Capstone 2 Project Proposal: Distinguishing between Icebergs and Ships using Satellite Imagery

Should this proposal prove insufficient, I propose to enter the Kaggle competition for distinguishing ships from icebergs. This project offers a dataset that can be more directly linked to empirical data, as its datasets provide raw data. It also offers a chance to demonstrate skills in identifying information contained in photos and to compare my results directly with other data scientists by entering the competition.

Competition: https://www.kaggle.com/c/statoil-iceberg-classifier-challenge

Data

The competition provides two datasets, a training set and a test set. The training set provides radar returns for 1,604 images of satellite images of ships and icebergs as well as a dummy variable that identifies whether or not the image contains an iceberg. Those that do not have an iceberg have a ship. This is a balanced dataset, with 46.95% of the images have icebergs. The training set has 8,434 images but does not contain labels.

Two radar band returns are provided with each image in both datasets. Each band return provides 5,625 specific data points that, together, provide a 75x75 pixel image. The data contained in these values provide provide return values for each pixel, measured in dB. Band 1 is the HH, or transmit/receive horizontally, return. Band 2 is the HV, or transmit/receive vertically, return. These values will be used as hyperparameters. With more than 11,000 hyperparameters, data reduction will, of course, be absolutely necessary. Background research on how radar returns are structured will be used to guide the selection in how many factors to seek through a PCA.

Basic Competition Parameters

The goal of the competition is to predict ship or iceberg. Within the rules of the competition, this is done by providing a probability estimate for whether or not an image in the test data contains an iceberg. Entry deadlines are in mid-January, 2018. I should have this project completed in advance of that and so will submit my results to the competition.

Anticipated Analytic Steps

Data Quality Review

The first step must be to review the completeness of the data provided. This will be surprisingly challenging, based on the size of the dataset. With more than 11,000 hyperparameters, it will be nearly impossible to review each potential parameter closely. Nevertheless, I will attempt to review normality and missing data. I will do this by collecting lists of hyperparameters that fail to meet a certain threshold of completeness, perhaps those with more than 5% missing data, and those that fail normality checks. I will have my code report to me those hyperparameters which fail these checks rather than reviewing the results manually for all 11,000 hyperparameters.

More important will be review of the training outcome variable. The training data includes a dummy variable that is coded 0 for ships and 1 for icebergs, with every image containing one or the other. I will check this variable for balance to see if controls for unbalanced training data is necessary.

Data Reduction

As indicated above, with more than 11,000 hyperparameters, data reduction is absolutely necessary. With each hyperparameter representing the same pixel in each image, data reduction will allow us to group the returns by color, putting black pixels together into one factor, white into another, and additional shades into yet other factors. This will allow the data to focus on factors such as shape and shade.

I will start data reduction by doing separate reductions for each image band. I will start with an unbound PCA, allowing the data to determine the number of factors, reviewing using Scree or Variance Explained plots to guide selection in the number of factors. Depending on this result, I may select a smaller number of factors, guided by choosing a number of shades to include in images.

Following the reduction, I should have three groups of hyperparameters: Band 1 PCA factors, Band 2 PCA factors, and the incidence angle of the images, included in the initial dataset.

Classification Training

With a data label provided, this process should proceed using supervised learning and hyperparameters derived from the data reduction. Training data contains 1,604 observations. I will split off 20% of this data to use as testing data, since the formal testing data does not include any labels. This will allow me to test my method prior to submitting a result to the competition. I anticipate stepping through a series of increasingly complex classification tools:

1. Logistic Regression - used since final submissions must report the probability that an image contains an iceberg (P(iceberg == 1|hyperparameter values))

2. SVM Classifier - useful for smaller datasets, less than 10,000 observations. Training will use approximately 1,300 observations, making smaller sample tools appropriate.
3. I will subsequently explore KNN and Ensembles, including Random Forests, to see if I can improve performance.
4. Finally, I will attempt a Neural Network model to, again, see if this improves performance.
5. Once I have found the best of these methods, I will explore different ways of structuring the data to improve performance on the best estimator. Options include:
    a. A single PCA that includes both radar bands
    b. Varying the number of factors in the PCA
    c. Conditioning radar returns by incidence angle after PCA
    d. Conditioning radar returns by incidence angle before PCA

At each step, I will prepare and submit a result to the competition and obtain a measure of logloss on the formal training data and ranking relative to other participants. I will be recording the logloss and ranking after each submission.

Deliverables

1. Competition ranking
2. Final report on findings with substantive conclusions about the use of radar images to find icebergs and ships
3. Slide deck detailing project results
4. Jupyter/IPython Notebook containing code