BIOL 419/519

Homework 6, Winter 2016

Due on Wednesday, March 2 by midnight

This homework is designed as an exercise not only to practice what we've learned in class, but synthesize/combine different ideas as well as push you to learn more related ideas on your own. It's less structured and more exploratory than the earlier homework assignments, and it's hopefully closer to what you'll do for your projects.

You are expected to do research, understand new algorithms, and figure out how to make use of code by reading documentation. These are key skills in any data science endeavor with real data! Wikipedia articles are often good resources, as are tutorials, posted lecture notes and blog posts. You are not expected to do it alone – I strongly encourage you to work in teams and discuss between yourselves.

Instructions: Please submit a write-up of your homework by uploading on Canvas. Your solutions must include the code your wrote to solve the problem as well as the output/answer. You should generate a write-up within Matlab using the "Publish" option to make a PDF file. Please also upload all .m files. **Failure to turn in all .m files will result in half credit for the entire assignment.**

In all exercises that ask you to plot or visualize something, label your plots for full credit.

Please seek help if you need it! You may ask questions at Friday's lab, come to office hours, and get together with your classmates to troubleshoot together.

Collaboration: As noted in the Syllabus, what you turn in should reflect your own understanding of the material. Collaboration with your classmates is encouraged, and I ask you to clearly indicate these collaborations in your homework.

1. (6 pts) **Classifiers 3 ways: Beyond linear discriminants**

   In this problem we're going to explore classification of some cancer microarray data. Download the `CancerMicroarray.mat` dataset from Canvas. This data was originally from here: `http://home.ccr.cancer.gov/oncology/oncogenomics/`.

   The data contains gene profiles of tumor samples obtained using cDNA microarrays. The tumors belong one of four (4) categories, which include neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL) and the Ewing family of tumors (EWS).

   Fun fact: There is a paper that uses a neural network to classify this same dataset by Khan *et al.*, which appeared in Nature Medicine in 2001. You'll find a copy of this paper on Canvas if you're curious.

   (a) (1 pt) How many tumor samples are there? How many genes are profiled for each tumor?

   (b) (1 pt) Compute the PCA of this data. Plot the cumulative sum of the variance explained as a function of number of principal components (PCs). How many PCs do you need to explain 95% of this data?

   (c) (2 pts) Using the scores of the first 10 PCs, train a linear discriminant classifier (LDA) and a quadratic discriminant classifier (QDA) on this data. To be specific, the input

should be the scores of the first 10 PCs on all tumor samples, and the output should be the type of tumor. What is the cross-validated accuracy of each classifier?

Hint: Use the `fitcdisc` and `predict` functions. Read the documentation for examples, and find optimal parameters to train different types of discriminant classifiers supported by these functions.

(d) (2 pts) Using the scores of the first 10 PCs, train a decision tree classifier.

Do some research to teach yourself how this kind of classifier works, then summarize in 1–2 sentences. What is the cross-validated accuracy of the decision-tree classifier?

Hint: Use the `fitctree` function, read the documentation to figure out how to use it and visualize the results.

2. (4 pts) **Discovering clusters with k-means**

(a) (1 pt) Load Fisher's iris data, which contains morphological measurements of a few related species of iris flowers. This dataset was first collected by Ronald Fisher in 1936, and remains in use today. The data comes with Matlab, so to load it, use the command `load('fisheriris');`

How many flowers are in the dataset? How many morphological measurements did Fisher take, and what are they? How many types of irises are in the dataset?

(b) (2 pts) Cluster the iris measurements using the `kmeans` function. Try $K = 2$, 3, and 4 clusters.

Plot the results of your clustering in three separate figures, one for each K. In each figure, show each flower as a dot and color the flowers according to which cluster they were assigned to. (How to do the plot with >2 measurements per flower and only 2 axes? For visualization purposes, PCA the data and show the scores on the first 2 PCs. Note that we're going to do this only for the plot, not for kmeans.)

Hint: Use the `scatter` function to make colored dot plots easier to manage.

(c) (1 pt) If we did not know how many species of irises were in the original dataset, how would we pick the best K among the ones we tried? Do some research and answer in 1–2 sentences.

3. (Extra Credit) **Density-based clustering**

In lecture, we came up with two general strategies for clustering: computing distances and computing densities. The kmeans algorithm uses the first approach of computing distances. There is another very popular clustering algorithms based on computing densities known as *dbscan*, or Density Based Spatial Clustering of Applications with Noise. This problem explores the `dbscan` algorithm.

(a) Do some research and read about dbscan. Summarize what it is and how it works in 2–3 sentences.

(b) Download a Matlab implementation of dbscan:

`http://yarpiz.com/255/ypml110-dbscan-clustering`.

(c) Run the code you downloaded on the included example data. Run kmeans clustering on the same example data and compare the results. What are the key differences in the results of the two algorithms?

(d) What are two reasons one may want to use dbscan over kmeans? What are two reasons one may want to use kmeans over dbscan?

4. (Informational) How many hours did you spend on this homework?