BIOE 210, Spring 2022

Homework 11

Due Monday, 4/11/2022 by 5:00pm.
Upload your answers to Gradescope. If submitting a single PDF,
you must mark the location of all answers.

Part I

1. Given the line $y = -3x + 4$:
   (a) Write a vector normal to this line.
   (b) What is the distance from the origin to the closest point on the line?
   (c) What is the closest point?
2. Given the hyperplanes

$$\begin{aligned} x_1 - 2x_2 &= 4 \\ -3x_1 + ax_2 &= b \end{aligned}$$

   (a) Find values for a and b such that the hyperplanes have a unique point of intersection. Plot the hyperplanes.
   (b) Find values for a and b such that the hyperplanes have infinite points of intersection. Plot the hyperplanes.
   (c) Find values for a and b such that the hyperplanes do not intersect. Plot the hyperplanes.

Part II: Machine Problem

A team of researchers used DNA microarrays to measure gene expression in a large set of breast cancer cell lines (Kao, et. al, *PLOS One* 4(7): e6146. doi:10.1371/journal.pone.0006146). In this exercise, you will use gene expression profiles from this study to build a classifier that differentiates between invasive and regular ductal carcinoma (IDC and DC).

1. Load the mat file `HW5_data.mat`, which contains the following variables:
   - `training_lines` is a Matlab table containing gene expression data for the IDC and DC cell lines. Each of the 8750 rows corresponds to a gene with variable expression across the cell lines. Each of the 28 columns represents a cell line. The following cell lines were classified as invasive (IDC) by a pathologist: BT474, BT483, BT549, EFM19, MDA134, MDA175, SUM102, T47D, UACC812, UACC893, ZR75_1, and ZR75_30. The remaining cell lines are noninvasive ductal carcinoma (DC).
   - `patient_samples` is a Matlab table containing gene expression values for the same 8750 genes from the training data. Each column corresponds to a different patient biopsy.
2. Build an SVM classifier that separates IDC from DC samples.
   - The Matlab command `fitcsvm` accepts numerical arrays, not tables, so convert your table with the function `table2array`.

- Pay attention to the dimensions of your inputs, especially what rows and columns correspond to in your data and for `fitcsvm`.

3. Perform both *k*-fold (with 4 folds) and leave-one-out cross validations using the command `crossval`. Using the function `kfoldLoss`, report the accuracy of your model using each validation method.

4. Repeat the cross validation five times for both the *k*-fold and leave-one-out methods. Does the accuracy change for either method? Why or why not?

5. Using the Matlab `predict` function, determine if each biopsy in the patient data set is invasive (IDC) or regular (DC) ductal carcinoma.

Remember to submit all code, outputs, and explanations for these problems.