Predicting agricultural productivity in California using satellite data and machine learning

Aakash Ahamed, Noah Dewar
CS229 Machine Learning Final Project

Abstract

Reliable forecasts of agricultural commodity production can have significant positive impacts on food security, water resources, and economic stability (Li et al, 2015). California is the most productive agricultural state and largest water user in the United States, making these estimate critical for responsible water management and agricultural decision making.

Open source satellite data provided by NASA and other agencies offers an unprecedented opportunity to measure the physical factors impacting agricultural yields with previously unattainable coverage, resolution and frequency. Recent advances in machine learning techniques facilitate improved statistical modeling of these complex, unstructured, data, and have the potential to greatly improve current methods.

In this study we explore the application of machine learning toward agricultural yield prediction from satellite imagery, and report accuracies that neural networks can predict agricultural production in California with 80-90% accuracy.

Introduction

California is the most productive agricultural region in the United States, producing over 46 billion USD in 2013 (NASS, 2013). Agriculture accounts for the vast majority of California's water use. In the wake of the 2010 - 2016 drought, water resources in California have come under unprecedented stress. Improved forecasts of agricultural yields can greatly benefit water managers seeking to optimize water use, growers seeking better management strategies, and economic agencies mandated to subsidize farm production and regulate commodity prices. Production forecasts are also of great interests to financial institutions who trade commodity futures.

Agricultural yield prediction, however, is not a trivial task. Yields vary as a function of a multitude of competing factors, including physical quantities like (1) precipitation, (2) temperature, (3) snowpack, and (4) surface runoff; management strategies like (1) irrigation, (2) fertilizer application, (3) harvest timing; and latent factors including (1) incomplete reporting, (2) market trends, and (3) pest and disease outbreaks.

We hypothesize that machine learning techniques can reconcile the complex, unstructured structure of satellite imagery and help elucidate trends and correlations between these data and agricultural yields. The inputs to our algorithm are satellite images capturing vegetation intensity and precipitation. Optical vegetation intensity images are filtered to encapsulate only agricultural areas in a given county. We then use a feedforward neural network to output a predicted agricultural productivity value at the county level for each year from 2000 - 2015.

Related Work

Many previous studies have attempted to forecast agricultural yields from satellite imagery (e.g. Quamby et al.,). However, traditional studies rely on physically based models, regressions, and moving

average techniques (e.g. Quamby et al., 1993; Bastiansen et al., 2003). Recent studies have successfully applied machine learning techniques to predict agricultural yields (e.g. You et al., 2017 use CNNs and deep gaussian process modeling). However, these investigations are typically performed in highly homogenous agricultural areas (e.g. US corn belt), and use data sources from a single satellite sensor. Few studies have explored the applicability of machine learning methods to predict agricultural yields across crop types in larger, heterogeneous areas.

Features and Data

This study used publicly available NASA satellite remote sensing data from the Moderate Resolution SpectroRadiometer (MODIS) to measure vegetation fluorescence (Tucker, 1979) and the Tropical Rainfall Measuring Mission (TRMM) satellite constellation to measure precipitation. All satellite data were accessed, filtered, and aggregated using the Google Earth Engine python API (https://earthengine.google.com/). Crop production data were accessed from the United States Department of Agriculture (USDA) National Agricultural Statistical Service.

Input Features: Satellite Data

MODIS satellite data (https://lpdaac.usgs.gov/dataset_discovery/modis/modis_products_table/mcd43a4) are available twice daily at 500m resolution for any area in the world from 2000 - present. MODIS data were subset to include only the spectral wavelengths for which vegetation has a strong signature (Tucker, 1979), and monthly per-pixel averages were computed in order to decrease data volumes. The Normalized Difference Vegetation Index (NDVI; Ticker, 1979) is used in order to normalize data from 0-1.

TRMM satellite data (https://pmm.nasa.gov/data-access/downloads/trmm) area available every 3 hours at 25km resolution for any area in the world from 1997 - present. Monthly per-pixel sums were computed for the time period of 2000 - present, in order to coincide with MODIS data.

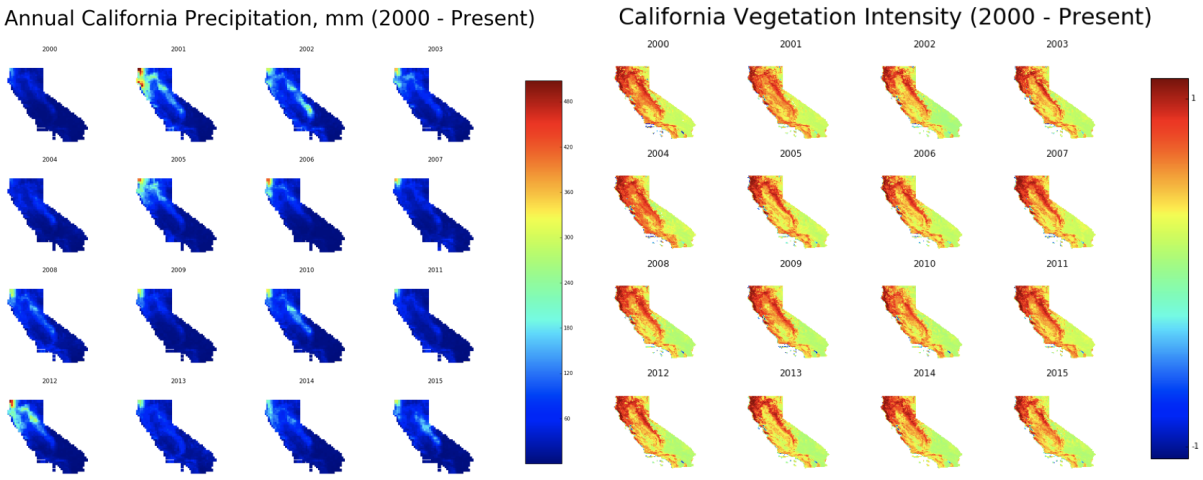

Figure 1: Example of TRMM precipitation data (left) and MODIS vegetation fluorescence data (right) used as input features in this study.

Training Data: County Level Agricultural Yields

Agricultural production data for California were accessed from the USDA NASS (https://www.nass.usda.gov/Statistics_by_State/California/Publications/AgComm/Detail/index.php). Data were preprocessed to filter out irrelevant crops, like pasture, horse, cattle, etc.

Preprocessing: Filtering out non-agricultural areas

Non agricultural areas were removed from all MODIS images in order to prevent signal dampening of vegetation. Data describing agricultural areas was obtained from the CA Department of Water Resources (https://gis.water.ca.gov/app/CADWRLandUseViewer/downloads/atlas_i15_CropMapping2014.zip).

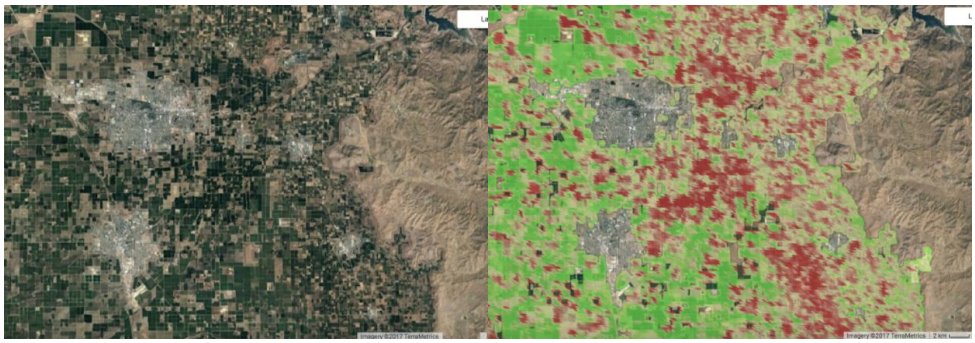

Figure 2: example of filtering of non-agricultural areas for Tulare County.

Methods

All methods described here were tested on two highly productive agricultural counties, Tulare and Fresno. Future work will involve applying the methods described below to all counties in California.

Naïve Bayes

The Naïve Bayes implementation presented here used a multinomial event model and Laplace Smoothing. Input features were MODIS vegetation fluorescence and TRMM precipitation. Fluorescence values for each year of MODIS data were binned in equal width bins starting at 0 and moving over by 0.025 per bin, with the last bin at 0.975 to 1. The value in each bin was the number of times a pixel with that value appeared in that years MODIS image. The TRMM data was binned in a similar manner over 10 bins and then scaled with the mean of all the bins and all years of the MODIS data. The yield data is discretized similarly in order to map data from continuous labels to discrete classes. Each year of yield data paired with MODIS data was an example in the training and test sets. The years of data used ran from 2000 to 2015 and included both Tulare County and Fresno County.

Leave one out cross validation was performed across all 32 examples from Tulare and Fresno. The error was calculated by comparing the middle value of the predicted yield bin for the left out year with its true value. Figure 3 shows the test and training accuracy for Tulare and Fresno on the left and on the right a plot of predicted production values from the hold one out cross validation versus true production values.

Neural network

The matlab neural net toolbox was used to construct and train a variety of feedforward neural nets for non-linear regression. Several different architectures and training functions were tested and evaluated by hold one out cross validation across all years of data from both Tulare and Fresno. The architecture used for the final model (Figure 3) was a 2 layer net with hyperbolic tangent activation functions for the hidden layers and a linear activation function for the output layer.

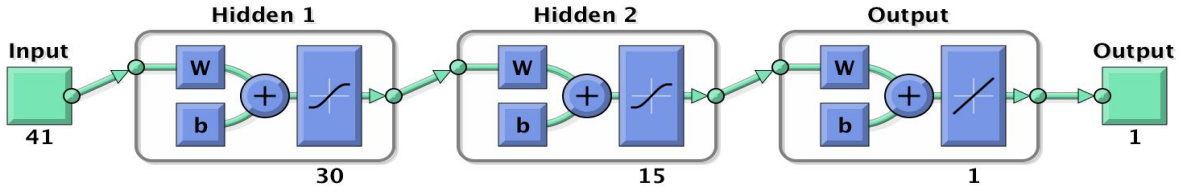

Figure 3: Architecture of the neural network.

Nets were trained with Levenberg-Marquardt optimization (LM), Gradient descent with momentum and adaptive learning rate backpropagation (GDX), and Bayesian Regularization optimization (BR). LM and GDX trained two orders of magnitude faster than BR, however, accuracies from LM and GDX were generally worse, so BR was chosen as the preferred method.

A Monte Carlo stochastic optimization scheme was run on all relevant hyperparameters (learning rate, batch size, number of epochs, etc) to try and examine sensitivity between the hyperparameters and the model accuracy. This was done by training a few thousand neural nets with hyperparameters that were randomly sampled from some uniform distribution of acceptable values. The resulting prediction accuracies for each set of hyperparameters were split into a top third and bottom two thirds of accuracy values. The difference between the distribution of hyperparameters in the top third and the bottom two thirds was then taken as the effect each hyperparameter has on ultimate model accuracy.

Results and Discussion

The methods described above were trained and tested on 2 of the most prolific agricultural counties (Tulare and Fresno) in order to save time and judge efficacy. Preliminary results of hold one out cross validation are promising, with overall test and training accuracies of 70% and of 73% for Naive Bayes (Figure 4) and 85% and 96% for the feedforward neural network with Bayesian regularization (Figure 5). The neural network results (Figure 5) show the test and training accuracies (left) for the final feedforward net (i.e. Figure 3), trained with BR optimization, for hold one out cross validation for all years of data from both Tulare and Fresno (Figure 5, left panel) as well as the predicted production values versus the true production values for the same model (right panel). The hold one out cross validation was performed by holding one year of data, and then training on all of the other years of data for both Tulare and Fresno, then holding out the next year, and so on until each year from both counties had individually been held out and tested upon.

Initial results, though promising, have substantial potential for improvement. Increasing the feature space or altering data aggregation schemes could improve results. Deepening the feedforward net so it contains more layers and training it on more counties is also likely to improve test accuracies. The application of convolutional neural nets could also serve to improve results, since spatial relationships are an important component of satellite data and yields. Deep convolutional networks could learn not just how vegetation fluorescence and other features of satellite imagery are linked to agricultural production,

but also what spatial groupings of the same features are common across counties and how they influence agricultural production.

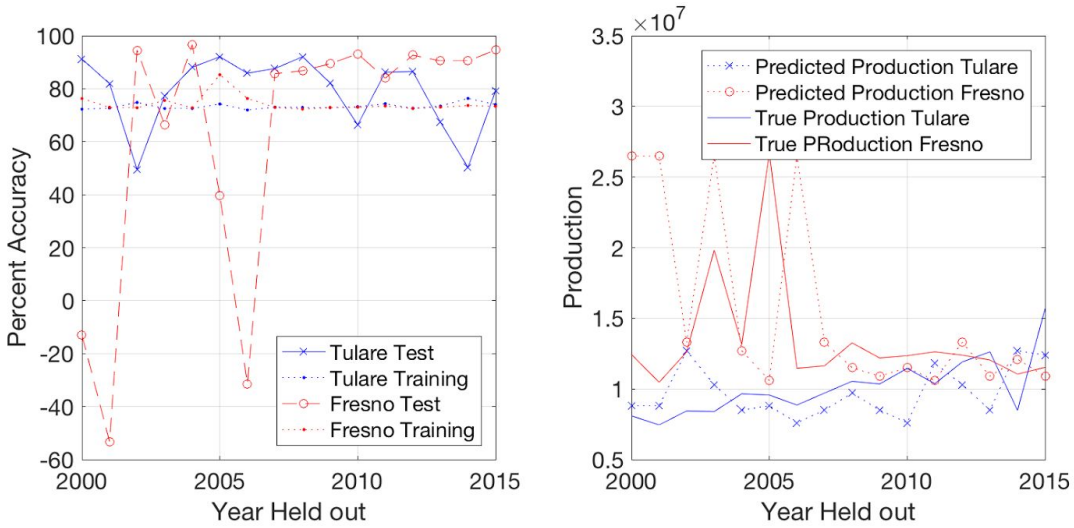

Figure 4: Hold one out cross validation results for Naive Bayes.

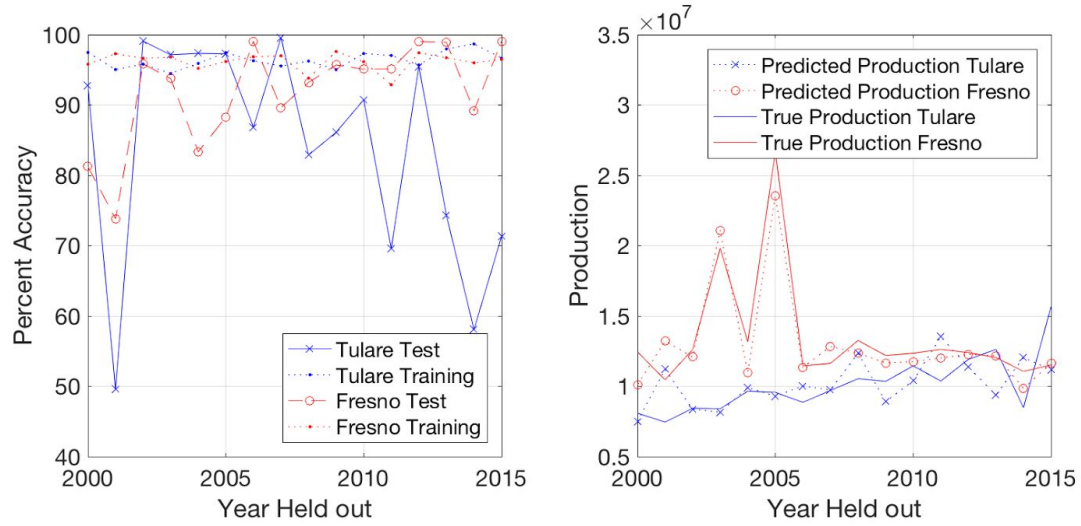

Figure 5, hold on out cross validation results for the final feedforward neural net.

Future Work

The study described here will continue in a number of directions. Additional counties, satellites, and variables will be considered in training. A sensitivity analysis for each variable and aggregation scheme will be examined using either a grid search and a stochastic search method. Convolutional neural networks will also be tested, and the performance of the further tuned feedforward neural net will be compared.

To summarize, despite favorable initial performance for simple models, enhanced performance could be achieved by (1) optimizing model parameters, (2) adjusting the preprocessing of input features, and (3) adding additional satellites and satellite derived variables as input features.

Contributions

The satellite imagery aggregation and processing was done by Aakash.
The adaption of the Naive Bayes code from problem set 2 was done by Noah
The background research into remote sensing products to be used for this project was done by Aakash.
The method of flattening and binning the NDVI and precipitation values to reduce the dimension of the dataspace was done by Noah.
The collection and processing of the agricultural production data from the USDA tables was done by Aakash.
The utilization of the matlab neural net toolbox to construct and train the feedforward nets was done by Noah.
The final report was written and edited by both Aakash and Noah.

References

Bastiaanssen, W. G., & Ali, S. (2003). A new crop yield forecasting model based on satellite measurements applied across the Indus Basin, Pakistan. *Agriculture, ecosystems & environment*, *94*(3), 321-340.

California Department of Water Resources. Report to the Governor's Drought Task Force – Groundwater Basins with Potential Water Shortages and Gaps in Groundwater Monitoring, April 2014 2014a. p.51. Sacramento, CA.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790-794

Li, T., Hasegawa, T., Yin, X., Zhu, Y., Boote, K., Adam, M., Bregaglio, S., Buis, S., Confalonieri, R., Fumoto, T., Gaydon, D., Marcaida, M., Nakagawa, H., Oriol, P., Ruane, A. C., Ruget, F., Singh, B.-, Singh, U., Tang, L., Tao, F., Wilkens, P., Yoshida, H., Zhang, Z. and Bouman, B. (2015), Uncertainties in predicting rice yield by current crop models under a wide range of climatic conditions. Glob Change Biol, 21: 1328–1341. doi:10.1111/gcb.12758

Quarmby, N. A., Milnes, M., Hindle, T. L., & Silleos, N. (1993). The use of multi-temporal NDVI measurements from AVHRR data for crop yield estimation and prediction. *International Journal of Remote Sensing*, *14*(2), 199-210.

Tucker, C. J. (1979). Red and photographic infrared linear combinations for monitoring vegetation. *Remote sensing of Environment*, *8*(2), 127-150.

You, J., Li, X., Low, M., Lobell, D., & Ermon, S. (2017). Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data. In *AAAI* (pp. 4559-4566).