Assignment 5, SVM & Trees

Ethan Tenison

4/25/2020

# Q1: Support Vector Machines in the abstract (20 points)

Energy poverty is at the center of a current discussion of equity. Electric Service Payment Assistance is one of the top reasons people call the [2-1-1 Navigation Center](https://www.unitedwayaustin.org/CNT2017/) operated by United Way for Greater Austin. Local organizations like [TEPRI](http://www.txenergypoverty.org/) are active in working to address energy poverty in Texas. Suppose you had access to geospatial data of the location of 2-1-1 calls in Central Texas. Some of these calls are related to energy bill assistance, others have other topics. Your goal is to uncover communities facing energy poverty.

### (a) In the context of the support vector classifier, discuss the role of margins, the tuning parameter, and the relationship between them.

The maximal margin classifier selects a hyperplane, out of an infinite number, that has the greatest perpendicular distance from all of the observations on the margin, called the support vectors. This is hypeyplane is called the maximimal margin hyperplane. This often leads to overfitting, as only a few number of observations determine the hyperplane. SVC use what is called a “soft margin”, where some observations are allowed to be on the wrong side of the boundary. This is where the tuning parameter comes in. The tuning parameter C is used to decide how many incorrectly labeled observations are allowed within the margins of the hyperplane. The soft margin is set by ensuring the sum of the error terms is less than or equal to C, which is chosen through cross validation to find the optimal point for the variance and bias trade off.

### (b) Suppose you use SVM with a radial kernel to classify regions by x,y coordinates. Explain the intuition behind how a radial kernel works. If, in the future, more calls related to energy bill assistance come from inside a region that you have identified as a communities facing energy poverty, how does this impact your SVM? Why? What does it mean if a call comes in that results in a change to the SVM?

Radial kernels help us to generate non-linear decision boundaries when our data is inherently not linear. Kernels in general create nonlinear boundaries by doing what’s called feature expansion, where we convert features to higher dimensions. This can be done with polynomial kernels, but as we increase the number of dimensions we have to deal with the problem of overfitting. We can solve this issue by using a radial kernel, which can have a smoother decision boundaries based on the tuning parameters. Based on the textbook, it also does better when the data is inherently circular.

# Q2: Regression trees in data (40 points)

A portion of residential energy consumption goes towards a legitimate end use, but there is plenty of room to increase energy efficiency in homes (for example, by repairing leaky ducts). Often, utilities themselves offer rebates for their customers to increase their energy efficiency, e.g. [Austin Energy’s free Weatherization program](https://savings.austinenergy.com/rebates/residential/offerings/home-improvements/weatherization). Targeting these resources to homes that could benefit from them can be difficult. Use the data from [Austin’s Energy Conservation Audit and Disclosure](https://data.austintexas.gov/Utilities-and-City-Services/Map-2009-2013-ECAD-Residential-Energy-Audit-Data/wqi4-2ivr) to devise a simple way to sort homes that identifies those with high levels of duct leakage. Please consult and adapt the included SAL\_A5q2 script to clean the data (and generate some nifty maps along the way).

### (a) Grow a decision tree to predict System 1 Duct Leakage in the training data on the basis of the

### sort of simple questions you could ask someone over the phone, e.g. “Where is your system?”;

### “Is your furnace gas or electric?”; “What sort of home are you in?” Grow a very large tree and

### print it (hint: add mindev = 0.001 and minsize = 5, you may need to adjust mindev).

###(b) Use cross validation to prune the tree. What is the best size? Print the pruned tree. Interpret the ### pruned tree. How would you use this to inform targeting weatherization resources to those that ### need them?

###(c) Suppose the top 20% of leakage need immediate attention. Predict leakage using the pruned ### tree and apply the classification threshold. Print the confusion matrix for the test data. Calculate ### and discuss the error types, rates, and consequences in the context of this problem.

###(d) Comment on the “interestingness” of your findings. Identify the audience for these findings. ###Craft a concise message to convey these findings in as interesting a way as you can.