Projecting Spring Arrival With Functional Temperature Data

Cecilia Earls

The onset of Spring is widely anticipated throughout the northern hemisphere. While Spring's arrival universally brings the promise of warmer weather, it has special significance for the farmer whose livelihood depends on the weather. If an early Spring can be predicted with a reasonable level of accuracy a significant amount of time before its arrival, that information could be utilized in agricultural planning to reduce costs, increase yield, and prevent waste.

This project's focus is to find the machine learning algorithm that can predict the onset of an early Spring most accurately and timely using temperature as a predictor. The *Data* section below describes the nature of the data used in analysis and how missing data is handled. *Framework* specifically outlines the assumptions and definitions used in this analysis. The following section presents the methodologies considered and provides a synopsis of results. The last section discusses which methodologies perform best under different circumstances and the direction of future work.

Data

Temperature and "greenness" data are recorded via satellite from 33 different geographic regions from January 2001 - December 2008 in 8 day intervals. Thus, for each year and each region, there are 46 records of temperature and "greenness". However, data collected via satellite are sensitive to cloud cover. Consequently, about 27% of the data are missing.

Temperature is recorded in Kelvin for each 8 day period. Missing temperature data are approximated by the average of the closest temperature observations available.

The "greenness" data are vegetation indices developed by Earth scientists that quantify the density of green leaf vegetation in our environment[3]. Within each year, for each of the 33 land plots, "Spring Arrival" is determined by the timing of the greatest increase in greenness in the period from early April to early June. To assure reasonable classification, each set of 8 years of data, for the 33 land plots, is looked at individually and classified

in relationship to the average onset of Spring's arrival for that particular plot. Determining a classification of "early Spring" versus "average or late Spring" is not sensitive to missing values because a range of time points map to each classification level. In this analysis, a response of 0 corresponds to "early Spring" and a response of 1 corresponds to "average or late Spring."

Framework

The focus of this project is to assess the ability to predict an early Spring based only on temperature data in the prior year. Clearly, if a classifier always predicts an early Spring in Florida and a late Spring in New York, the model essentially is an expensive indicator of geographic location. Consequently, each year/plot combination is classified in relation only to the other greenness profiles for that particular plot over the 8 years of data collection. Simply said, in Florida, the model should predict an early Spring for Florida, not an early Spring for New York.

In general, for each plot, "Spring Arrival" varies within a time period of approximately 15-30 days. If the variation within the 8 years for a particular geographic plot is not at least 15 days, all years are classified as "average or late Spring."

For most of the 33 geographic plots, the 8 records of greenness are variable enough to delineate between an early Spring arrival versus an average or late Spring arrival. However, clear indication of an early Spring arrival is demanded of the data to be classified accordingly. The final data consist of 223 observations of which 48 are classified as "early Spring."

For all methodologies except Gaussian Discriminant Analysis, temperature is considered only after the observations for the time period of interest are smoothed into a functional data format[2]. This data transformation filters out some of the erratic fluctuations of temperature to uncover a smoothed temperature trend which is utilized as feature data. In some cases, such as the logistic regression model, it will be more convenient (and sometimes necessary) to use a discrete approximation to the smoothed temperature function. Gaussian Discriminant Analysis is the only methodology for which results are based on actual temperature observations that are not smoothed.

Finally, to obtain consistent and comparable results, the same training and testing sets are used for all methodologies with one exception. Local Gaussian Discriminant Analysis (LGDA) uses subsets of the original training and testing sets in order to distinctly separate the data into two groups for which the observations either have an average Winter temperature below 274 degrees Kelvin or an average Winter temperature above 294 degrees Kelvin. Observations that have an average Winter temperature between

274 and 294 degrees Kelvin are excluded from the LGDA analysis.

Methodologies and Results

The following three sections describe in detail the classification procedures used in this investigation. The table in the *Results* section is a synopsis of classification performance as indicated by the true positive rate (TPR) and false positive rate (FPR), for each methodology, determined by classifying the testing data.

Logistic Regression

Logistic regression[1] is the first model considered as an "early Spring" classifier. Two different sets of features are considered: the finite evaluation of the temperature function (FLR), and the finite evaluation of the derivative of the temperature function(DLR).

Originally, temperature data from August to March is considered for FLR features. After an initial run, the surprising and interesting development is that the most significant [1] features of this model are temperatures in late November through early December. Consequently, if logistic regression classifies the data accurately, by the end of December, the probability of an early Spring in the next year can be estimated.

Next, the derivative of each temperature function is considered as feature data(DLR). After varying the range of features needed for accurate predictions, the final feature set is the evaluation of the derivative of the entire temperature function from August to March.

The basic logistic regression model set-up is:

$\pi_i \equiv$ probability of a late or average Spring

$\pi_i = \frac{1}{1+e^{-\eta_i}}$

$\eta_i =$inprod$(\beta, features_i)$

Classification of the testing data is determined as follows (hats indicate estimates of the above parameters):

$\hat{y}_i = 1$ if $\hat{\eta}_i \geq$ C and $\hat{y}_i = 0$ if $\hat{\eta}_i <$ C

The parameter C above is subjectively determined using the training

[1]Significance is determined by the results of a Wald test on the estimated regression coefficients.

data, with an eye for an "ideal" balance of true positives versus false positives. If C is low, an average or late Spring will always be our estimate which gives a low true positive rate and low false positive rate. At the other extreme, if C is large, an early Spring will always be our estimate which results in a high true positive rate and a high false positive rate. For this analysis, C is set equal to 1, as determined reasonable based on the training data.

In practice, the value of C is uneccessary; decisions can be made at the individual level using the estimated probability of an early Spring, $1 - \hat{\pi}_i$.

KNN

The second approach considered for classifying data by the timing of Spring's arrival is a simple k-nearest neighbor algorithm[1].

Under this model, for each smoothed temperature function of the testing data, the inner product of the difference of this temperature function and temperature profiles of the training data were compared to determine the testing datum's k nearest neighbors. If the average response of the k-nearest neighbors was greater than or equal to one-half, the datum was categorized as "average or late Spring," otherwise the datum was categorized as "early Spring." A final value of k=3 to classify the 55 testing observations is determined by 4-fold cross-validation on the training data.

Results are obtained on the full range of temperature data from August to March(FKNN) as well as on a reduced feature set of temperatures from August to the beginning of February(RKNN). As presented below in the *Results* section, kNN with the reduced feature set actually performs slightly better.

Gaussian Discriminant Analysis

For this project, Gaussian Discriminant Analysis[1] is considered in two different ways. First, the entire testing set is classified using GDA with parameters determined by the original training set. In addition, to discover whether a more "local" variation of GDA might perform better, the training and testing sets are split into two groups based on average temperature from mid-November to late February. Then, GDA is performed individually on the two separate groups(designated LGDA-High and LGDA-Low). A small number of observations for which neither the designation of high average temperature or low average temperature is appropriate are excluded from the training and testing sets.

While, as seen below, splitting the data by low and high average temperature did not induce better performance overall; this portion of the analysis demonstrates that GDA is remarkably accurate for areas with low average temperature in Winter.

Results

The following table summarizes results over all methodologies. For each technique, the range of temperature data presented is determined as "ideal" for that methodology. Ideal is defined as the smallest range of temperature data for which classification performance is unaffected.

Methodology	Temperature Range	TPR	FPR
Baseline*	N/A	.18	.18
FLR	Mid-Nov. to Early Dec.	.55	.20
GDA	Mid-Nov. to Late Feb.	.64	.11
LGDA-High	Mid-Nov. to Late Feb.	.33	.15
LGDA-Low	Mid-Nov. to Late Feb.	1.00	.09
RKNN	Aug. to Early Feb.	.55	.09
FKNN	Aug. to March	.55	.11
DLR	Aug. to March	.82	.25

*The baseline true positive and false positive rates are determined by randomly assigning each test observation as "Early Spring" with a probability of 22% and as an "Average or Late Spring" with a probability of 78%. These probabilities are the maximum likelihood estimates of the chance of an "Early Spring" or an "Average or Late Spring," as determined on the training data.

Assessment and Future Work

The goal of this analyses is to determine a timely and accurate algorithm for predicting whether an early Spring is to be expected for a particular geographic region.

Clearly, the results show temperature data can be used to predict the nature of Spring's arrival with some accuracy. Determining the method which performs "best" is dependent on the time for which the prediction is desired.

The best overall accuracy is achieved by logistic regression with the evaluated derivative function as features, but this method also requires the most temperature data of all methods considered. If an early prediction is desired, logistic regression with the evaluated temperature function as features does reasonably well with a 55% TPR and a 20% FPR. However, if a prediction 2 months before May is adequate, Gaussian Discriminant Analysis is preferable with a higher accuracy rate and much lower FPR than logistic regression.

While, ideally, the desired procedure is invariant to geography, the results imply there is some characteristic of areas with low average Winter temperature that lends observations in this group to accurate classification under Gaussian Discriminant Analysis. If only these areas are considered, GDA classifies all early Springs appropriately on the testing data with a very low FPR. This result is quite remarkable, and further work needs to be done to examine why the classification performance in these areas are substantially superior to performance in areas with a higher average Winter temperature.

In summary, the results of this investigation are quite promising. However, additional research needs to be done to verify these preliminary results and further improve on classification accuracy.

References

[1] Bishop,C. (2006), *Pattern Recognition and Machine Learning,* Springer Science + Business Media, LLC.

[2] Ramsay, J., Hooker, G., and Graves S.(2009), *Functional Data Analysis with R and Matlab*, Springer.

[3] Weier, J., and Herring,D.,"Measuring Vegetation," *Earth Observatory*, NASA, 19 October 2010, http://earthobservatory.nasa.gov/Features/Measuring Vegetation/.