



Digital Paleography of the Ávila Bible



Classification of handwriting in a medieval manuscript

Alexander Berry
Brown University

Tuesday, October 22, 2019

Data set, code, figures, and reports are available at:

<https://github.com/ABerry057/avila/>

Problem Description:

Technical Details

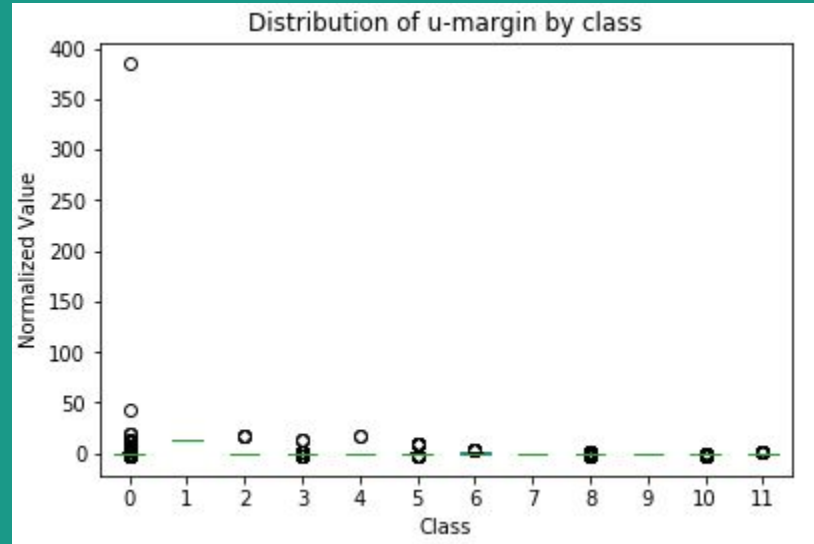
- Classification problem with 12 classes (copyists)
- 10 features collected from images, all continuous
- 20,867 data points, split evenly into training set and test set
- Data set comes from the UCI ML Repository:
<https://archive.ics.uci.edu/ml/datasets/Avila>



Fig. 3. The number of peaks in the pixel projection histogram on the horizontal axis for a row.

Preprocessing

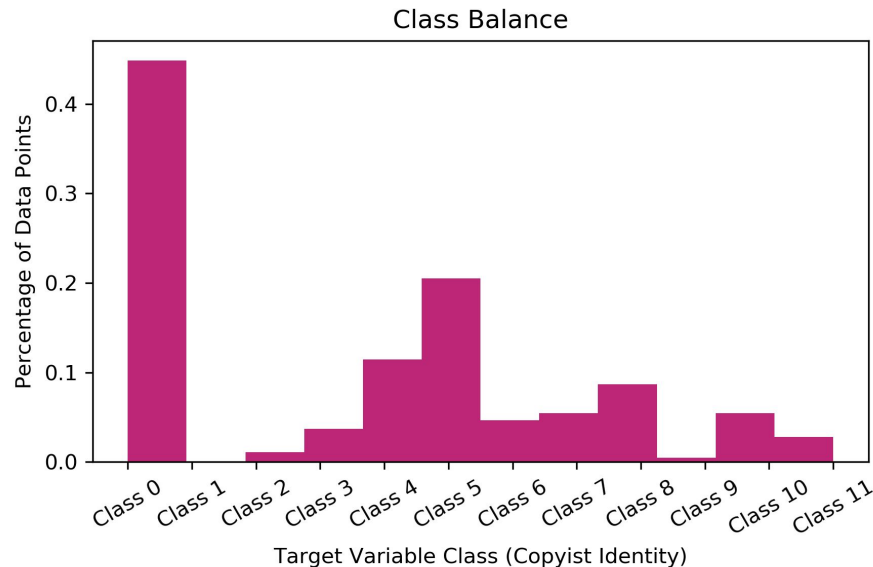
- All the features are continuous, they were normalized with Z-normalization method
- There were no missing values, no need for dropping data points, imputation, etc.
- Used LabelEncoder from sklearn to convert the target variable from letters A, B, C, D, E, F, G, H, I, W, X, Y to integers 0-11
- Discovered “broken” data point

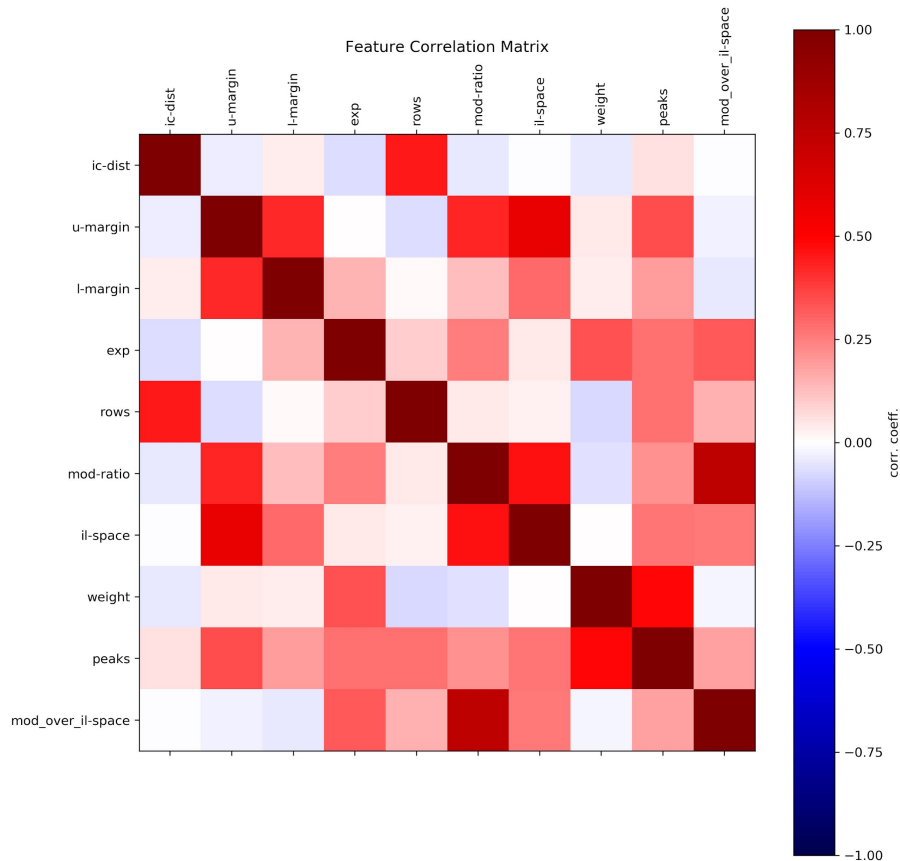


Index	ic-dist	u-margin	l-margin	exp	rows	mod-ratio	il-space	weight	peaks	rod_over_il-spac	class
6619	0	386	50	0.168104	0	53	83	0.275032	44	0.63802	0
9402	-3.4988	43.1337	-3.21053	-5.44012	-4.83284	-7.45026	-11.9355	13.1731	-5.48622	-6.71932	0

Exploratory Data Analysis, I: Class Balance

- Classification problem, so what is the balance?
- Definitely imbalanced!
 - 0 makes up ~41% of data points
 - Next biggest class is 5 at ~19%
- Might need to do additional preprocessing to even out distribution

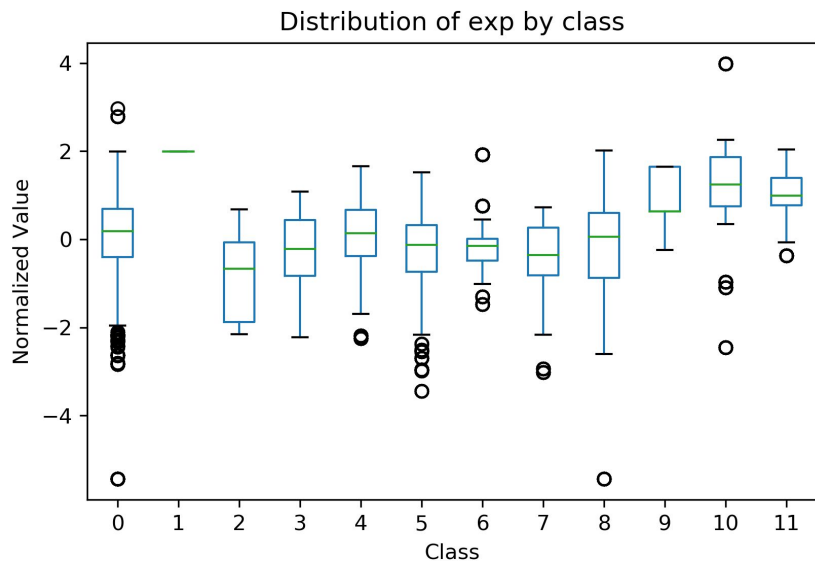




Exploratory Data Analysis, II: Correlation Matrix

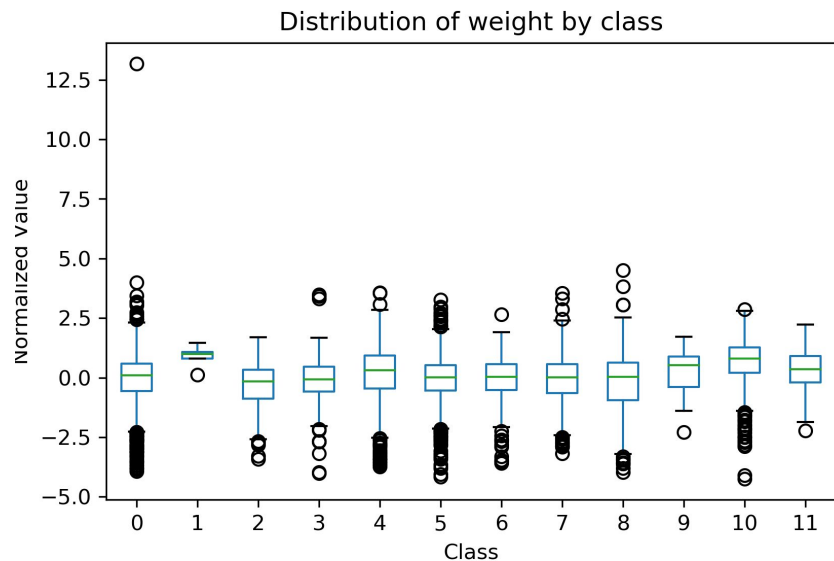
- Feature correlation matrix does not suggest any unexpected relationships
- Can you find the engineered feature?
 - Look at the feature names
- Interlinear space and upper margin
 - Similar, but distinct ideas

Exploratory Data Analysis, II: Boxplot by Class



- Exploitation is the density of script in a column
- Measures how much of the vertical space a copyist uses
- One of the most distinct ways to classify different handwriting

Exploratory Data Analysis, IV: Boxplot by Class



- Weight is the density of script in a row
- Measures how much of the horizontal space a copyist uses
- Used with exploitation



Exploratory Data Analysis, V: Boxplot by Class

- Peaks per row is simple but important
- Reliably distinct for different copyists
- Easily visualized
 - See pixel projection histogram below

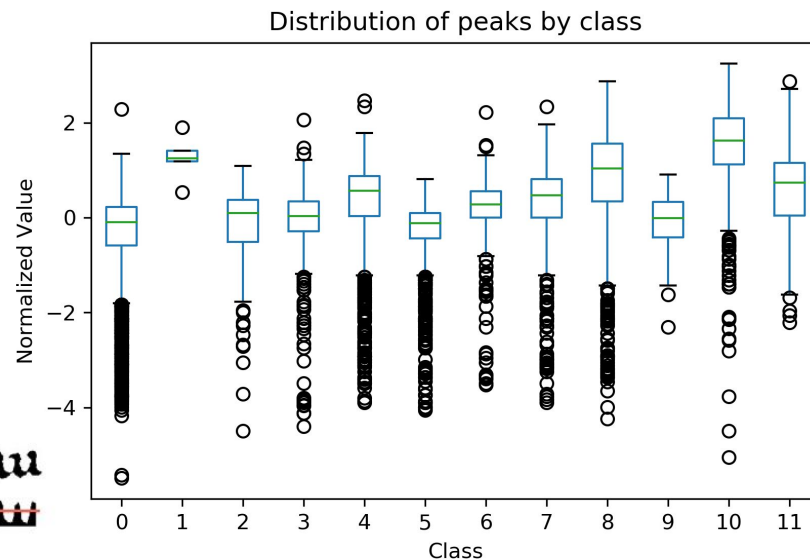
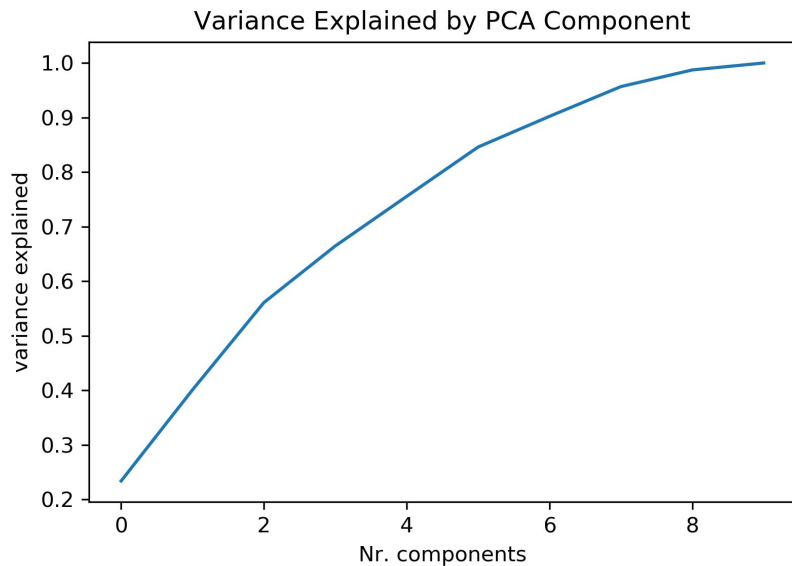


Fig. 3. The number of peaks in the pixel projection histogram on the horizontal axis for a row.

Exploratory Data Analysis, VI: PCA

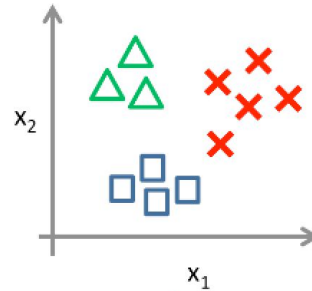





- Used PCA to investigate variance explained by features
- All of the features are useful, no “plateau” of explained variance
- Features selected by original researchers?
 - Also help from manual paleographers

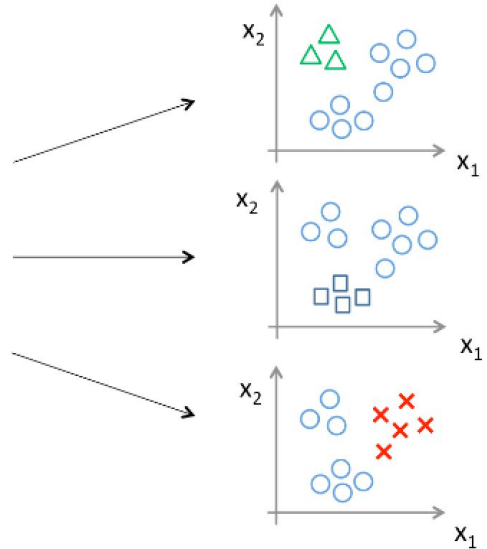
Next Steps

- Classification problem, so will be considering decision trees, SVM, logistic regression, etc.
- Need to bootstrap more/fewer samples to correct for imbalance
- One-vs-All iterative approach
 - Class 0 vs not Class 0 -> Class 5 vs not Class 5 (or Class 0) -> ...

One-vs-all (one-vs-rest):



Class 1: 
Class 2: 
Class 3: 





Thank you! Questions?

A Method for Scribe Distinction in Medieval Manuscripts Using Page Layout Features

Claudio De Stefano¹, Francesco Fontanella¹,
Marilena Maniaci², and Alessandra Scotto di Freca¹

¹ Dipartimento di Automazione, Elettromagnetismo, Ingegneria dell'Informazione e
Matematica Industriale

Via G. Di Biasio, 43 – 03043 Cassino (FR) – Italy

² Dipartimento di Filologia e Storia

Via Zamosch, 43 – 03043 Cassino (FR) – Italy

University of Cassino

{destefano,fontanella,mmaniaci,a.scotto}@unicas.it

Classification of handwriting in a medieval manuscript

Alexander Berry

Brown University

Tuesday, October 22, 2019

Data set, code, figures, and reports are available at:

<https://github.com/ABerry057/avila/>