Machine Learning on pH Strips

BioE 134 Final Project, Fall 2017

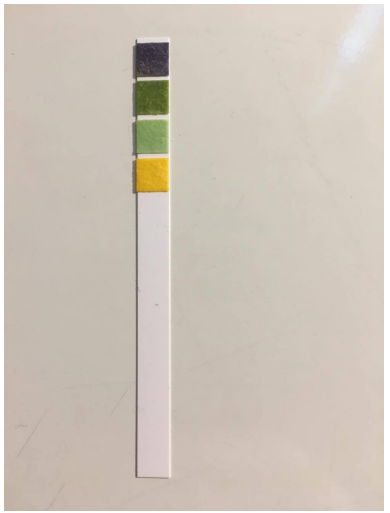
Matthew Sit and Rudra Mehta

Data Collection

Matthew Sit

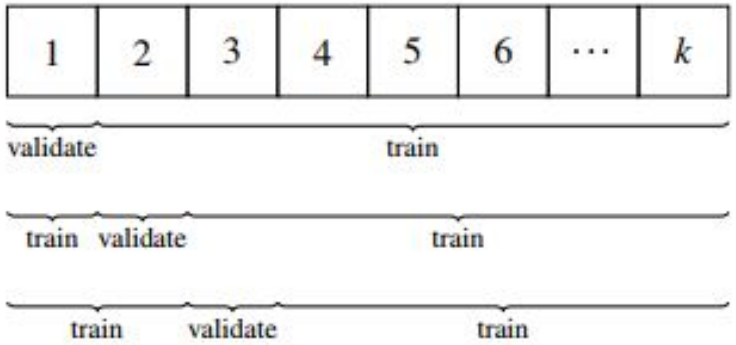
Data Collection

9 different buffers,
10 trials of each.



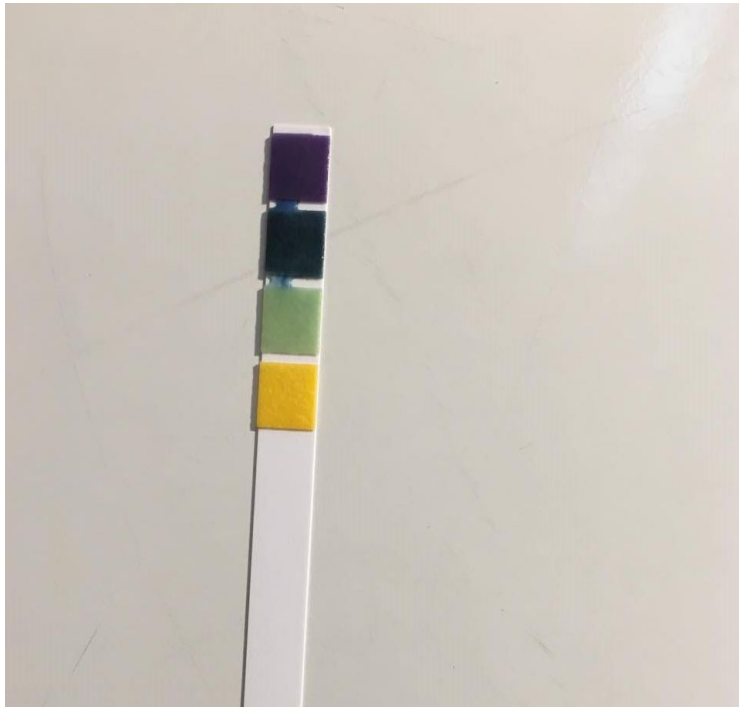
Problem: 90 is not enough data points

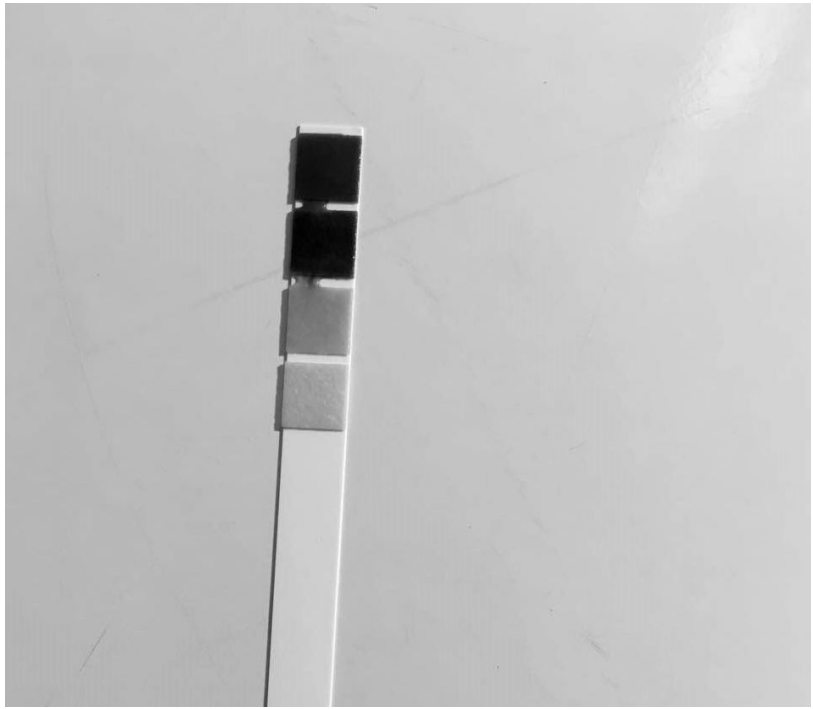
Solution: Implemented k-fold cross validation to receive full mileage from collected data.



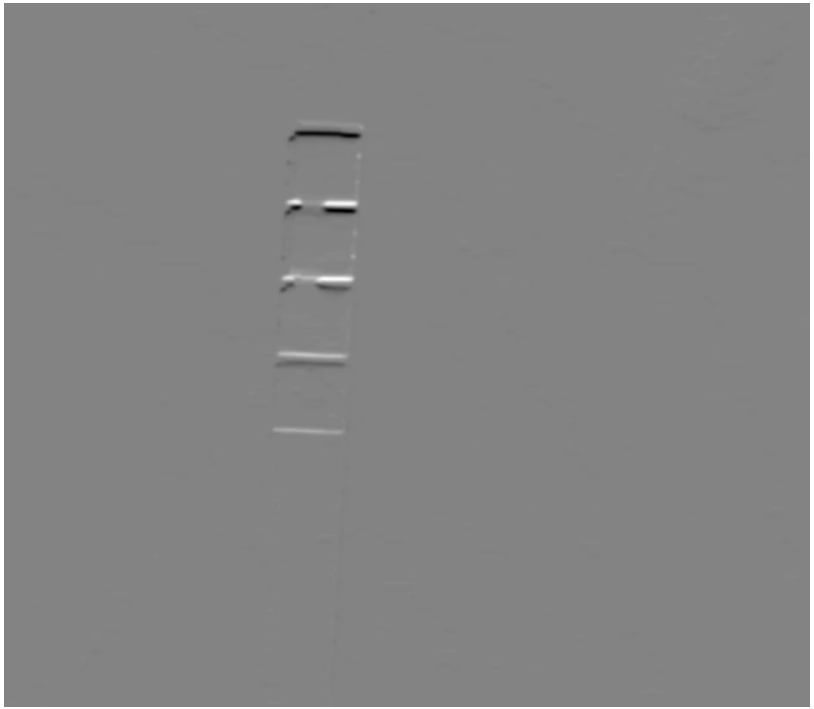
Manual Feature Extraction

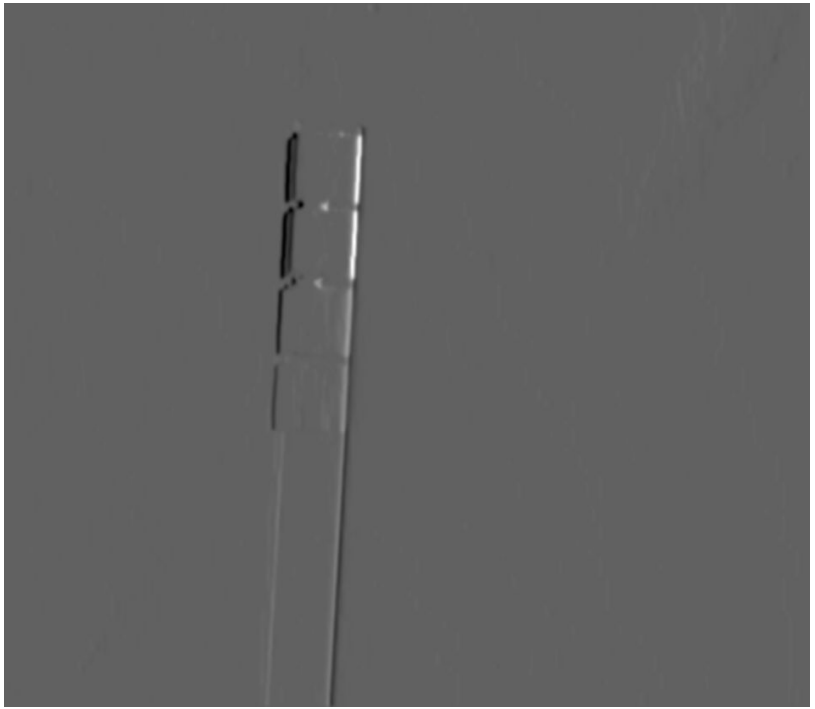
Rudra Mehta



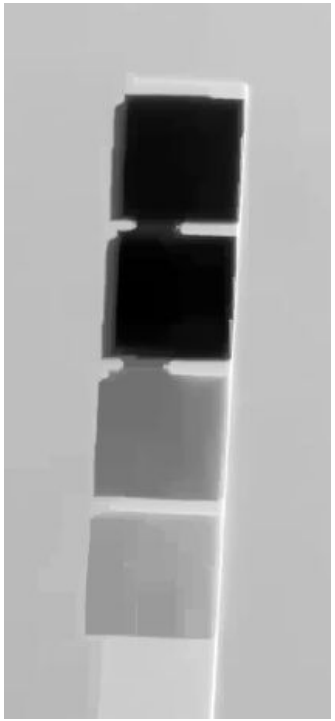


Original, grayscale images

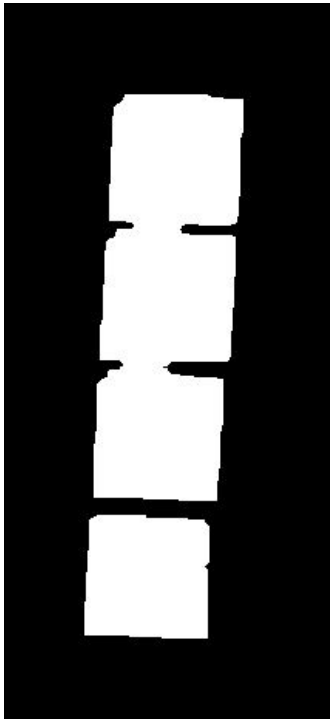


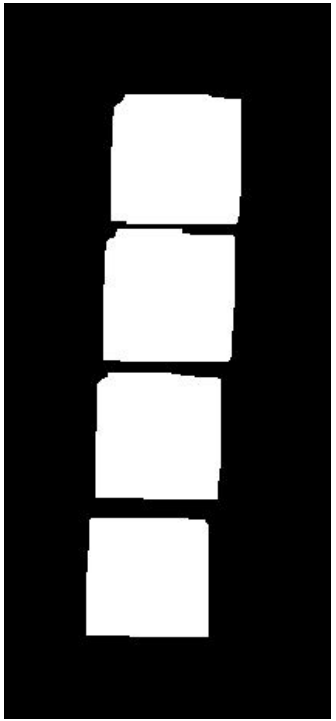


Horizontal and vertical edge maps - gradient convolutional filter

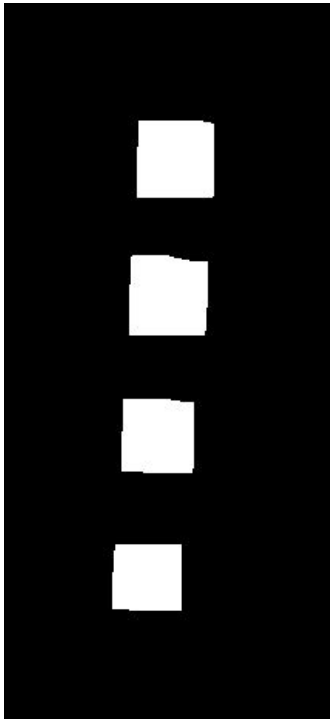


Cropped images using edge maps

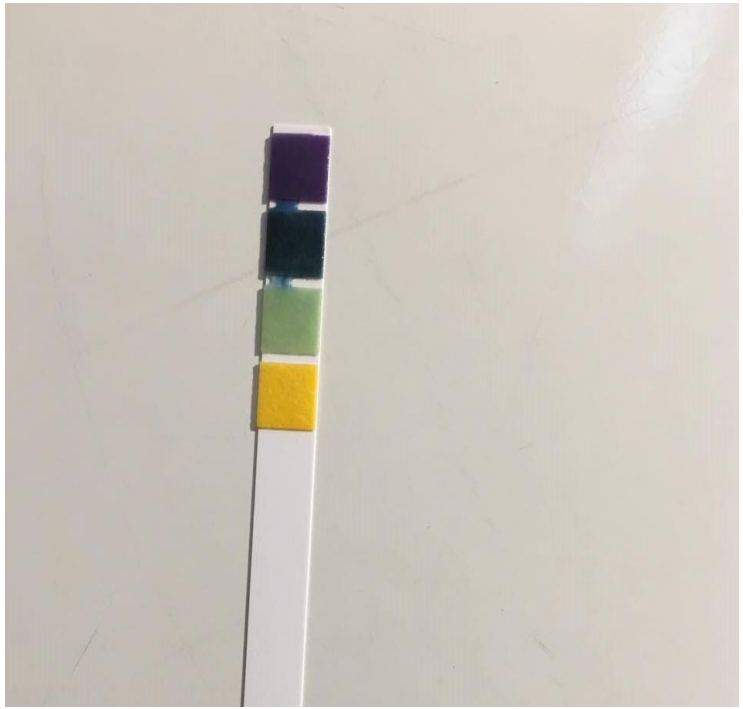




Binarize image, open to disconnect different squares



Get the center areas of the squares, and apply this final mask to the original image



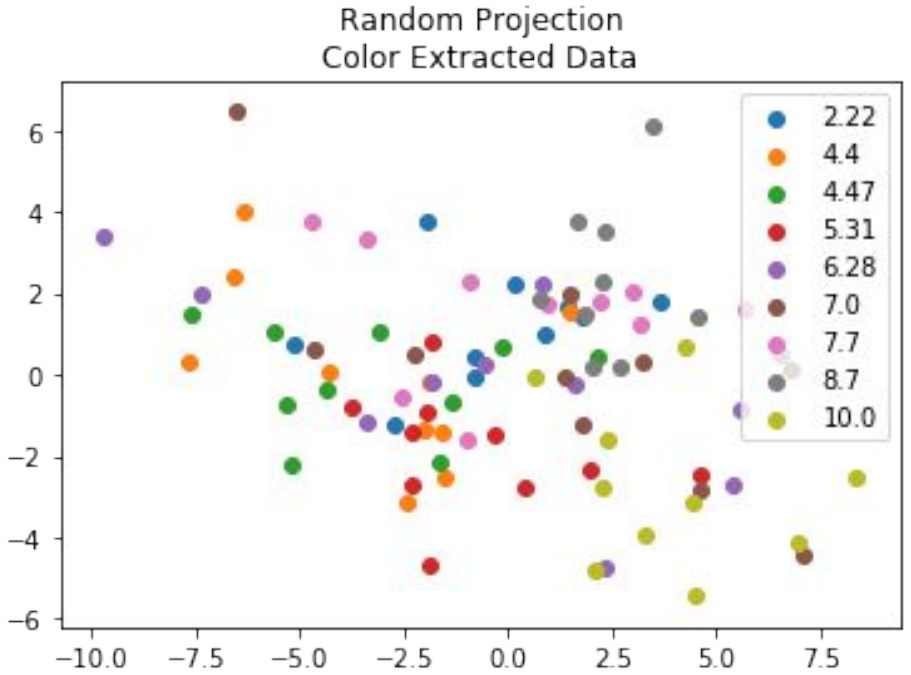
Original, final images

Projections to Reduce Dimensionality

Matthew Sit

Random Projection

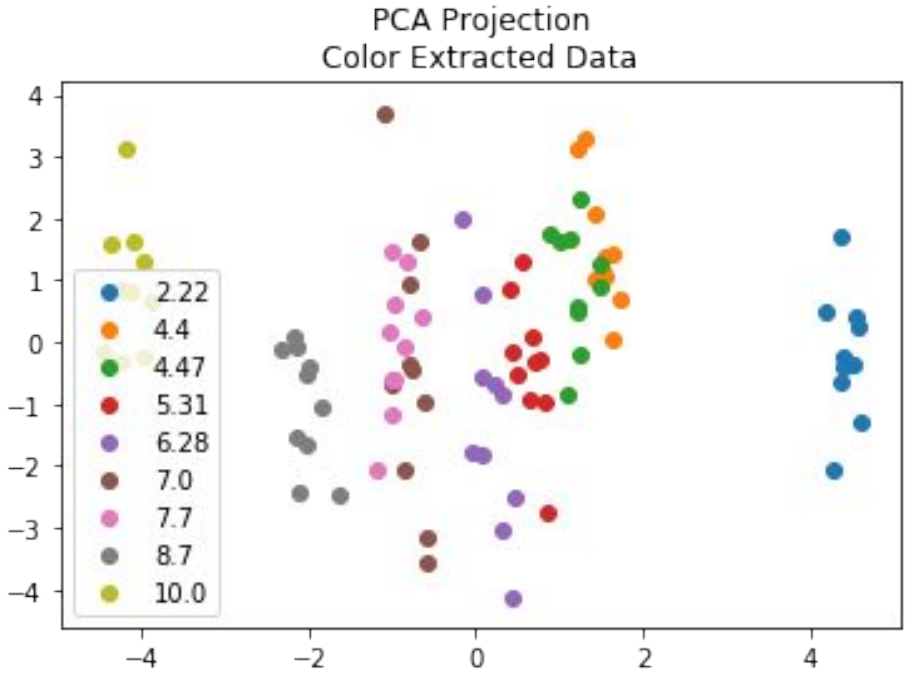
Flatten dimensions to a randomly chosen set of 2D axes.



PCA Projection

Flatten dimensions to a set of orthogonal 2D axes such that those axes chosen maximize the variance still illustrated by our data.

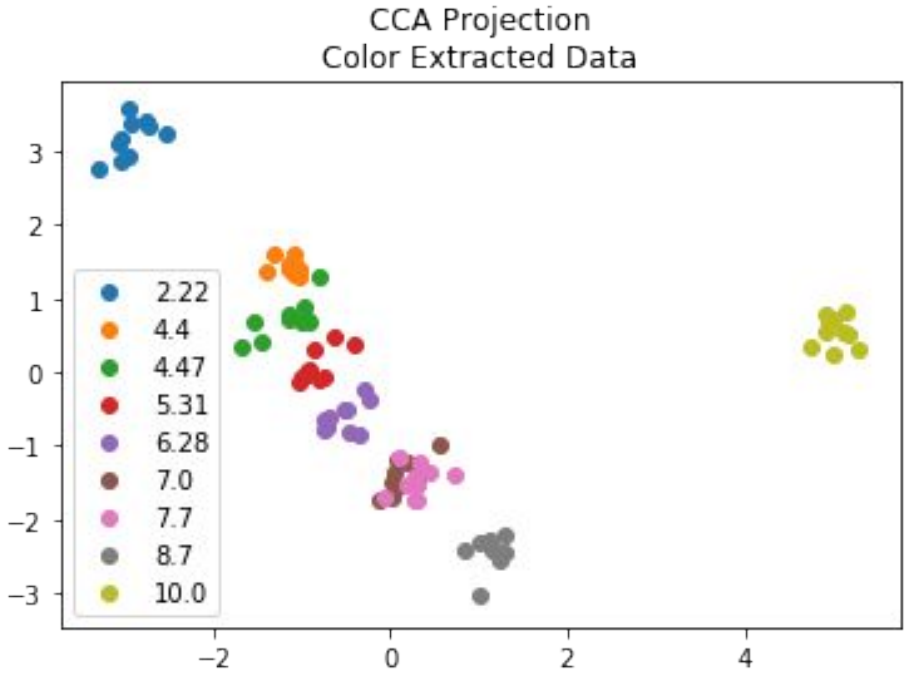
(Done through Principal Component Analysis.)



CCA Projection

Flatten dimensions to a set of 2D axes such that those axes chosen maximize correlation after projection.

(Done through Canonical Correlation Analysis.)

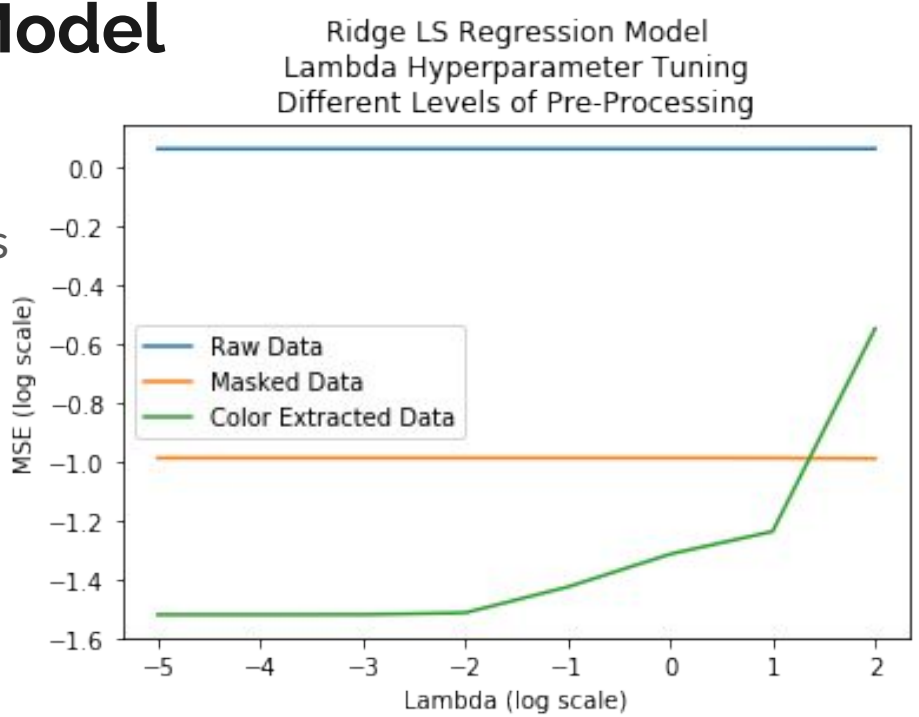


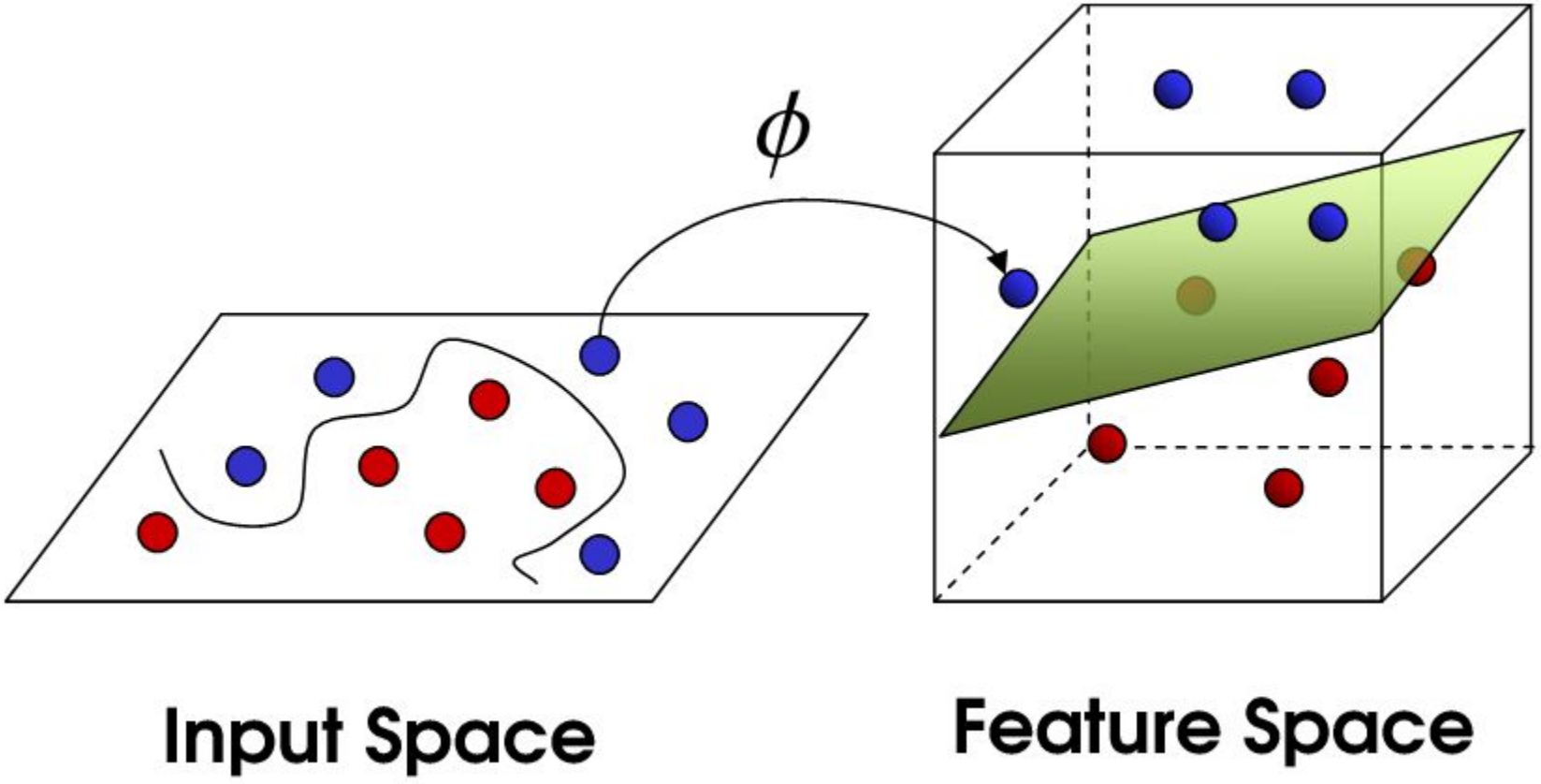
Hyperparameter Tuning the Least Squares Model

Matthew Sit

Ridge Least Squares Model

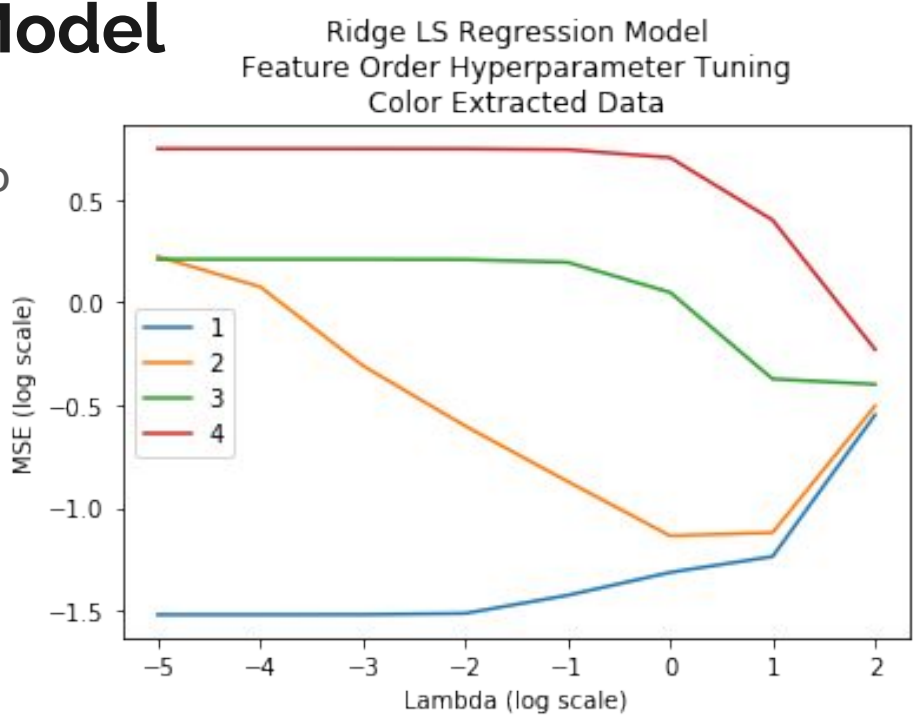
Prediction via linear regression.
Penalize linearly dependent weights to prevent overfitting.




ϕ
Input Space
Feature Space

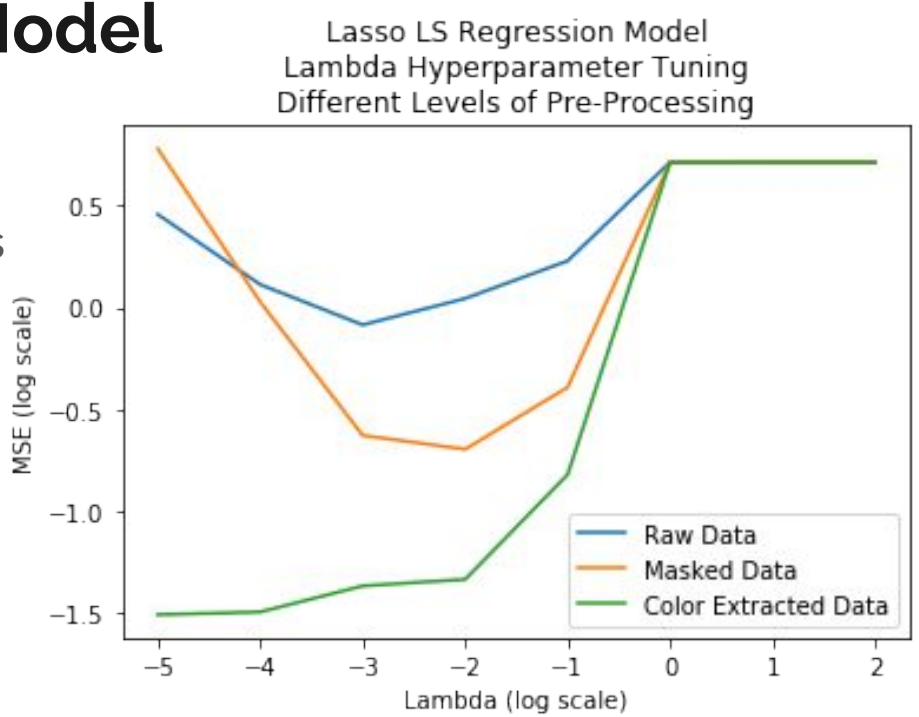
Ridge Least Squares Model

Lift features to higher dimensions to improve linear separability.



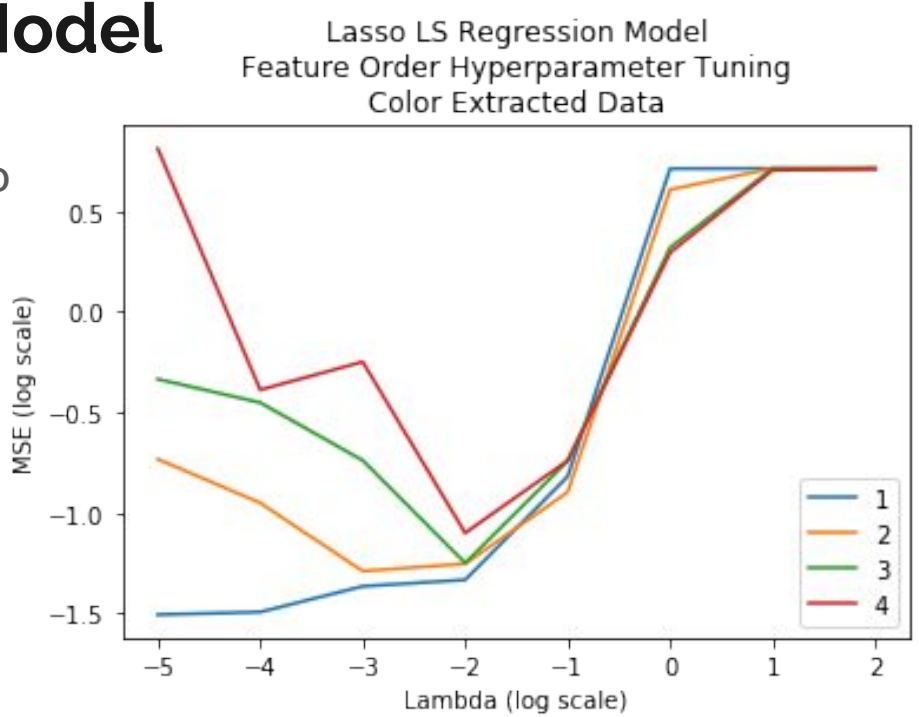
Lasso Least Squares Model

Prediction via linear regression.
Push less important feature weights to zero to eliminate them.



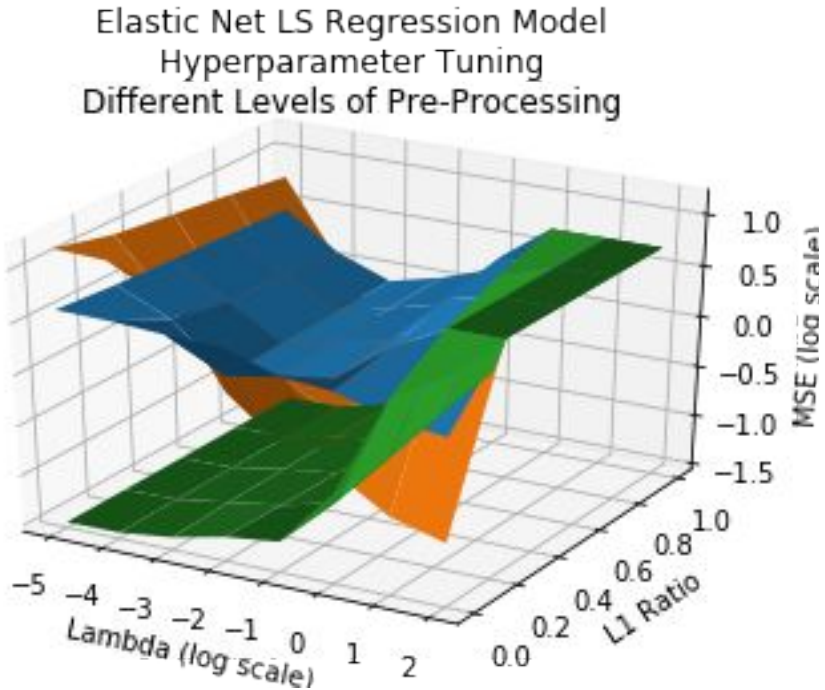
Lasso Least Squares Model

Lift features to higher dimensions to improve linear separability.



Elastic Net Model

Combine best of both ridge and lasso models.



More Complex Models

Rudra Mehta

Regression vs Classification

Regression

- More realistic - real pH can be decimals
- Buffered sample solutions were given with decimal pH's

Classification

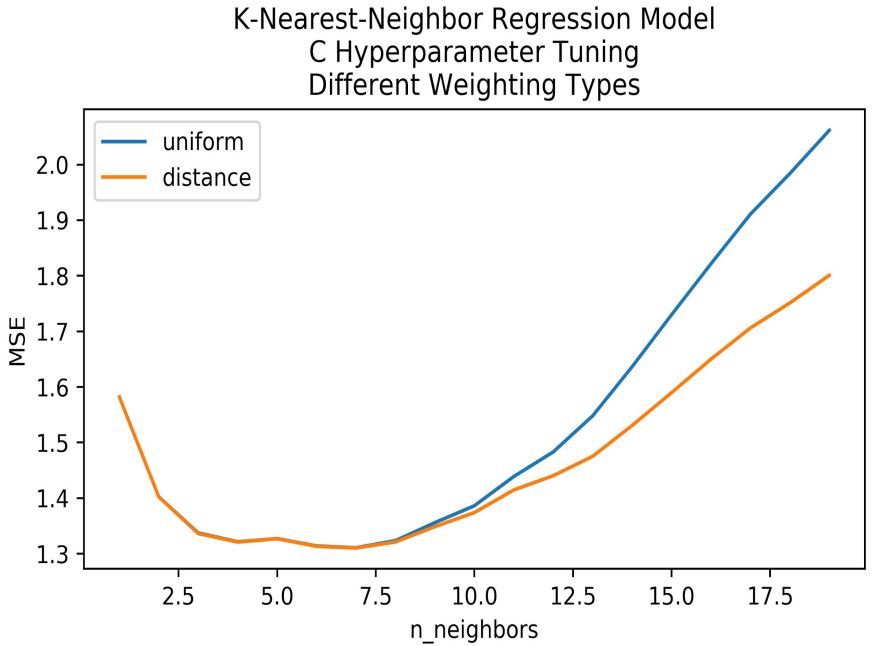
- Practical - a human using a pH strip would classify the pH as an integer
- Have to round the samples to closest integer

Models Tested

- Regression
  - Ridge
  - Lasso
  - Elastic Net
  - K-nearest-neighbors
- Classification
  - K-nearest-neighbors
  - SVM
  - LDA
  - QDA

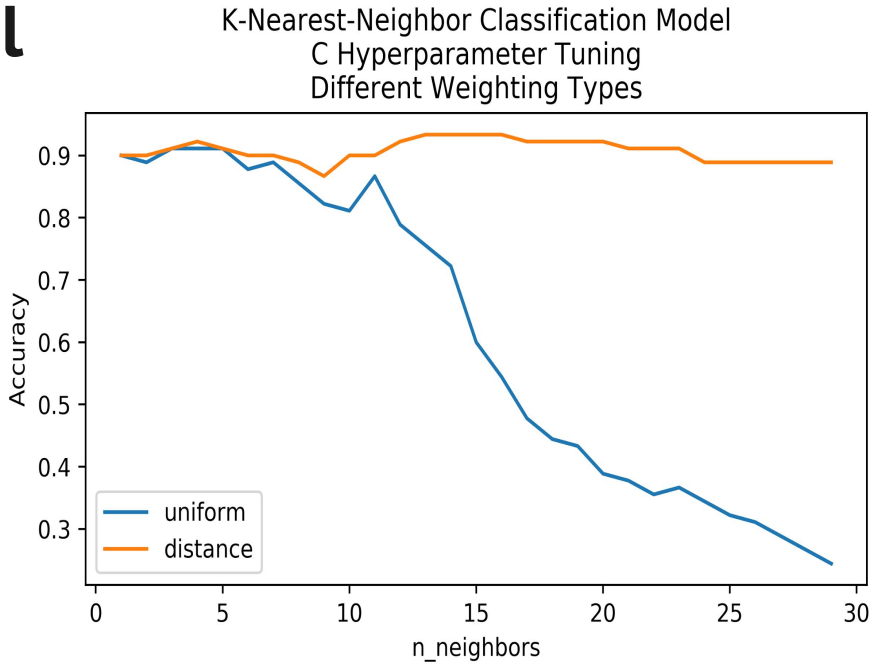
Regression KNN Model

Prediction based on similar samples.



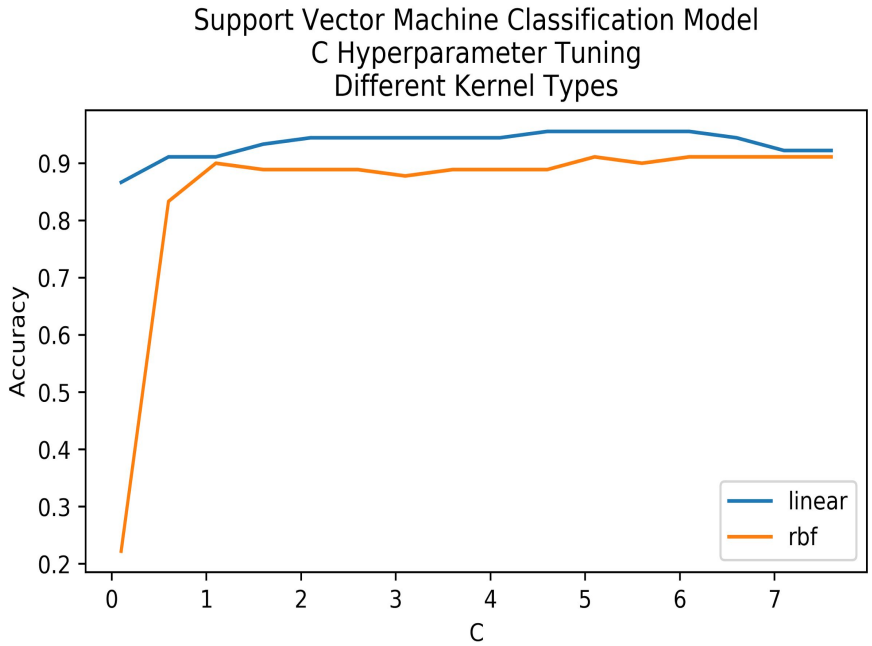
Classification KNN Model

Prediction based on similar samples.



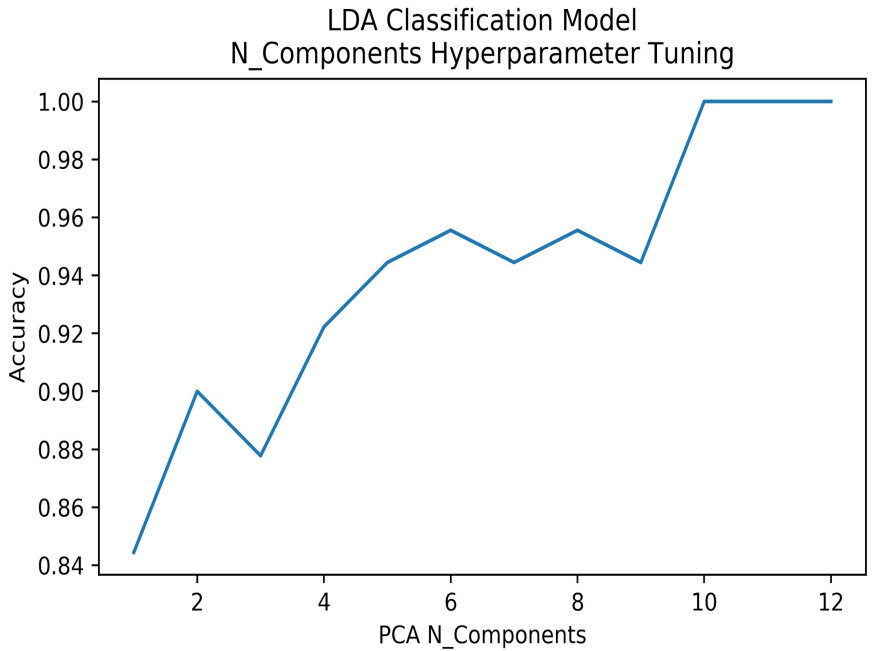
SVM Model

Classification via maximizing the margin between classes



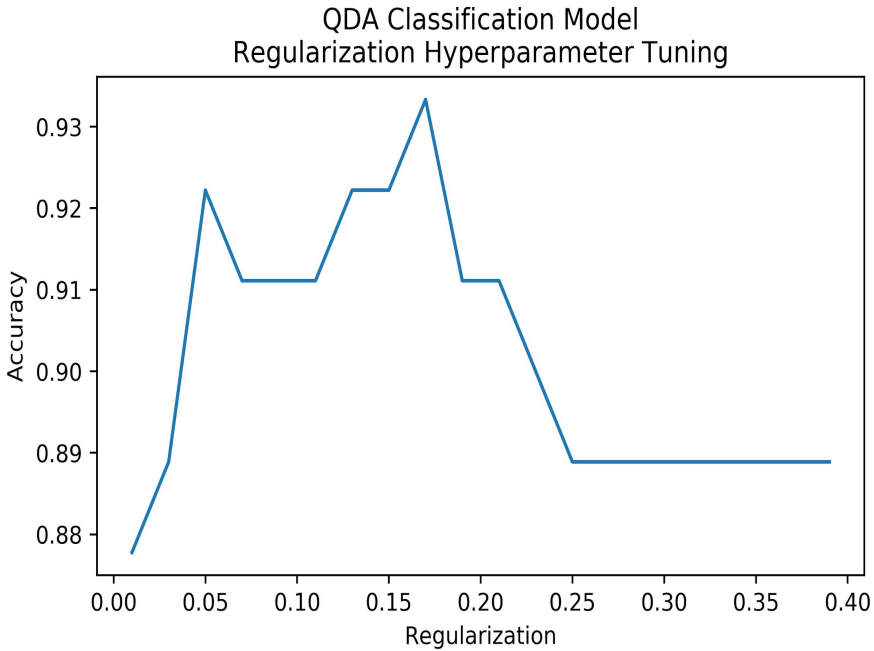
LDA Model

Classification via probabilities, using the same covariance for all classes



QDA Model

Classification via probabilities, using a unique covariance for each class



Grid Search Results

- Regression (MSE)
  - Ridge: 0.21175
  - Lasso: 0.1356
  - Elastic Net: 0.1295
  - KNN: 1.310
- Classification (Accuracy)
  - KNN: 93.33
  - SVM: 95.56
  - LDA: 100.00
  - QDA: 93.33

Pipeline Overview

Rudra Mehta

Function Features

Train or Test:

- Train: Choose best model using GridSearchCV from input training data
  - Save the best model to a file
- Test: Load given model, predict values for input images
  - Write values to a file

Currently the input images are the post-processed colors, need to add feature extraction code to this function