. This is equivalent to identify type of the underlying function [in this case $\sin(2\pi x)$ and some shape parameter [in this case $2\pi x$].

a) Introductory activity: plot these data sets [y versus x, x is generated as a uniform distribution between 0 and 1].
b) Try to do a polynomial fit to the target data set with polynomials 1-9 and compare your results with Fig. 1.4 of the text.

For your analysis:

a. Discuss the outcome as you increase the polynomial order and the relationship between the outcome and the number of data points used.
b. For polynomial order n=9 and number of data points N=100 compare the noise σ of your model with the quantity E_{RMS} (Eq. 1.3)
c. For N=100, divide your sample in training (25%) and test (75%) sets. Plot the root-mean-square errors on the training and test sets for various values of M.

Activity 2: Linear regression with Regularization

Now introduce a regularization term in your calculation according to Eq. 1.4 and evaluate the fit parameters for lnλ=-18 and lnλ=0. Compare your results to fig. 1.7

For your analysis:

a) Describe the underlying reason why fig. 1.7 (left) and (right) are so different.
b) Reproduce Fig. 1.8 and explain the reasons underlying the worse outcome when λ is close to 0 or very high.

Activity 3: Practice on Bayes Theorem (no coding necessary)

Consider a rare disease A on which we know:

1. The probability of contracting it is 0.001
2. A test for the disease gives: $P(+|A) = 0.98, P(+|notA) = 0.03$ [false positive]

Do you need to get worried if you get (+) as a test result? In other words, what is the posterior probability?

For your report:
Give your answer and identify clearly:

a) The likelihood function
b) The Bayesian prior
c) How the normalization term is obtained

Activity 4: Bayesian curve fitting

In this activity, you will use the M=9 polynomial as the model to be fitted to the noisy sinusoidal function data set you generated in Activity 1.

Code the procedure described in Section 1.2.6 [formulae 1.68-1.72] to produce Fig. 1.17. Please note that you will use M=9, and the parameters $\alpha = 5 \times 10^{-3}$ and $\beta = 11.1$ are considered input values.

For your analysis:

a) Note that $\beta = 11.1$ should be expected from this data set. Please explain why based on the physical significance of β in the system that you are modeling.
b) Compare this approach with the frequentist fitting approach