Team Checkpoint 1

Forest Cover Team B

Annie Condon, Matt Hayden, Matt Robertson, Yvette Gonzalez

Contents

1 Introduction

The U.S. Forest Service relies on an accurate understanding of its forests' composition in order to best protect and manage the forest land. Conducting accurate inventory of forest composition by direct observation or remotely sensed data is often too expensive and time consuming to do at large-scale. Predictive analytics, however, can be employed to use the results of a small-scale survey to create a model that can be applied across a large region, using descriptive features extracted from maps of the area.

In this paper, our objective is to predict the forest cover type given a set of cartographic features and a variety of mulitnomial classification models. Our models will be evaluated using predictive accuracy.

2 The Modeling Problem

Our modeling problem is to predict the forest cover type as a multiclass classification problem based on the associated features. Our forest cover type is defined as

While some classification algorithms naturally permit the use of more than two classes, others are by nature binary algorithms. In these cases we will have to extend the exisiting binary classifiers to solve multi-class classification problems.

Table 1 shows the seven different kinds of forest type found in our data set, along with the number of observations for each type.

Table 1: Cover Types

Cover Type	Number of Observations
Spruce/Fir	211840
Lodgepole Pine	283301
Ponderosa Pine	35754
Cottonwood/Willow	2747
Aspen	9493
Douglas/Fir	17367
Krummholz	20510

Our modeling problem is a multiclass classification problem. Blurb on multiclass classification

What types of modeling methods work on these kinds of problems

3 The Data

The actual forest cover type for a given 30 x 30 meter cell was determined from US Forest Service (USFS) Region 2 Resource Information System data. Independent variables were then derived from data obtained from the US Geological Survey and USFS. The data is in raw form (not scaled) and contains binary columns of data for qualitative independent variables such as wilderness areas and soil type.

4 Exploratory Data Analysis