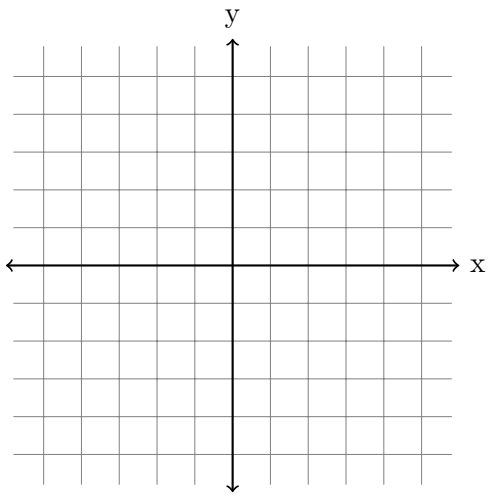
IST 597: Homework 5

Justin Silverman

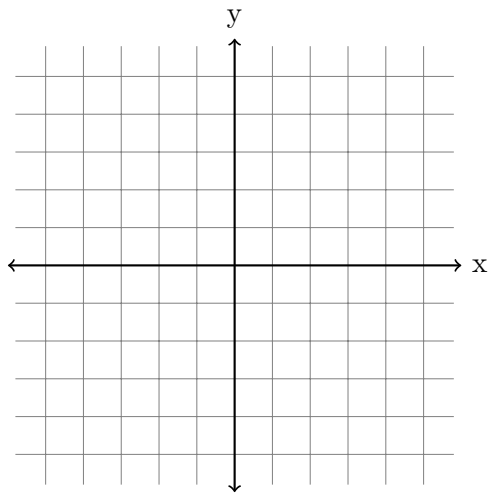
Problem 1 (20 points)

Consider trying to learn the relationship between y (target) and x (predictor).

a. Using "$\circ$" to denote samples from a training set and "$\times$" to denote a single test point, draw a dataset (and test point) where linear regression fails but a Gaussian process with an RBF kernel would do very well.



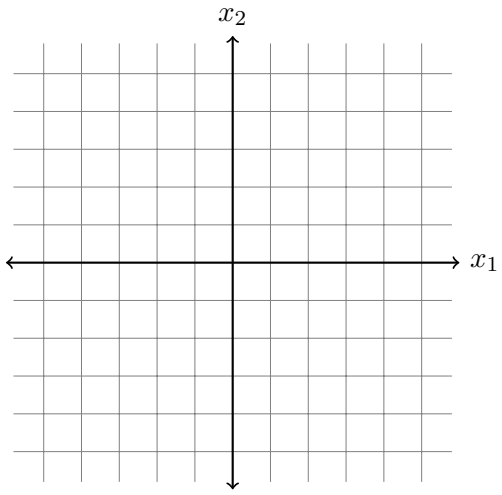
b. Now draw the opposite, a situation where linear regression does very well but a Gaussian process with an RBF kernel fails. BUT, I want you to draw this for a situation where the underlying process is actually non-linear. In other words, draw a setup where linear regression (without any transformations of x, just simple linear regression) does a better job at predicting the non-linear relationship a Gaussian process with an RBF kernel. Non-trivial solutions only, e.g., a setup where its clear that the linear-regression model "just got lucky at that one test point" will not relieve full credit. Really think about when a simple linear regression model may do better than a GP with RBF kernel for modeling a non-linear relationship.



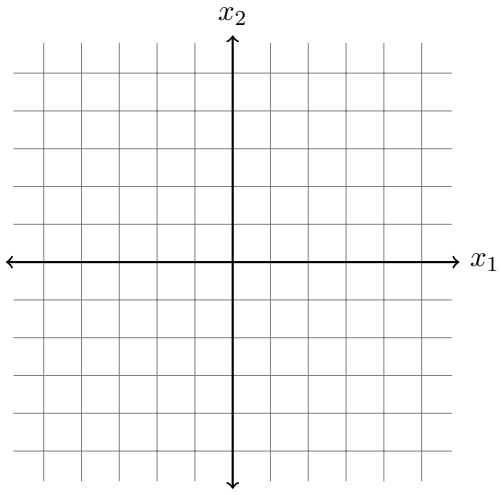
Problem 2 (20 points)

Consider trying to learn to classify two classes "$\circ$" and "$\times$" using two predictors, x_1 and x_2.

a. Using "$\circ$" to denote samples from one class and "$\times$" to denote samples from the other class, draw a dataset where Random Forests, Neural Networks, and Gaussian Process Classification do well, but logistic regression does poorly.



b. Now draw a case where Logistic Regression does well but standard implementations of random forests or decision trees do poorly.



Problem 3 (50 points)

You are going to be building a model to predict wine quality (high/medium/low) and type (red/white) using the other variables given in the dataset. The dataset is given in `data_train_hw5_problem3.csv` and I am asking you to predict the labels `quality` and `type` for the training set `data_test_hw5_problem3.csv`.

a. (0 point) Do a brief exploratory data analysis. Notice anything unusual? How many observations are in each of the quality levels? (Please avoid telling me unimportant details like the number of observations or which column has more 7's – just highlights that directly impact modeling).
b. (10 points) This is an multivariate and ordinal multi-class classification task (`low`, `medium`, `high` for `quality` and `red`, `white` for `type`). How can you accomplish this task? How are you going to encode the quality variable? (For ideas you may want to look at this manuscript (but there are other ways as well): https://www.cs.waikato.ac.nz/~eibe/pubs/ordinal_tech_report.pdf. A summary of this approach is given in this blog post: https://towardsdatascience.com/simple-trick-to-train-an-ordinal-regression-with-any-classifier-6911183d2a3c.). How are you going to handle the two dimensional prediction task (e.g., predicting two different variables)?
c. (5 points) How are you going to deal with the inequality of the class labels (the fact that there are many more `medium` examples than `high` or `low`).
d. (25 points) Go ahead and implement your proposed scheme. As before, submit your predictions as a csv file with two columns (the first `quality`, and the second `type`) with the file name `[your last name]_predictions_hw5_problem3.csv` (do not include row numbers). Also, make sure that you use the same labels as in the training data: submissions with `1`,`2`,`3` rather than `low`, `medium`,`high` will receive no credit as they are ambiguous.
e. (10 points) Try to use the learned model for inference. What features were the most predictive? What is the most important factor in deciding what makes a good or bad wine? Explain how you came to that answer.

Problem 4 (10 points)

Describe, intuitively, the relationship between overfitting in machine learning and the multiple hypothesis testing problem in machine learning.

Please keep your answers short.