BE 700 Fall 2021 Homework : Review and Networks

This homework is LONG but aims to help you assemble the knowledge we covered and test your basic understanding. It is not hard but will hopefully encourage you to review again the material. The second part is a quick programming exercise in networks.

Part I: Review and Understanding (short answers please, long answers will be skipped).

Questions 1-9 are easy if you understood the lectures and know the answer. 1-5 minutes each. Question 10 requires research and may take several hours of work.

1. Most machine learning algorithms solve an optimization problem during training. What is it mathematically?
2. We use 1-NN and k-NN to learn on a random set of points in D dimensions with random labels. We train and test on the same set. K is odd. What accuracy do you expect from this experiment and why?
3. We use a decision tree algorithm on the same set.
   1. What size decision tree do you expect to get?
   2. What will be the information gain expected to be achieved at the root node?
   3. Explain information gain and state succinctly (or via an example) when it might be better than simple accuracy as an impurity measure for decision trees.
   4. What optimization problem is solved by the decision tree algorithm?
4. We have a linearly separable two class data in D dimensions. We now add one irrelevant dimension. We populate the irrelevant column in the data with random data. The labels do not change.
   1. What optimization problem is solved by the perceptron learning algorithm?
   2. What accuracy do you expect a perceptron will achieve on the modified dataset?
   3. Can you modify your perceptron algorithm to achieve very high accuracy with high probability on the modified data set.
   4. What if we add D1 < < D irrelevant dimensions with completely random data. Can you repeat (b) for this problem.
   5. What if we now add a new dimension (attribute) Xi+ where the data is just 2Xi where Xi is one of the old dimensions. How about many such copies of old columns?
   6. What if we add a new column Xij+ where the data is just Xi+Xj.
5. Can you propose a two layer neural network (perceptron that will perfectly classify a non-linearly separable set) in 2D (draw the set and define the neural net).
6. Gene regulation is dependent on transcription factors.
   1. What happens to cells if we knock out all transcription factors out.
   2. What effects will we see on gene expression if we knock one transcription factor.
   3. What effects will we see if we knock one transcription factor out and suppress translation.
   4. What are kinases and why are they relevant to cancer. What happens when we use a drug to inhibit a kinase?
7. Propose a new way to cure cancer based on cancer hallmarks we discussed.
8. Cancer cells are associated with genome instability.
   1. State several specific markers of this instability we might observe in cancer cells and how they might benefit cancer cells.
   2. Can you come up with a creative way to exploit these consequences to fight cancer?
9. Cancer cells are associated with poor differentiation.
   1. Is poor differentiation a good or bad prognostic marker and why?
   2. How can we build an ML based diagnostic for poor differentiation? Are any available on the market.
10. Insulin Resistance is a poor marker of health with a number of long term serious disease consequences. What are they? E.g. Alzheimer’s. How can we explain some of these severe consequences? List at least three. This problem requires a lot of research and requires a one page write up.

Part II. (Requires two hours of work – perhaps more, perhaps less).

1. Perform a correlation analysis (both spearman and pearson) on the data in your two homework (cancer and diabetes). What do you conclude from the differences in the correlation structure for the two diseases.
2. Produce networks by thresholding the correlation with different thresholds.
3. Plot the degree distribution for different thresholds. What do you conclude from this analysis.
4. What does correlation on gene expression in different types of cancer data sets (discuss at least three) tells us about systems biology of cancer