Assigment 1

EECS 4404/5327

Winter '19

The assignment is due on Friday, February 8, before noon.
Submit a paper copy of your assignment to the dropbox in Lassonde!

1. **Bayesian Reasoning I**
   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)

2. **Bayesian Reasoning II**
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 θ 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 θ, and the probability that it reports that someone doesnt have the disease when they don't is also θ. Formally:

$$p(R = 1|D = 1) = p(R = 0|D = 0) = \theta$$

Finally, an α-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 a general formula for the posterior probability that the person actually has the disease, and simplify it in terms of θ and α. Which value do you get for $\alpha = 0.001$ and $\theta = .95$?

(b) After the results of the first test come back positive, the doctor runs it a second time. Again, it comes back positive. Derive the posterior probability that the person actually has the disease after this second round of testing assuming the two test results are independent and simplify in terms of θ and α. Again, in addition to the general expression, report the values you for $\alpha = 0.001$ and $\theta = .95$.

(c) Analyze under which conditions the posterior probability of having the disease after two positive tests is larger than after only one positive test. How does it depend on α and θ?

(d) Now we would like to use the above insights for employment of intelligent systems: Say security at an airport would like to use a machine learning based system to identify travelers smuggling illegal substances based on expressions of their face while going through security. Say the system was trained to 95% accuracy and we can expect 0.1% of travelers to be smuggling illegal substances. Would installing multiple cameras have the same effect as multiple blood tests? Explain you reasoning!

(3 + 3 + 3 + 3 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. In this question you will prove that if $\lambda > 0$, then $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{Id})$ is invertible.

(a) Show that every eigenvector of the matrix $\mathbf{X}^T\mathbf{X}$ is also an eigenvector of the matrix $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{Id})$.

(b) Show that all eigenvalues of $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{Id})$ are strictly positive.

(c) Use the above results to conclude that $(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{Id})$ is invertible.

(3 + 3 + 2 marks)

4. **Linear Regression**

In this question you will implement linear least squares regression as discussed in class. Print out your code and submit it with you assignment.

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 write-up.

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. Each time train polynomial of degree 20 for regularization parameters λ so that $\ln(\lambda) = -1, -2, \cdots - 20$. This time plot (and include) the empirical loss as a function of i. Compare and discuss the two curves you get for ERM and RLM.

Step 4 - cross validation
Implement 10-fold cross validation for ERM. 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 degrees $W = 1, 5, 10, 20$ plot the data along with the ERM learned models. Do the same for models learned with RLM with a fixed regularization parameter $\lambda = 0.001$ (while varying the degree as for ERM). Discuss the plots. Which degree seems most suitable? What is the effect of adding the regularizer here?

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)