# Assigment 1

## EECS 4404/5327

**The assignment is due on Thursday, February 8, before 10:30am (before class).**

1. **Bayesian Reasoning I**
   Consider a test which detects if a person has a disease. Let $R$ denote the outcome of the test on a person, $D$ denote whether the person actually has the disease and $\theta$ be the likelihood that the test gives the correct result. That is, the probability that it reports that someone has the disease ($R = 1$) when they actually do ($D = 1$), is $\theta$, and the probability that it reports that someone doesnt have the disease when they don't is $\theta$. Formally:
   $$p(R = 1|D = 1) = p(R = 0|D = 0) = \theta$$
   Finally, an $\alpha$-fraction of the population actually has this disease, that is, the prior probability of a person having this disease is $p(D) = \alpha$.

   (a) A patient goes to the doctor, has the test performed and it comes back positive. Derive the posterior probability that the person actually has the disease, and simplify it in terms of $\theta$ and $\alpha$.

   (b) After the results of the first test come back positive, the doctor runs it a second time. This time it comes back negative. Derive the posterior probability that the person actually has the disease after this second round of testing and simplify in terms of $\theta$ and $\alpha$.

   (c) Suppose $\theta = .99$ and $\alpha = 0.001$. Suppose 1000 patients get tested, and they are all negative. On average, how many of these patients actually have the disease? I.E., what is the expected number of false negatives?

   **(3 + 4 + 3 marks)**

2. **Bayesian Reasoning II**
   A terrible crime has been committed and blood is found on the crime scene, that must come from the person who committed the crime. Only 1% of the population (in the city which has 1000000 inhabitants) have this type of blood. A suspect is identified and tested positive for this blood type.

(a) The prosecutor says: there was only 1% chance that he had this blood type if they were innocent, so there is now 99% chance they are guilty.
What is wrong with this argument?

(b) The defendant says: There are 10000 people in this city with this blood type, so the chance of being guilty is only 1/10000.
What is wrong with this argument? Can you come up with a scenario, where it would be valid?

(c) Further investigations are being conducted, and more evidence collected. The search is narrowed down to 10 suspects. One of these 10 must have committed the crime. A first suspect of these is chosen (at random), the test conducted and it comes back positive. The judge says: "Given how this whole case developed, I have learned my lesson about using Bayes rule now. We can send this person to jail. We know:

$$p(B) = 1/100, \quad p(G) = 1/10,$$

where $B$ is the event that the blood test comes back positive and $G$ is the event that the person was guilty. We also know $p(B|G) = 1$, due to the evidence on the crime scene. Now we get

$$p(G|B) = \frac{p(B|G)p(G)}{p(B)} = \frac{1 \cdot \frac{1}{10}}{\frac{1}{100}} = 10$$

Now this seems convincing...!"
Is the judge correct?

**(3 + 3 + 4 marks)**

3. **Linear Algebra**
We have seen in class that the solution to regularized least squares regression is given as a solution to the linear system

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{Id})\mathbf{w} = \mathbf{X}^T\mathbf{t}$$

where $\mathbf{X}$ is the design matrix and $\mathbf{Id}$ is the identity matrix. Prove that if $\lambda > 0$, then $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{Id})$ is invertible.

**Hint:** What do you know about the eigenvalues and eigenvectors of $\mathbf{X}^T\mathbf{X}$? What can you infer about $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{Id})$?

**(10 marks)**

4. **Linear Regression**

**Step 1 - load the data**
The data is stored in two files, `dataset1_inputs.txt` and `dataset1_outputs.txt` which contain the input values (i.e., values $\mathbf{x}_i$) and the target values (i.e., values $t_i$) respectively. These files are simple text files which can be loaded with the load function in Matlab/Octave. Plot the outputs as a function of the inputs (ie plot the datapoints, not a curve) and include this plot in your report.

**Step 2 - ERM**
For degrees $W = 1, \ldots 20$, fit a polynomial of degree $W$ to the data using (unregularized) least squares regression. For each learned function, compute the empirical square loss on the data and plot it as a function of $W$. Include this plot in your report. Which value of $W$ do you think would be suitable?

**Step 3 - RLM**
Repeat the previous step using regularized least squares polynomial regression with $\lambda = 0.001$. Again plot (and include) the empirical loss as a function of $W$. Compare and discuss the two curves you get for ERM and RLM.

**Step 4 - cross validation**
Implement 10-fold cross validation for RLM. That is, randomly divide that data into 10 chunks of equal size. Then train a model on 9 chunks and test on the 10th that was not used for training. For each model you train, average the 10 test scores you got and plot these again as a function of $W$. Which value of $W$ do you think would be suitable?

**Step 5 - visualization**
For the degree that you chose based on the previous questions and for $W = 1, 5, 10, 20$ plot the data along with the ERM and RLM learned models. Discuss the plots.

**Step 6 - bonus**
Repeat the steps above (or whatever else you may find suitable) to come up with a polynomial regression vector $\mathbf{w} = (w_0, w_1, \ldots w_W)$ for the data in `dataset2_inputs.txt` and `dataset2_outputs.txt` (to be posted a few days before the submission deadline). Submit the weights vector. Your submitted weight vector will then be tested on an independent test set generated by the same process.

Please submit the weights as a 21-dimensional vector $\mathbf{w} = (w_0, w_1, \ldots w_{20})$ to be applied to the data as $w_0 + w_1 x + w_2 x^2 + \ldots + w_{20} x^{20}$; if you choose $W < 20$ just set the appropriate weights to 0. Submit this vector as a text file with each weight on a line.

**(2 + 5 + 5 + 5 + 3 marks + 5 bonus marks)**