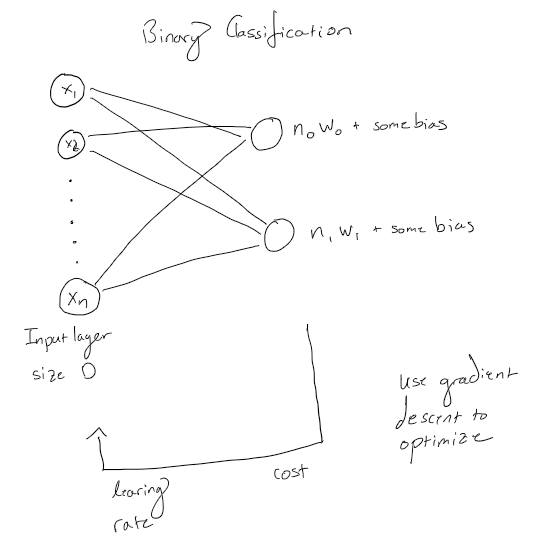
Vibhav Jha BE700 HW 4

Part I: Review/Short Answers

1. We are minimizing the loss function.
   1. For example **min(C(ŷ (i), yi)) = - (yilog(ŷ(i)) + (1-yi)log(1-** **ŷ (i)))** for a neuron with a sigmoidal activation function.
2. There should be rather high accuracy so long as D dimensions is not very large. In general, the 1-NN will have lower accuracies compared with k-NN the larger D is.
3. *Decision Tree*
   1. The decision tree should be at most a size of D if there are no repeat decision rules used.
   2. The information gain will be the highest of all the other attributes.
   3. Information gain is essentially a measure of the entropy/change of entropy of a feature. It may be better to use simple accuracy when the data is well known or balanced.
   4. In general decision trees look to do a local search for either the minimum entropy or maximize the information gain for each node.
4. Linearly separable two class and one irrelevant dimension
   1. The perceptron aims to minimize iteration error.
   2. Expect lower accuracy than when compared to just the linearly separable two class data as perceptrons struggle with non-linearly separable data.
   3. Instead of using one perceptron, I would set a 1 v rest neural network, where each input is connected to each output, this may should be able to bypass the non-linearity.
   4. I expect accuracies to be around 0.5 as there may no longer be any clear boundaries.
   5. The accuracy will be close to (b), as it really is just scaled dimensional data which will provide similar decision boundaries.
   6. By adding Xij+, there will be a significant drop in overall accuracy, maybe around 0.5 as the data is significantly more confusing.
5. Two Layer NN
   1. Using a 1 v rest approach and minimizing cost as well as utilizing gradient descent, it is possible to classify a non-linearly separable set. Depending out what the classification out is, there can be N output neurons, dependent on number of classes. The activation function can vary, though the most popular for these kinds of problems is ReLU.



1. Gene regulation
   1. If all transcription factors are knocked, cells will die shortly. No new proteins/RNA will be produced.
   2. It may be a cascading effect where one transcription factor could have a large role in decreasing overall gene expression, or there could just be a slight decrease in overall gene expression.
   3. The specific gene will then be silenced, there will be no expression at all.
   4. Kinases are phosphorylation catalysts, often used in signaling pathways. In cancer deregulated kinases can lead to the progression of cancer. Inhibiting a kinase could halt the proliferation of cancer (basically stopping phosphorylation), however the impacts on surrounding networks that utilize the same or similar kinases could be unknown.
2. One of the cancer hallmarks is to evade apoptosis, if there was a method to tag what is different in cancer cells such as the damaged chromosomes and generate an artificial signal to the body such that it can recognize the defective cells again.
3. Cancer Cells and Genomic instability
   1. Some markers are the weakening of the DNA damage checkpoint or the p53 pathway, where if this is suppressed then cell apoptosis does not function correctly, allowing cancer cells to freely proliferate. Another marker is depletion of Shugoshin function (Sgo1) where chromosomal separation and mitosis is terminated prematurely, allowing damaged chromosomes to pass through unchecked. This further proliferates cancer cells.
   2. Yao Y, Dai W. Genomic Instability and Cancer. J Carcinog Mutagen. 2014;5:1000165. doi:10.4172/2157-2518.1000165
4. Cancer Cells and Differentiation
   1. Poor differentiation is a bad prognostic marker as the cells will be more abnormal, thus allowing for much more rapid spread.
   2. Earlier this year, Zhou et al 2021 published an article in nature scientific reports showing a novel method to predict lymph node metastasis of poorly differentiated cancer. If this is efficient, it could be possible to then train a simple algorithm based on this poorly differentiated data to classify good and poor differentiation.
   3. Zhou, CM., Wang, Y., Ye, HT. et al. Machine learning predicts lymph node metastasis of poorly differentiated-type intramucosal gastric cancer. Sci Rep 11, 1300 (2021). https://doi.org/10.1038/s41598-020-80582-w
5. Insulin Resistance and Disease

Insulin Resistance is associated with a few long-term diseases. The most obvious is type-2 diabetes however after developing type-2 diabetes there are many other diseases that can co-occur. Higher risk of stroke, heart disease (hypertension, atherosclerosis, cardiomyopathy, coronary artery disease etc.), fatty liver, kidney dysfunction and Alzheimer’s. The most common cause of insulin resistance in obesity which lead to high free fatty acids and decreased lipolysis1. Obesity by itself is not just a cause of insulin resistance, it also contributes to dysregulation of adipocytes, or in other words, how the body stores excess nutrients. It is not just insulin resistance that leads to long term diseases, but also other comorbidities that could be the cause of or caused by insulin resistance.

*Overall Heart Disease and Insulin Resistance*

In the heart, most energy is supplied by fatty acid oxidation, with a supplement being provided by glucose oxidation. In the presence of insulin resistance there is a greater reliance on fatty acid oxidation1. This reliance on fatty acid also help to cause greater insulin resistance. It seems as though energy imbalance is the biggest issue that arises from insulin resistance thus leading to various heart diseases. For example, AMP-activated protein kinase activity is reduced with insulin resistance, and could possibly lead to mitochondrial dysfunction in the heart1. Alongside energy imbalance it is observed that systemic inflammation and insulin resistance is linked due to an increased circulation concentration of inflammatory cytokines1. So, with increased fat, insulin resistance there can be observed myocardial inflammation contributing to decreased heart health.

It has been shown that in severe chronic heart failure, patients are more insulin resistant2. While this relation has not been fully explained, it is shown that insulin resistance contributes to oxidative stress which in turn impairs calcium uptake (lowers activation rate of ryanodine receptor thus decreasing calcium uptake; causes diastolic/contractile dysfunction), decreases metabolic flexibility (a decrease of glucose oxidation thus decreasing cardiac efficiency), decreases ATP production (thus increasing apoptosis) as well as cardiomyocyte-endothelial nitric oxide signaling impairment (impaired generation of NO causes hypoxia and thus myocardial cell death)3.

*Kidney Dysfunction and Insulin Resistance*

With insulin resistance specific dysregulation of polyol flux, an increase in glycation end products, imbalance of protein kinase B, an increase in mitogen activated protein kinase all lead to glomerulus damage4. The increase of glycation end products is triggered through hyperglycemia. From hyperglycemia kidney damage comes from increased concentrations of transforming growth factor β in which there is increased production of the extracellular matrix4.

*Alzheimer’s and Insulin Resistance*

In Alzheimer’s there are dual factors leading to brain insulin resistance. One is peripheral insulin resistance linked by excess cytotoxic ceramide production. Similar mechanisms as described above in the heart disease section also apply here, but specifically looking towards the decreased cerebral glucose utilization associated with Alzheimer’s, with increased insulin resistance there is a significant reduction in cerebral blood flow5. This can be considered as brain diabetes5. Two main mechanisms are seen throughout: a loss of insulin/IGF responsive neurons and reduced membrane receptor expression of the insulin/IGF ligand-receptor.

Citations

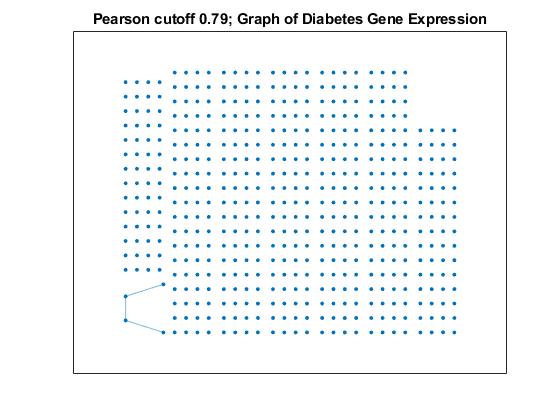
1. Abel ED, O'Shea KM, Ramasamy R. Insulin resistance: metabolic mechanisms and consequences in the heart. Arterioscler Thromb Vasc Biol. 2012;32(9):2068-2076. doi:10.1161/ATVBAHA.111.241984
2. J. W. SWAN, C. WALTON, I. F. GODSLAND, A. L. CLARK, A. J. S. COATS, M. F. OLIVER, Insulin resistance in chronic heart failure, European Heart Journal, Volume 15, Issue 11, November 1994, Pages 1528–1532, <https://doi.org/10.1093/oxfordjournals.eurheartj.a060425>
3. Aroor AR, Mandavia CH, Sowers JR. Insulin resistance and heart failure: molecular mechanisms. Heart Fail Clin. 2012;8(4):609-617. doi:10.1016/j.hfc.2012.06.005
4. S. De Cosmo, C. Menzaghi, S. Prudente, V. Trischitta, Role of insulin resistance in kidney dysfunction: insights into the mechanism and epidemiological evidence, Nephrology Dialysis Transplantation, Volume 28, Issue 1, January 2013, Pages 29–36, <https://doi.org/10.1093/ndt/gfs290>
5. de la Monte SM. Insulin resistance and Alzheimer's disease. BMB Rep. 2009;42(8):475-481. doi:10.5483/bmbrep.2009.42.8.475

Part II: Networks

For simplicity correlation thresholds of 0.25, 0.5 and 0.79 were chosen for all. In the diabetes set only 500 genes were chosen to alleviate computational load from this implementation. Code is attached as a .zip [BE700HW4\_vjha.m for cancer set; B700HW4\_diabetes\_vjha.m, as well as new\_expressiondata.mat, and all raw figures]

From the Pearson and spearman correlation, it was clear to see in the diabetes data set there were very low correlations overall, with only a few out of 500 having significant spearman’s ρ (a significant value is considered to be over 0.8). What is more interesting is that when decreasing the threshold to 0.5 there was a small cluster than were very well connected. There were many lowly correlated values showing an extremely well-connected graph for both the Pearson and spearman correlation methods. In the cancer set the same general trends follow that very few if any are significantly correlated at 0.79 or 0.5, but large clusters begin to appear at 0.25. A key difference is that using spearman cutoff of 0.25, two separate large clusters were formed, as compared to a fully connected graph from the diabetes set. Taking a finer look at the graphs and the degree distributions for each, the larger diabetes set was more evenly distributed in terms of degrees. The cancer set was not as well connected and often there were gaps of connections. From these results, it appears that diabetes has a larger system wide as compared to localized cancer clusters. [Figures on following pages].

The higher the correlation between genes in cancer sets indicate areas of similarity and thus should be identified in studies as markers of cancer. Looking at those specific groups can allow for a deeper understanding of the overall low, middle and high-level networks associated with gene regulation and cancer.



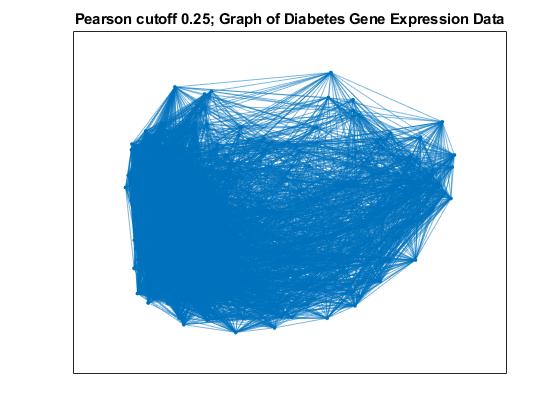
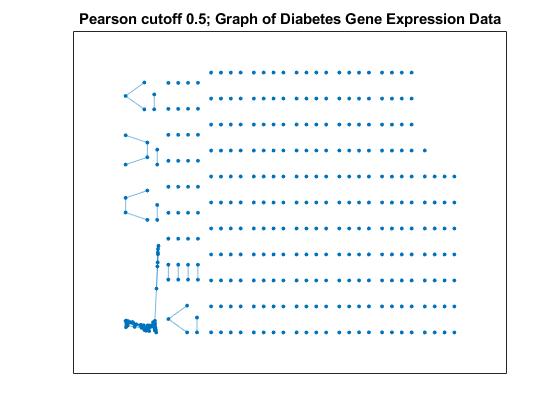


Figure 1. Diabetes Gene Expression Graphs based on Pearson correlation thresholds of 0.25 (top left), 0.79 (top right) and 0.5 (bottom left). Only 500 genes were used. Very few genes were correlated (0.79 graph) and the number marginally increased to have one large cluster and a few other connected areas. At 0.25 threshold the graph becomes nearly fully connected, indicating that this visualization is suboptimal.

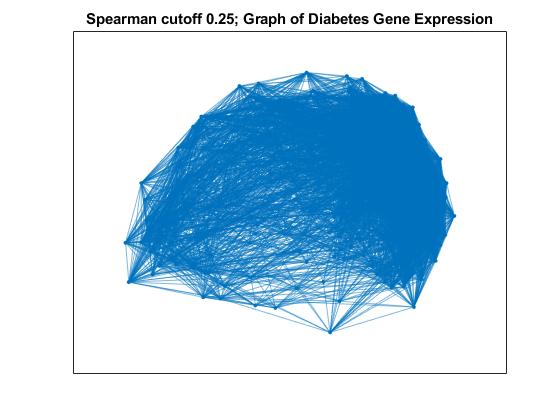
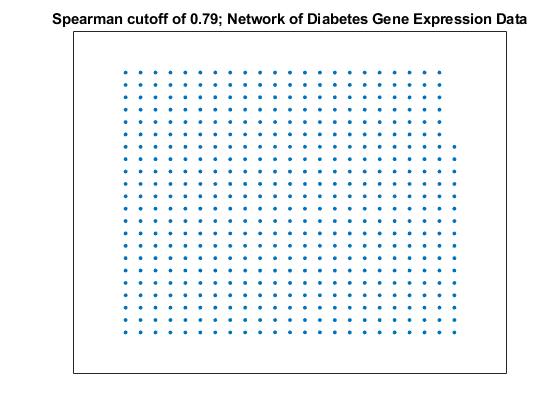
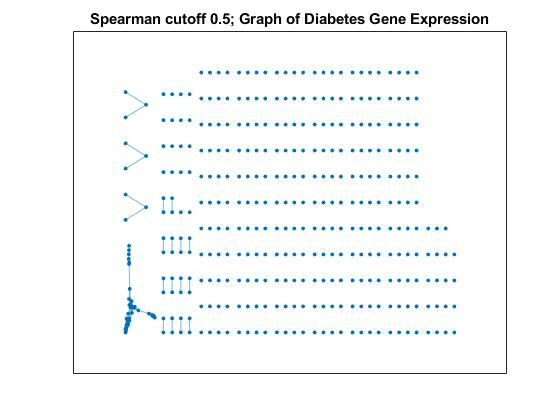


Figure 2. Diabetes Gene Expression Graphs based on Spearman correlation thresholds of 0.25 (top left), 0.79 (top right) and 0.5 (bottom left). Only 500 genes were used. No genes were correlated (0.79 graph) and the number greatly increased to have one large cluster and a few other connected areas. At 0.25 threshold the graph becomes nearly fully connected, indicating that this visualization is suboptimal.

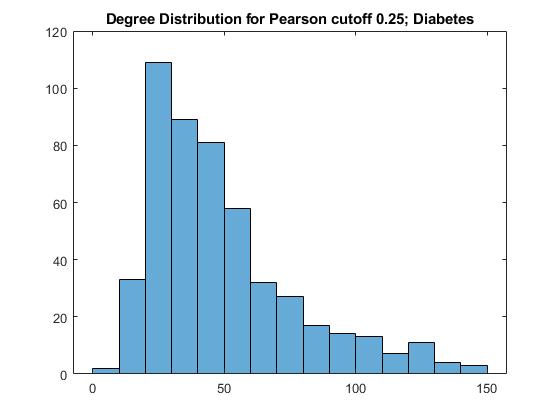
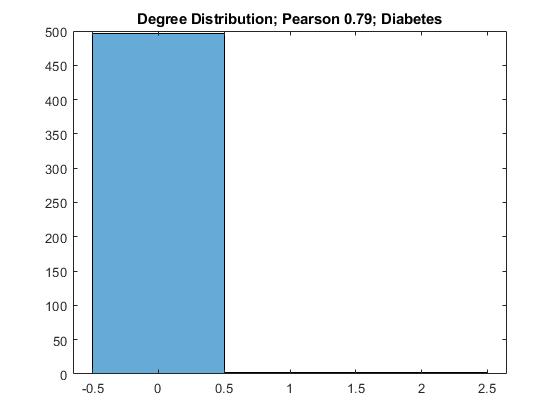
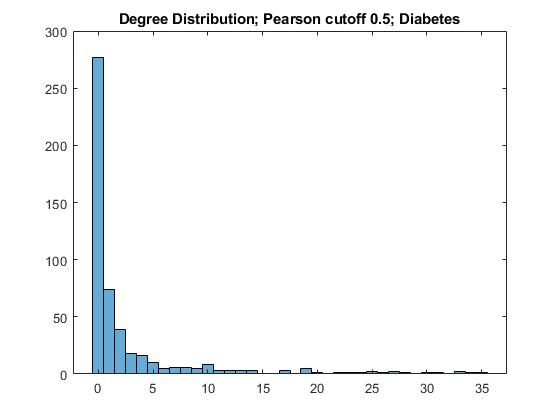


Figure 3. Degree distributions for the graphs using Pearson cutoff on diabetes data set. For 0.79 there only a few connections, with 0.5 there are significantly more connections, with certain genes connected at high as 35 times. At 0.25 the graph shows a significant amount of connections, with a majority connections around 20 to 70.

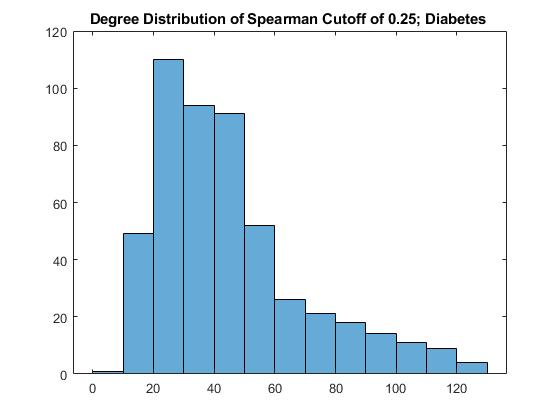
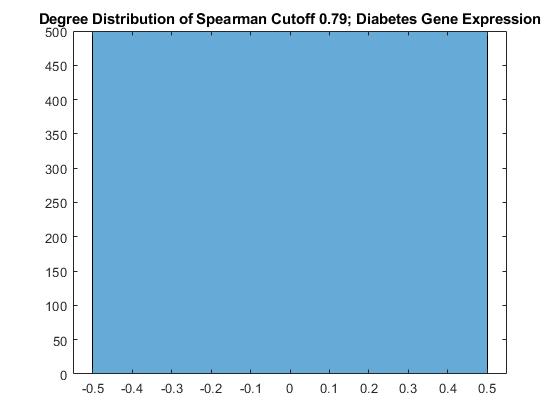
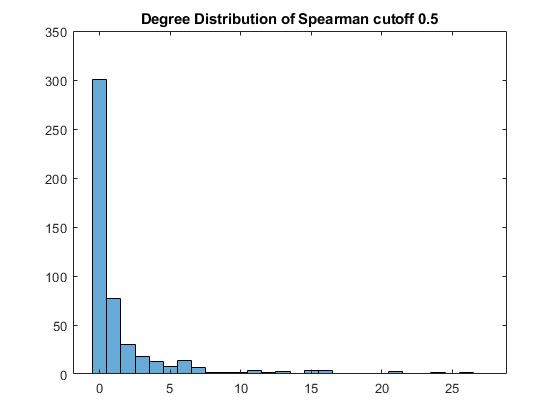


Figure 4. Degree distributions for the graphs using Spearman correlation cutoffs on diabetes data set. For 0.79 there were no connections, with 0.5 there are significantly more connections, with certain genes connected at high as 25 times. For 0.25 the connections were better spread out among the genes, with the highest reaching about 130 connections.

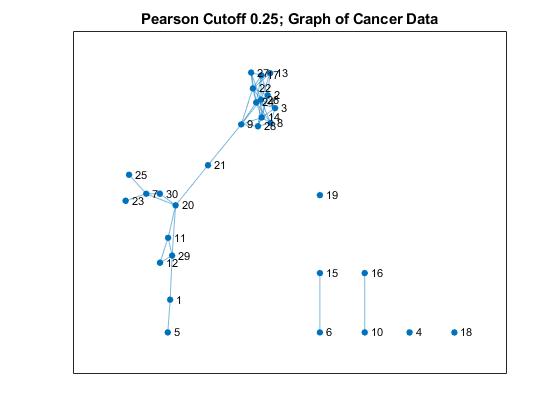
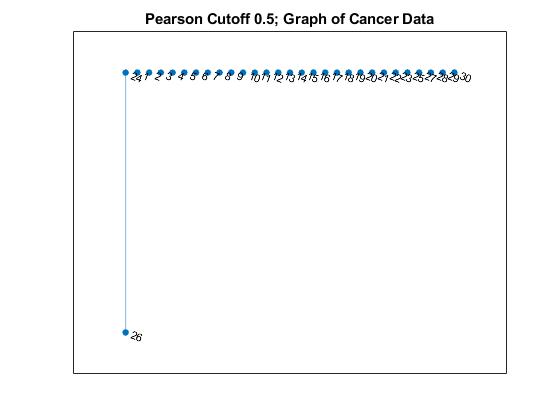
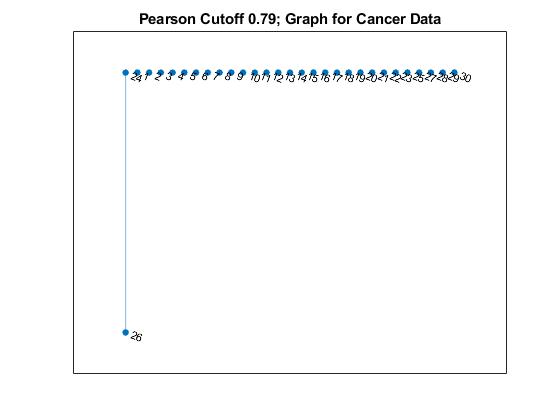


Figure 5. Cancer Gene Expression Graphs based on Pearson correlation thresholds of 0.25 (top left), 0.79 (top right) and 0.5 (bottom left). Only one gene was correlated in both 0.79 and 0.5 cutoff thresholds. At 0.25 threshold the graph is well connected, however 7 genes are outside the main network.

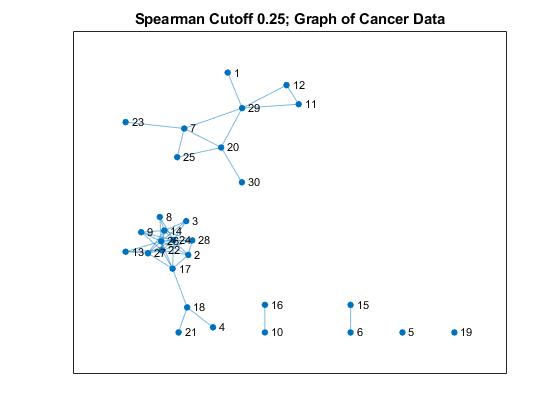
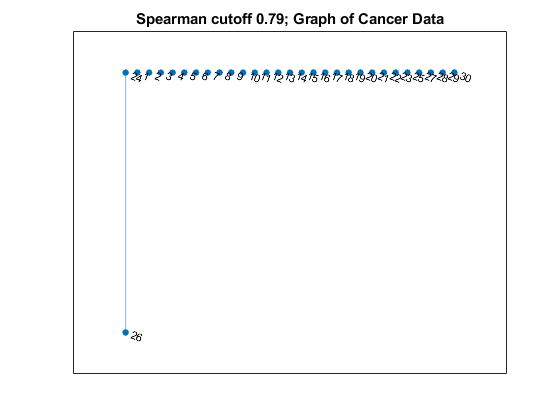
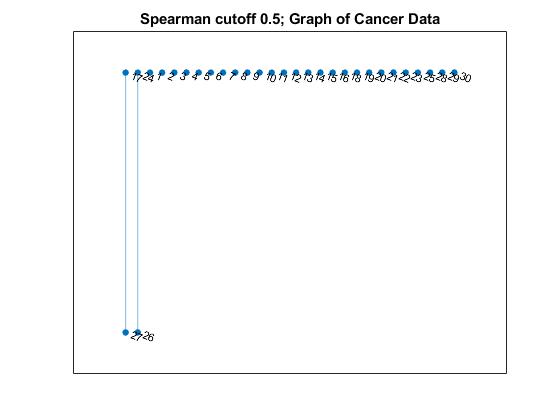


Figure 6. Cancer Gene Expression Graphs based on Spearman correlation thresholds of 0.25 (top left), 0.79 (top right) and 0.5 (bottom left). Only one gene was correlated in both 0.79 cutoff threshold. Only one additional correlation was identified at the lower threshold of 0.5. At 0.25 threshold the graph shows two separate clusters of connections.

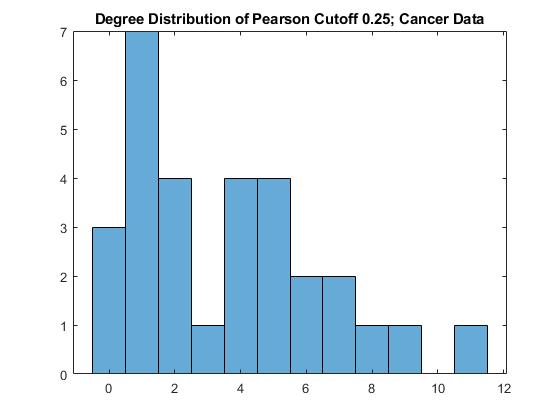
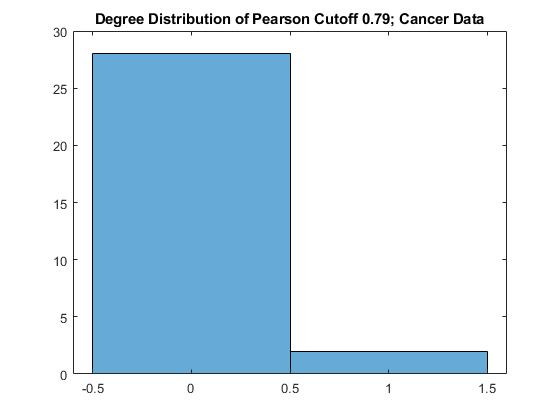
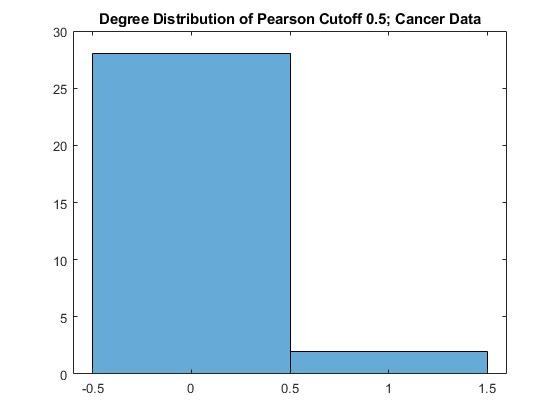


Figure 7. Degree distributions for the graphs using Pearson cutoff on the cancer data set. For 0.79 and 0.5 there one or two connections. At 0.25 it is not an even distribution, with most connections being 1,2,4 or 5.

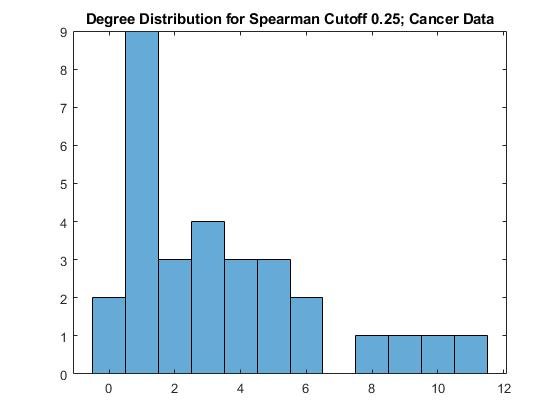
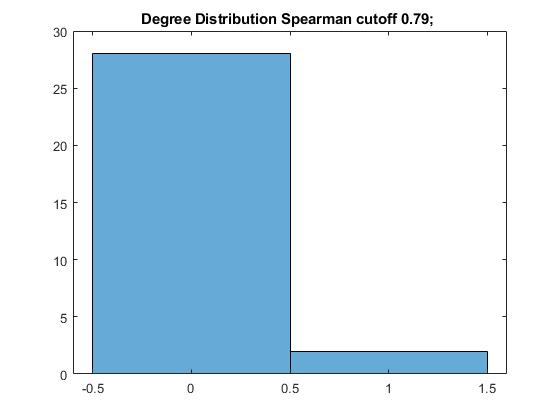
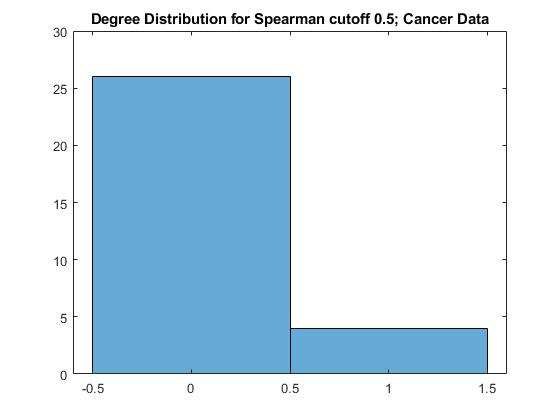


Figure 8. Degree distributions for the graphs using Spearman cutoff on the cancer data set. For 0.79 and 0.5 there are one or a few connections. At 0.25 it is not an even distribution, with most connections being 1 or 3.