Alex Berry
Data 1030
Semester Project Proposal
2019.09.24

Introduction

I propose a project that will attempt to assign patterns of medieval handwriting from the 12th-century Bible of Ávila to one of the twelve copyists who created the manuscript. The target variable is the identity of the copyist who wrote the script represented by a given data point. This is a classification problem with twelve classes. This problem is interesting because it is an application of machine learning to a question in the humanities that requires years of experience and labor to answer when performed manually. A computational approach to the characterization of medieval handwriting can replace qualitative measurements with quantitative ones, and can analyze and categorize observations more rapidly and systematically than in the past. I am interested in this project because my undergraduate history advisor was a paleographer specializing in medieval manuscripts and I have always wanted to apply a computational way of thinking to some of the problems she encountered in her work.

Dataset

This dataset is from the UCI Machine Learning Repository and consists of 20867 data points and 10 features. There are no missing values.

Feature	Feature Type	Description
Intercolumnar distance	Page layout	Distance between two columns on the page
Upper margin	Page layout	Distance between top edge of page and script
Lower margin	Page layout	Distance between bottom edge of page and script
Exploitation	Column	Density of script in a column
Row number	Column	Number of rows per column
Modular ratio	Column	Level of a row at which a center zone occurs
Interlinear spacing	Column	Distance between center zones in rows
Weight	Row	Density of script in a row
Peak number	Row	Number of peaks in hortizontal density of a row
Modular ratio/interlinear spacing	Row	Ratio of modular ratio to interlinear spacing

This dataset's entry on the UCI ML Repository describes its origin in the paper "Reliable writer identification in medieval manuscripts through page layout features: The 'Avila' Bible case" by C. De Stefano, M. Maniaci, F. Fontanella, and A. Scotto di Freca in *Engineering Applications of Artificial Intelligence*, Volume 72, 2018, pp. 99-110. This project's main goal was the verification of the

classification power of certain specifically devised features concerning page layout. The authors found that these some of these features are very effective at classifying authorship. I hope to have similar results once this project is complete.

Preprocessing

As available on the UCI ML Repository, this dataset is already preprocessed using the standard scaling method and divided into two datasets: a training set containing 10,430 samples, and a test set containing the 10,437 samples. However, the target variable column's entries are letters A-I and W-Y, so I used a LabelEncoder to transform this column's entries from letters to integers 0-11. This format of integers for the target classes is necessary for classification. The preprocessed dataset has 10 features and 1 target feature.