An Introduction to Machine Learning Using Principal Component Analysis

Christina Branson

Toanl Nguyen

Abstract

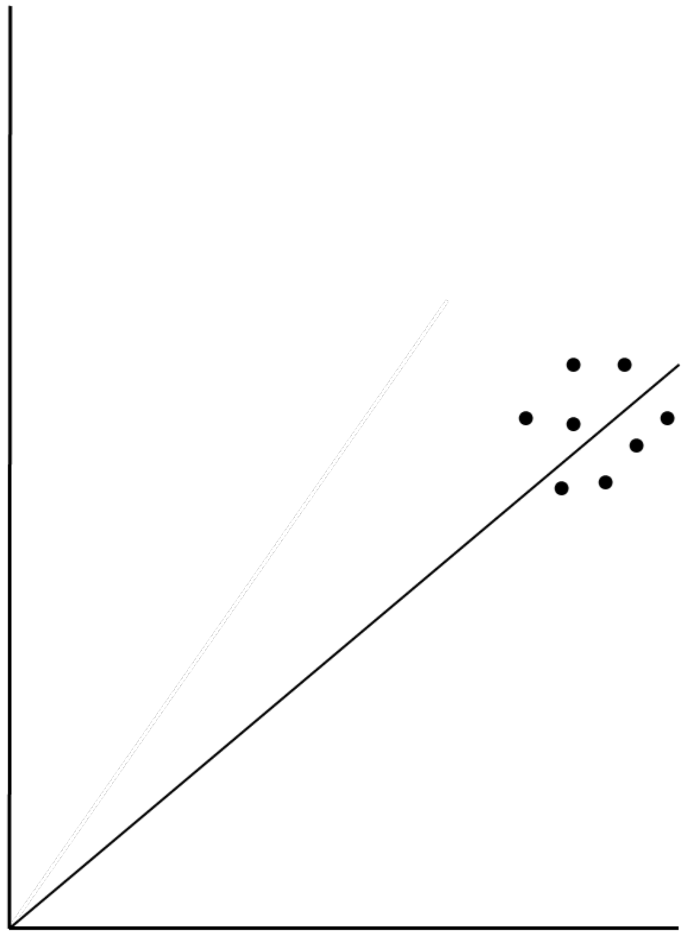
Humans are natural classifiers. From a young age, we learn without effort how to tell the difference between objects: cats versus dogs, 0's versus 1's. However, instructing a computer to classify similar objects can be challenging. A familiar example is the use of Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHAs) on the internet. These tests take advantage of a human's ability to easily read and reproduce a string of letters and/or numbers and a computer's struggle to do so.

Machine Learning is a subject of mathematics and computer science which aims to help computers "learn" how to classify objects. We will be exploring an elementary learning technique called Principal Component Analysis (PCA), a dimensionality reduction algorithm which transforms a set of vectors to a more representative basis. We will discuss the proof of the PCA Theorem and how it relates to machine learning, which will include a variety of topics from Linear Algebra including Schur's Theorem, basis transformation and orthogonal projections. Some applications of Principal Component Analysis will be presented, including our work in classifying hand-written digits.

Introduction To Machine Learning

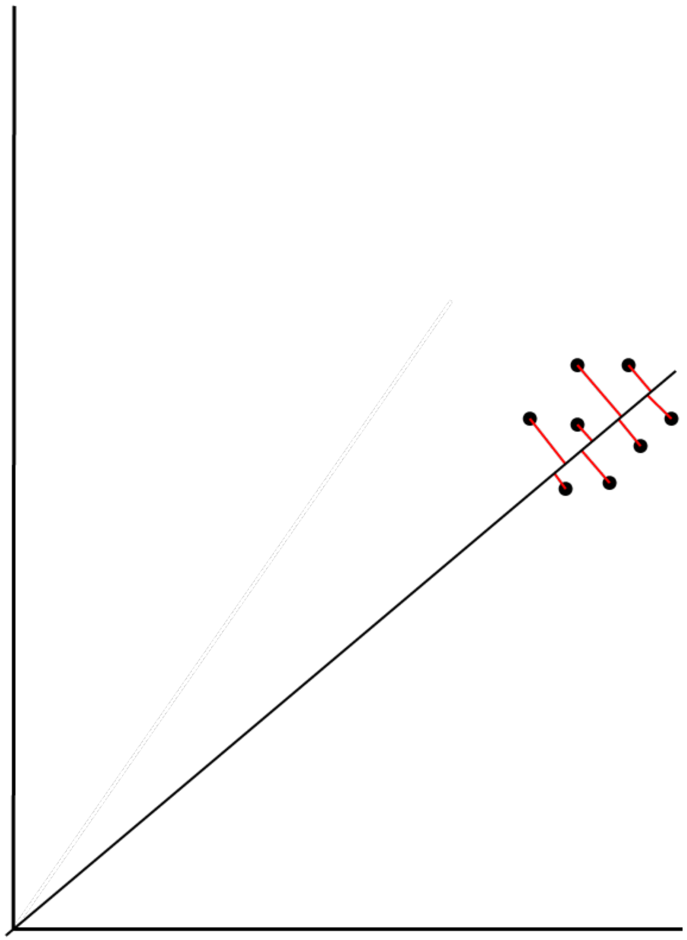
Machine learning deals with the algorithms and techniques used to help computers better classify objects. There is a huge variety of algorithms currently in use. One of these techniques, Principal Component Analysis (PCA) is a dimensionality reduction algorithm. The goal of PCA is to get project the data into a much lower-dimensional subspace that better represents the data.

Introduction To Machine Learning



- We have a set of vectors from a dataset X, and a subspace that is “close” to all the vectors.

Introduction To Machine Learning



- When a vector is not in the subspace, we can still come close to representing it by some vector in the subspace, Px. We call |x-Px| the error.
- Our goal is to find the subspace that minimizes the sum of the errors.

Preliminaries

Let $\{x_i\}_{i=1}^n$ be a collection of vectors in $\mathbb{R}^m$
For a subspace $W \subset \mathbb{R}^m$, $P_W x_i$ is the orthogonal projection of x_i into W.
Define

$$E(W) = \Sigma_{i=1}^n ||x_i - P_W x_i||^2$$

Theorem

Let $\{x_i\}_{i=1}^{n}$ *be a collection of vectors in* $\mathbb{R}^m$*,*

$$A = \left[x_1|x_2|...|x_n\right]$$

Let $1 \leq p < m$*. Let* $\{u_1, ...u_p\}$ *represent an orthonormal collection of vectors and* $U = span\{u_1, ..., u_p\}$*. If*

$$E(U) = min\{E(W)|W \text{ } p\text{-dimensional subspace of } \mathbb{R}^m\}$$

Then $\{u_1, ...u_p\}$ *are the eigenvectors of* AA^* *that correspond to the* p *largest eigenvalues of* AA^**.*

Proof

Let W be a p-dimensional subspace of $\mathbb{R}^m$ with orthonormal basis $\{w_1, ..., w_p\}$. Then $P_W x_i = \Sigma_{j=1}^{p} \langle x_i, w_j \rangle w_j$ yields

$$||P_W x_i||^2 = \Sigma_{j=1}^{p} \langle x_i, w_j \rangle^2$$

Since

$$\begin{aligned} ||x_i - P_W x_i|| =& \langle x_i - P_W x_i, x_i - P_W x_i \rangle \\ =& ||x_i||^2 - 2\langle x_i, P_W x_i \rangle + ||P_W x_i||^2 \\ =& ||x_i||^2 - ||P_W x_i||^2 \end{aligned}$$

We may write

$$\begin{aligned} E(W) &= \Sigma_{i=1}^{n} ||x_i - P_W x_i||^2 \\ &= \Sigma_{i=1}^{n} ||x_i||^2 - \Sigma_{i=1}^{n} ||P_W x_i||^2 \\ &= \Sigma_{i=1}^{n} ||x_i||^2 - \Sigma_{j=1}^{p} w_j^* A A^* w_j \end{aligned}$$

Since $\Sigma_{i=1}^{n} ||x_i||^2$ is fixed, minimizing $E(W)$ is achieved by maximizing $\Sigma_{j=1}^{p} w_j^* A A^* w_j$

Let $\alpha(w_1, ..., w_p) = \Sigma_{j=1}^{p} w_j^* AA^* w_j$. Then α is a polynomial in the components of all the w_js and we can maximize it using Lagrange multipliers with orthonormality constraints. Let us begin with w_1.

$$L = w_1^* AA^* w_1 - \lambda_1(w_k^* w_k - 1)$$

Differentiating wrt to w_1,

$$\begin{aligned} 0 =& AA^* w_1 - \lambda_1 w_1 \\ \Leftrightarrow AA^* =& \lambda_1 w_1 \\ \Leftrightarrow & \ w_1 \text{ is an eigenvector of } AA^* \end{aligned}$$

Noticed we just maximized $w_1^* AA^* w_1 = w_1^* \lambda_1 w_1 = \lambda_1 w_1^* w_1 = \lambda_1$

Next, take

$$L = w_2^* AA^* w_2 - \lambda_2(w_k^* w_k - 1) - \beta(w_2^* w_1 - 0)$$

Similarly, we differentiate wrt w_2:

$$\begin{aligned}
0 =& AA^* w_2 - \lambda_2 w_2 - \beta w_1 \\
=& w_1 AA^* w_2 - w_1 \lambda_2 w_2 - w_1 \beta w_1 \\
=& w_1^* AA^* w_2 - w_1^* \lambda_2 w_2 - w_1^* \beta w_1 \\
=& (AA^* w_1)^* w_2 - \lambda_2 w_1^* w_2 - \beta w_1^* w_1 \\
=& \lambda_1 w_1^* w_2 - 0 - \beta \\
=& \beta \\
\Rightarrow& AA^* w_2 = \lambda_2 w_2
\end{aligned}$$

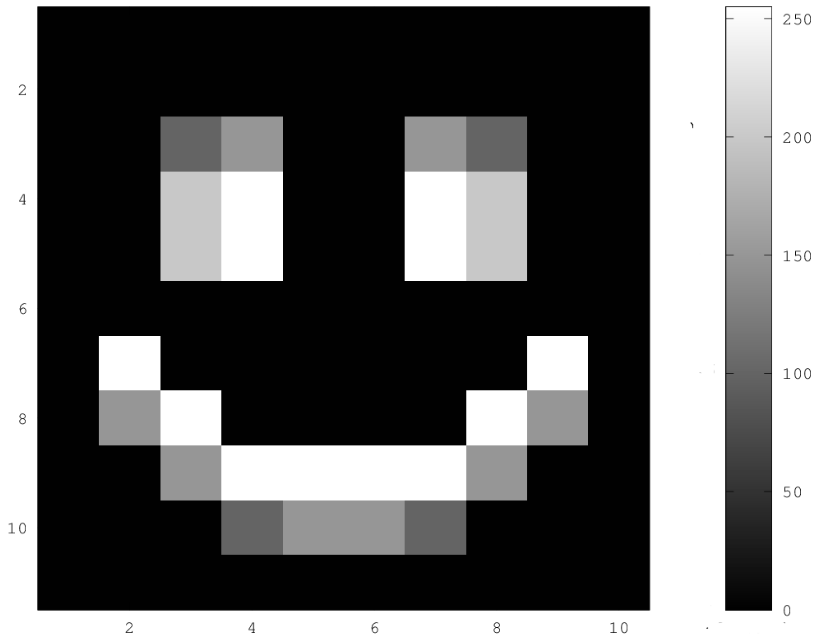
And we continue this inductively.

Returning to $E(W)$, we now have

$$\begin{aligned} E(W) =& \Sigma_{i=1}^{n} ||x_i||^2 - \Sigma_{j=1}^{p} w_j^* A A^* w_j \\ =& \Sigma_{i=1}^{n} ||x_i||^2 - \Sigma_{j=1}^{p} \lambda_j \end{aligned}$$

which is maximized by choosing the p largest eigenvalues of AA^*.

Applications



Images can be thought of as an object in mathematical space.

For example, the image seen here of size 10px by 10px can be represented by a computer as a matrix in M_10x10(R) where each element is an integer between 0 and 255.

By restructuring the data, we can also express it as as an element of R100.

0	0	0	0	0	0	0	0	0	0
0	0	100	150	0	0	150	100	0	0
0	0	200	255	0	0	255	200	0	0
0	0	200	255	0	0	255	200	0	0
0	0	0	0	0	0	0	0	0	0
0	255	0	0	0	0	0	0	255	0
0	150	255	0	0	0	0	255	150	0
0	0	150	255	255	255	255	150	0	0
0	0	0	100	150	150	100	0	0	0
0	0	0	0	0	0	0	0	0	0

The images we'll be looking At today are 28px by 28px scans of handwitten digits. We'll be treating them as vectors in R784.

Application: Visualization

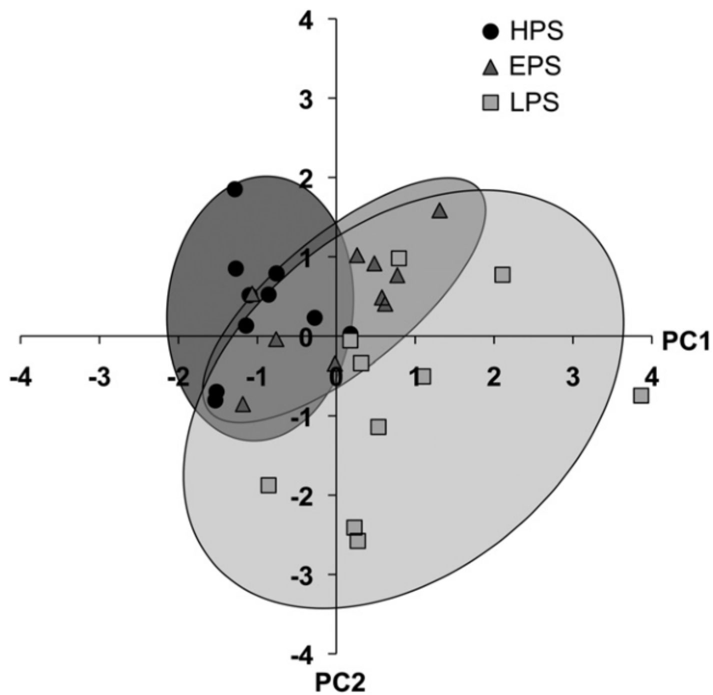


FIGURE 4 Result of principal component analysis of metabolic pools for *D. melanogaster* raised on the HPS, EPS, and LPS diets. Graph of principal component 1 (PC1) and 2 (PC2) with density ellipses (0.90) for each diet. PC1 and PC2 explain 46.9 and 37.6% of the variance, respectively.

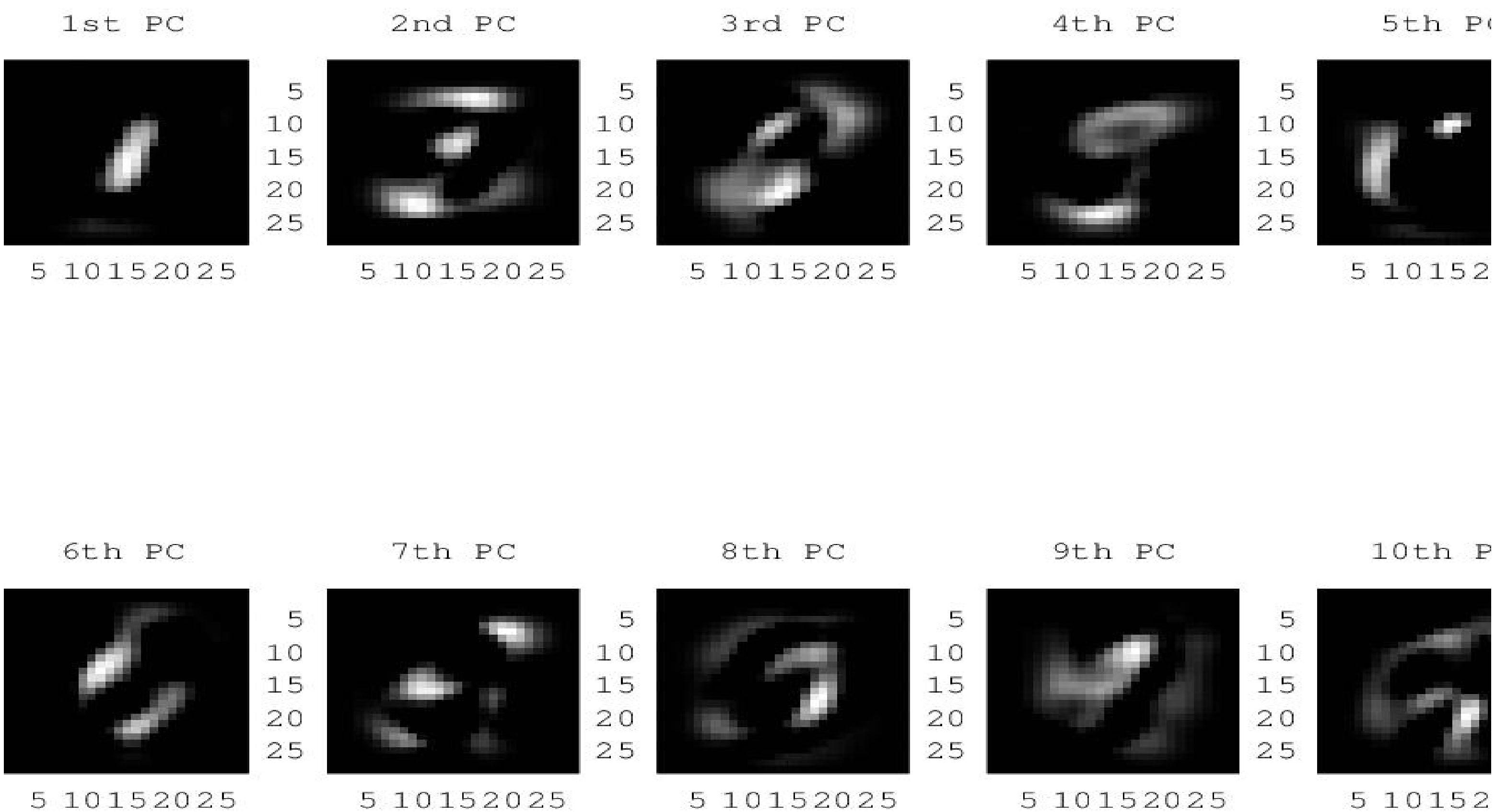
- Brought to me last week by a student needing to make her own figure.

Dietary Protein and Sugar Differentially Affect Development and Metabolic Pools in Ecologically Diverse Drosophila, Matzkin, et al, 2011

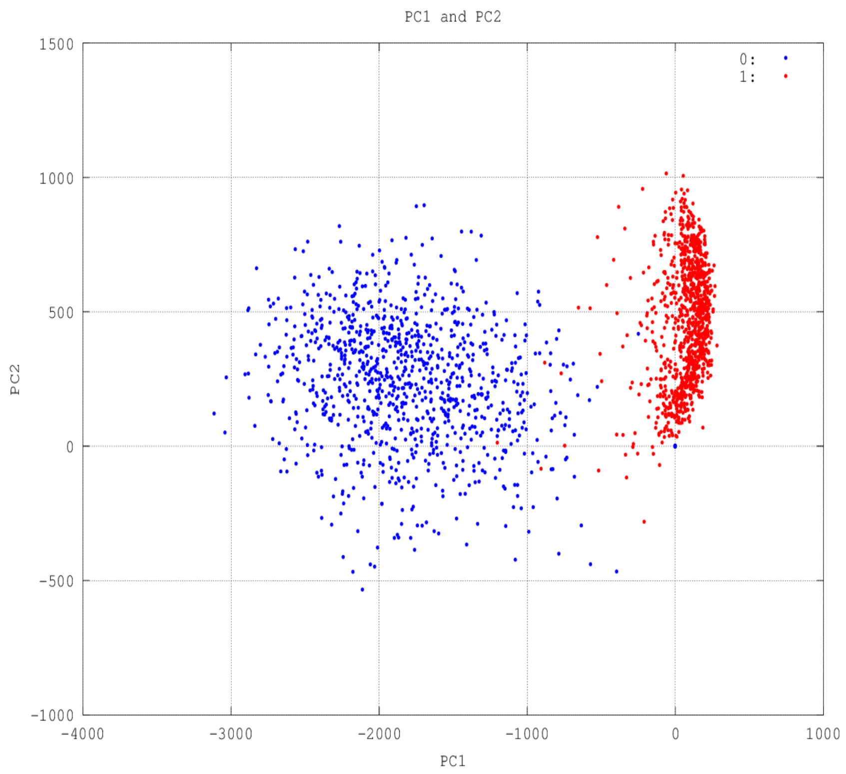
Application: Visualization

- Because of the dimensionality reduction inherent to PCA, it makes it a good tool for visualization.
- In our case, we’re looking at 784 dimensional data. With our human brains and years of practice with looking at numbers, we can differentiate between data easily. Not so with computer.
- With PCA however, we can pretty easily project our 784 dimensional data into 15 dimensions while still maintaining integrity.

Application: Visualization



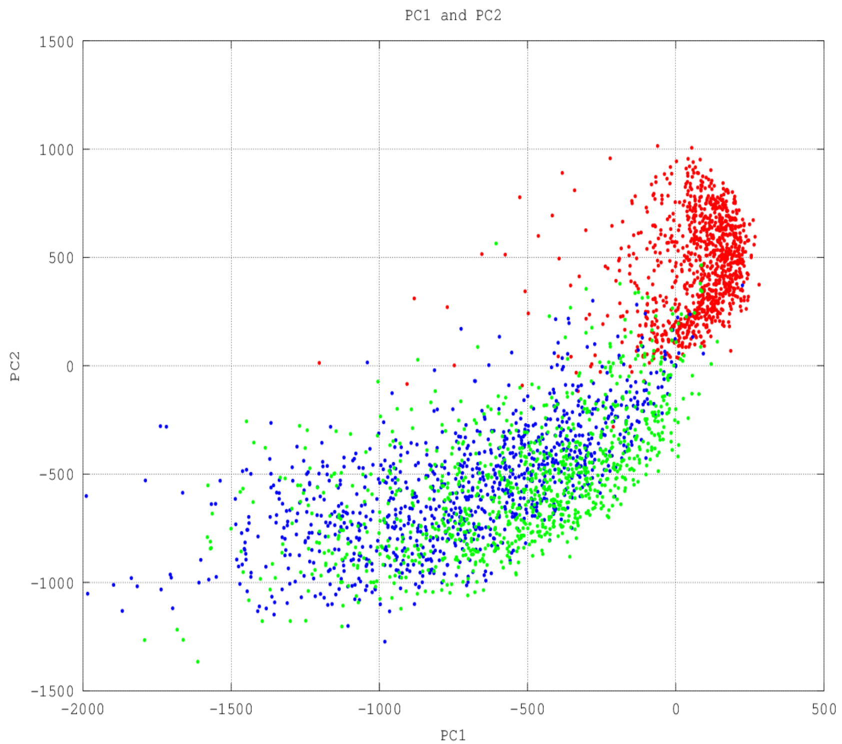
Application: Visualization



This graph shows the 1s and 0s projected into the two-dimensional space defined by the first two principal components.

The two digits are quite separable. The ones mostly have a small positive coefficient for PC1, whereas the zeroes have a large negative coefficient for PC1.

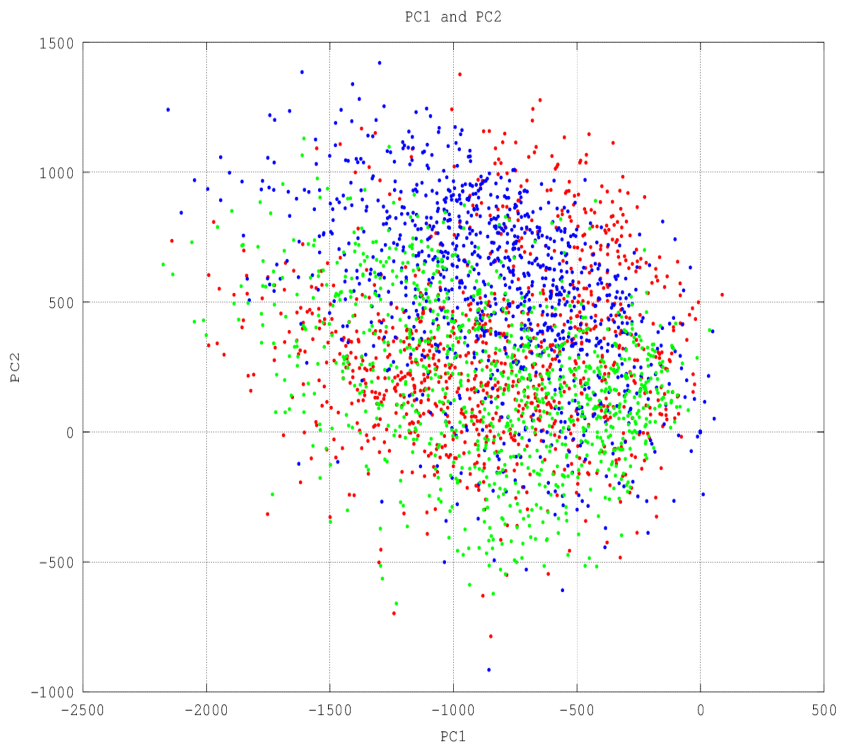
Application: Visualization



This graph shows 1s, 4s and 7s. They're grouped close to each other, due to their common features: the strong vertical line. The 4s and 7s also live below the x-axis and left of the y-axis.

Despite these promising connections, the 4s and 7s don't distinguish between each other well.

Application: Visualization

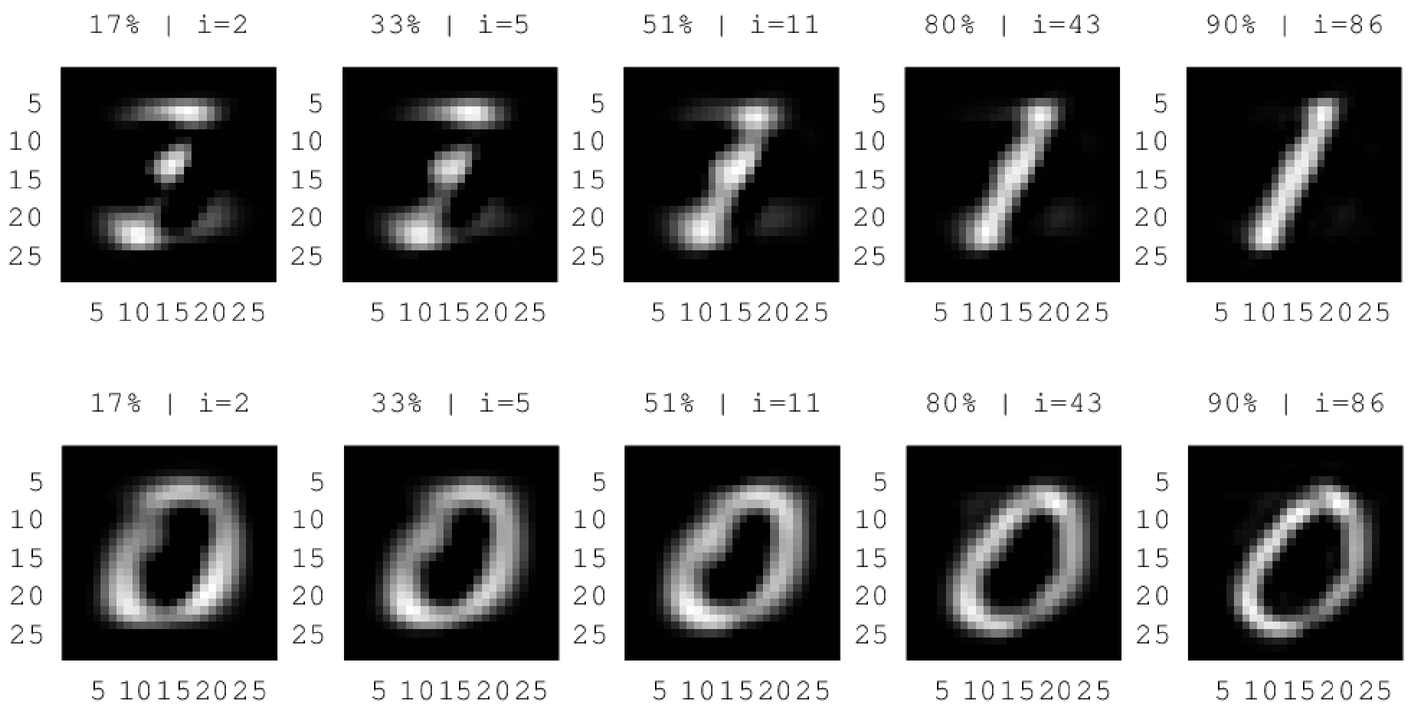


And then the 2s, 3s, and 5s. There's really not a lot that we can do to tell these apart in only two dimensions.

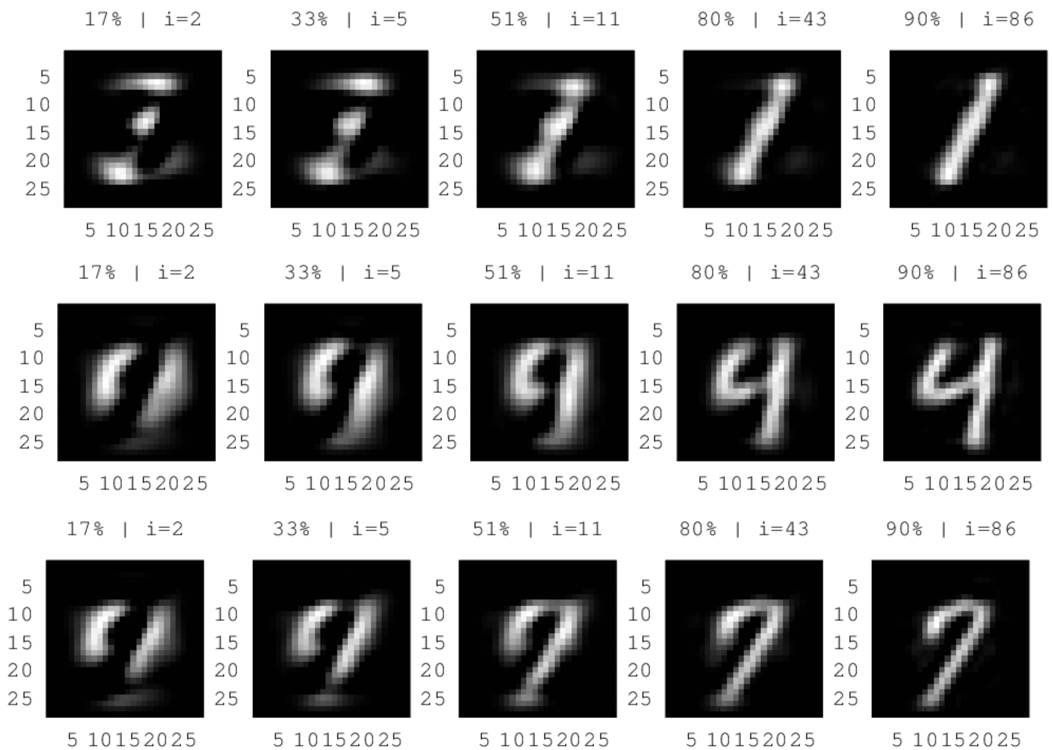
Why do some digits work better than others?

Application: Visualization

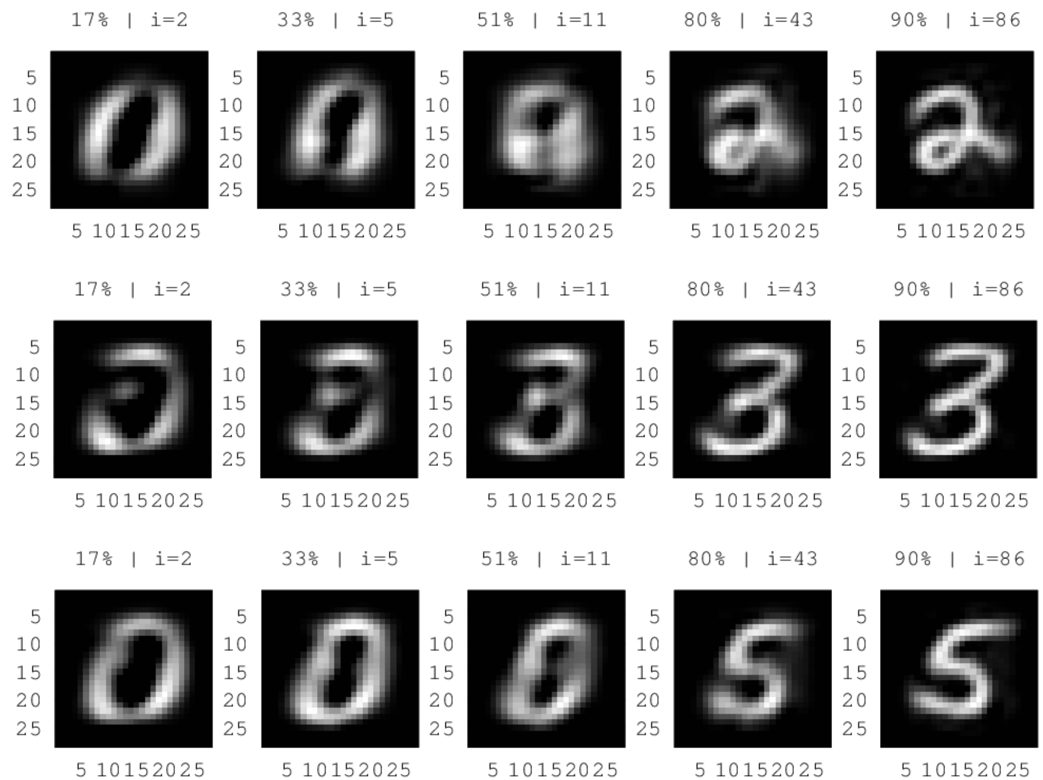
These images show the projections of a sample digit 1 and a sample digit 2. The left-most pictures show the projections with only 2 principal components. Unlike the graph from the biology paper in which the weights of the first two eigenvalues was 80%. It's only at 17% for the first two. In order to reach the 80% mark, we need to represent the data with the first 43 principal components.



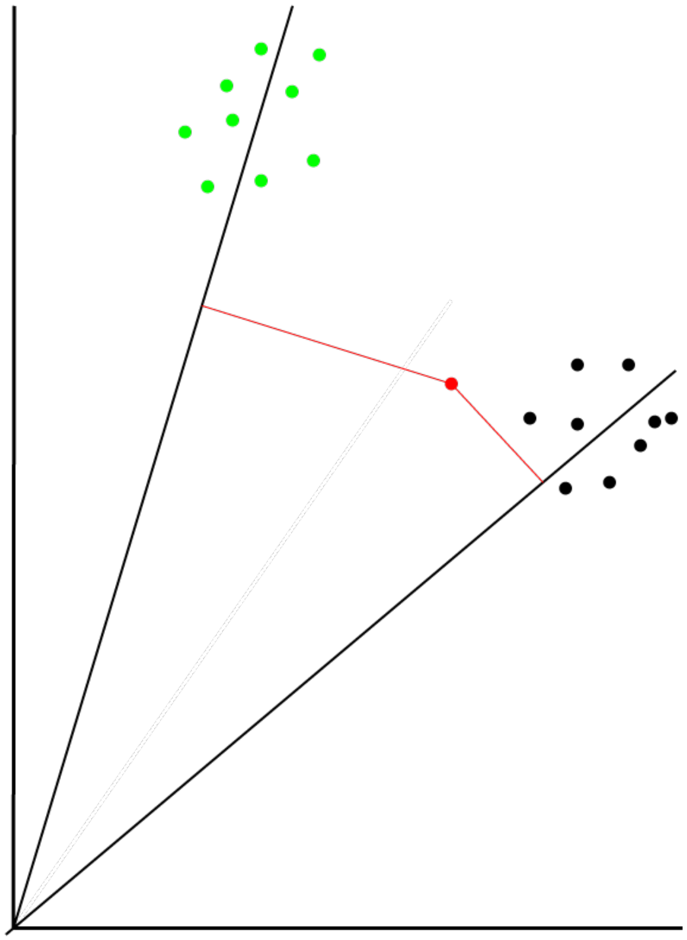
Application: Visualization



Application: Visualization



Application: Classification

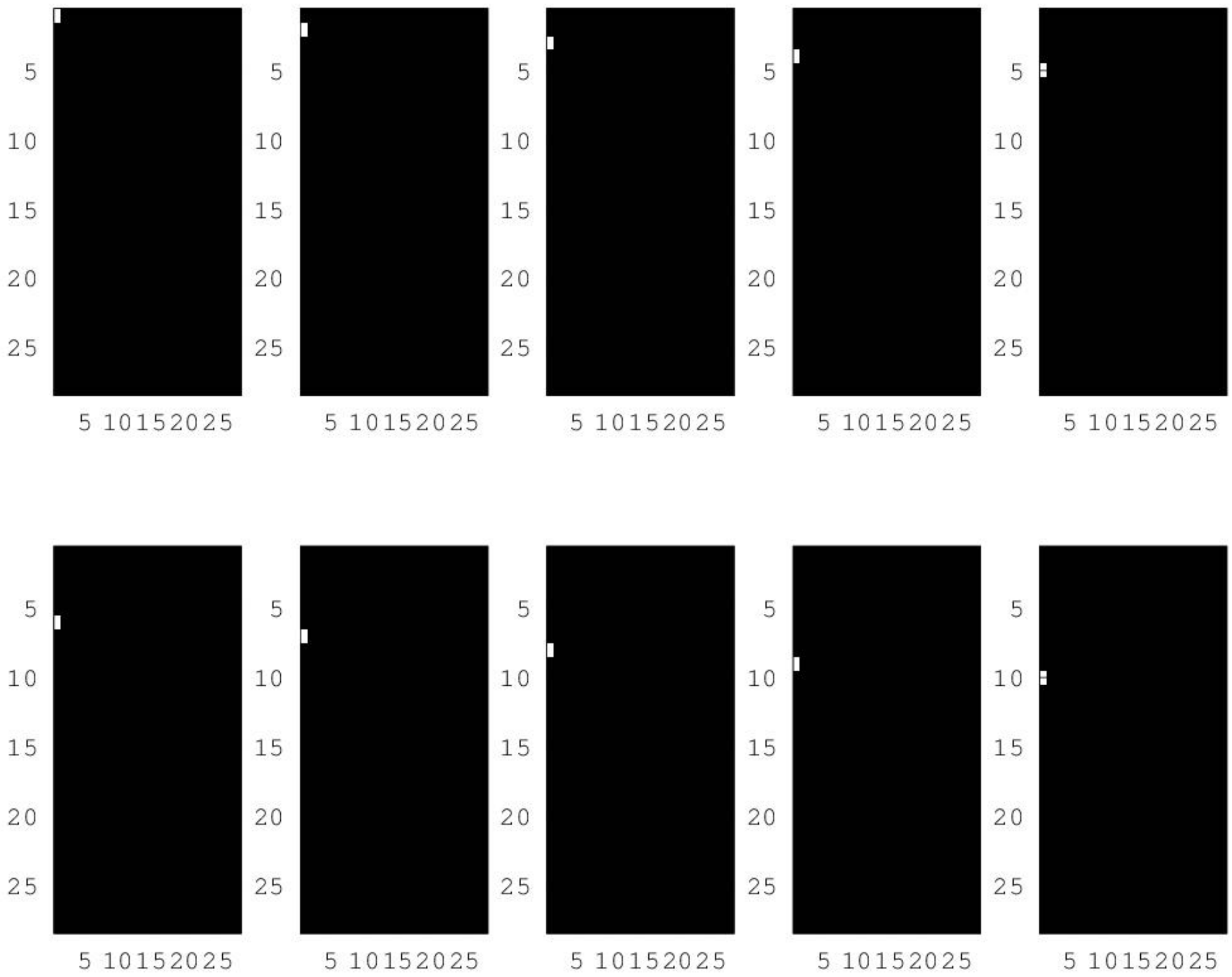


- Now suppose we have another dataset Y with its own defining subspace.
- We wish to classify the new vector, a.
- We would say that the

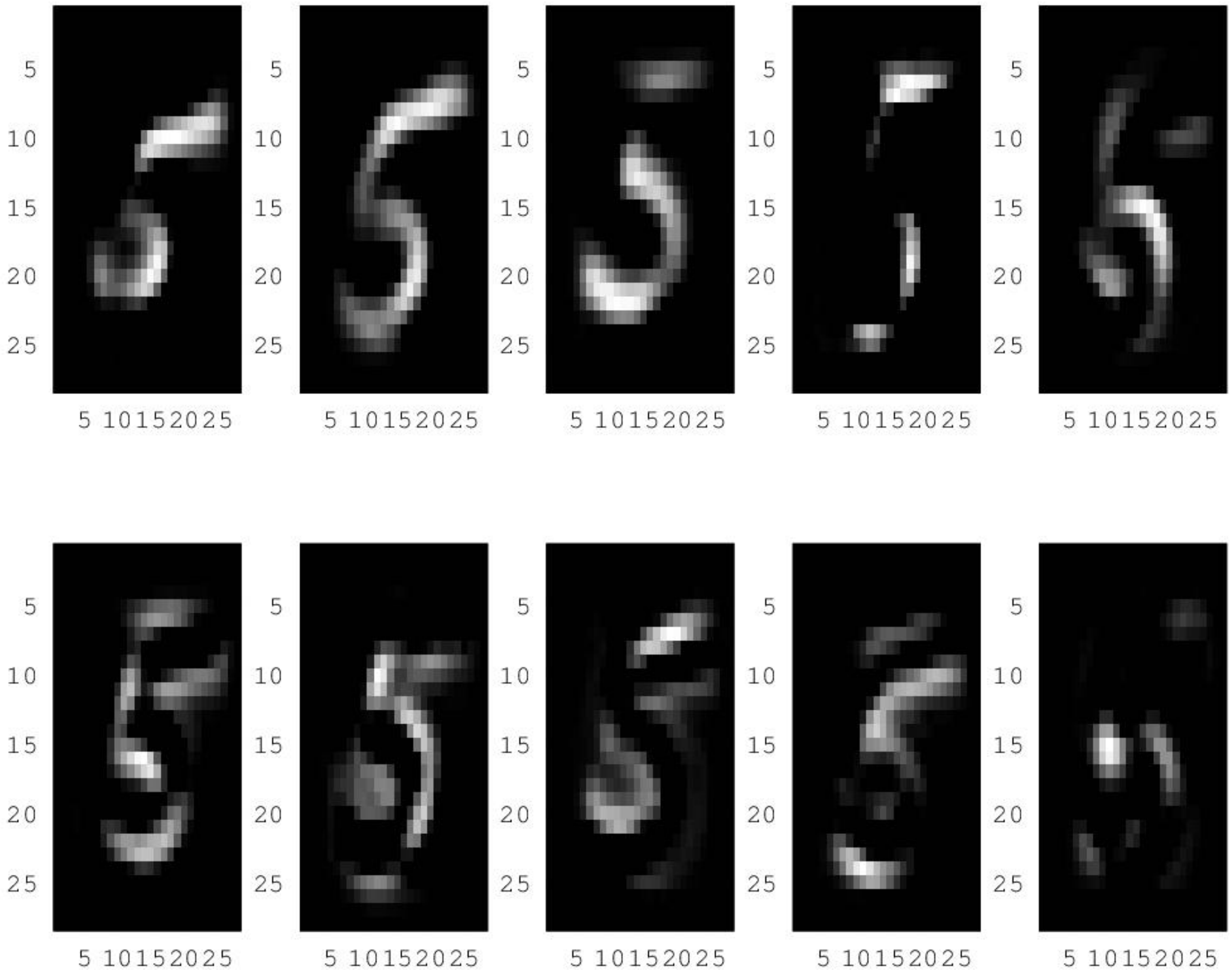
  $e_X = |P_X a - a|$ and

  $e_Y = |P_Y a - a|$ and classify the new vector as type X since $e_X < e_Y$

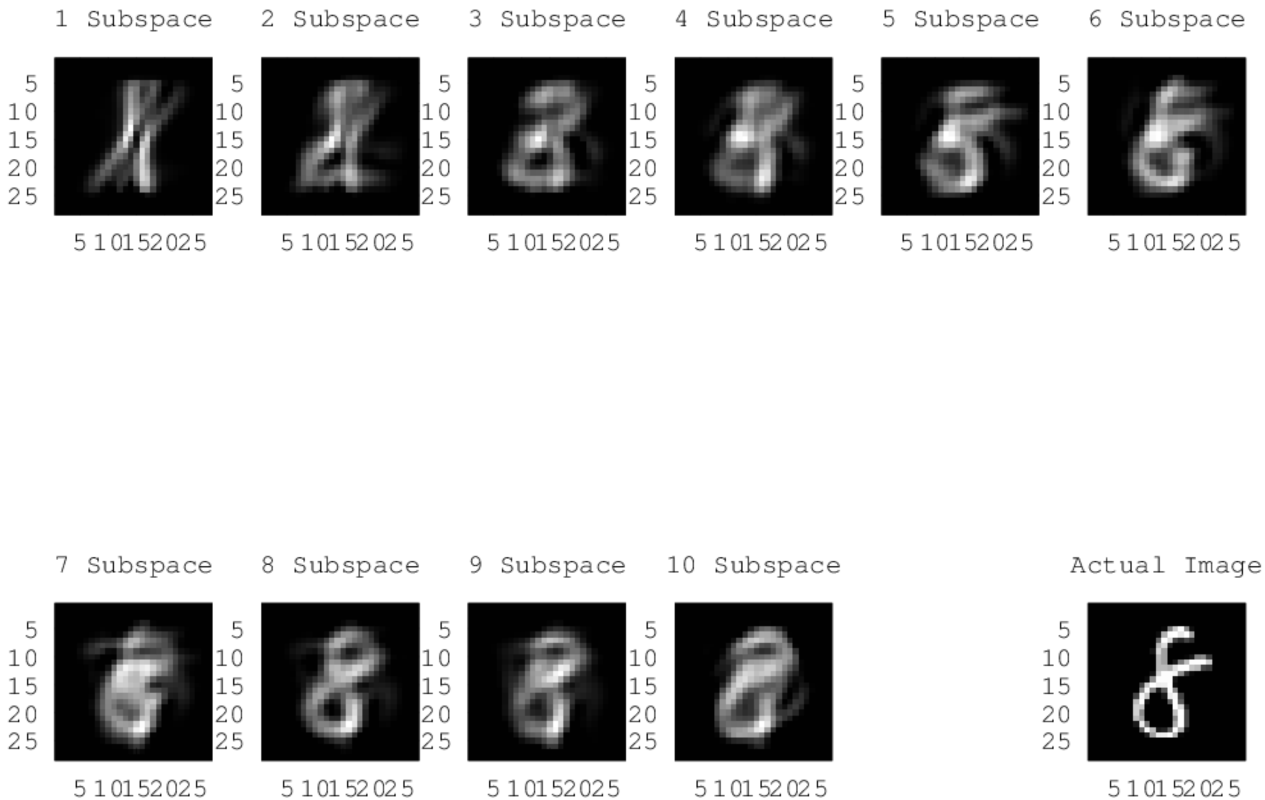
Application: Classification



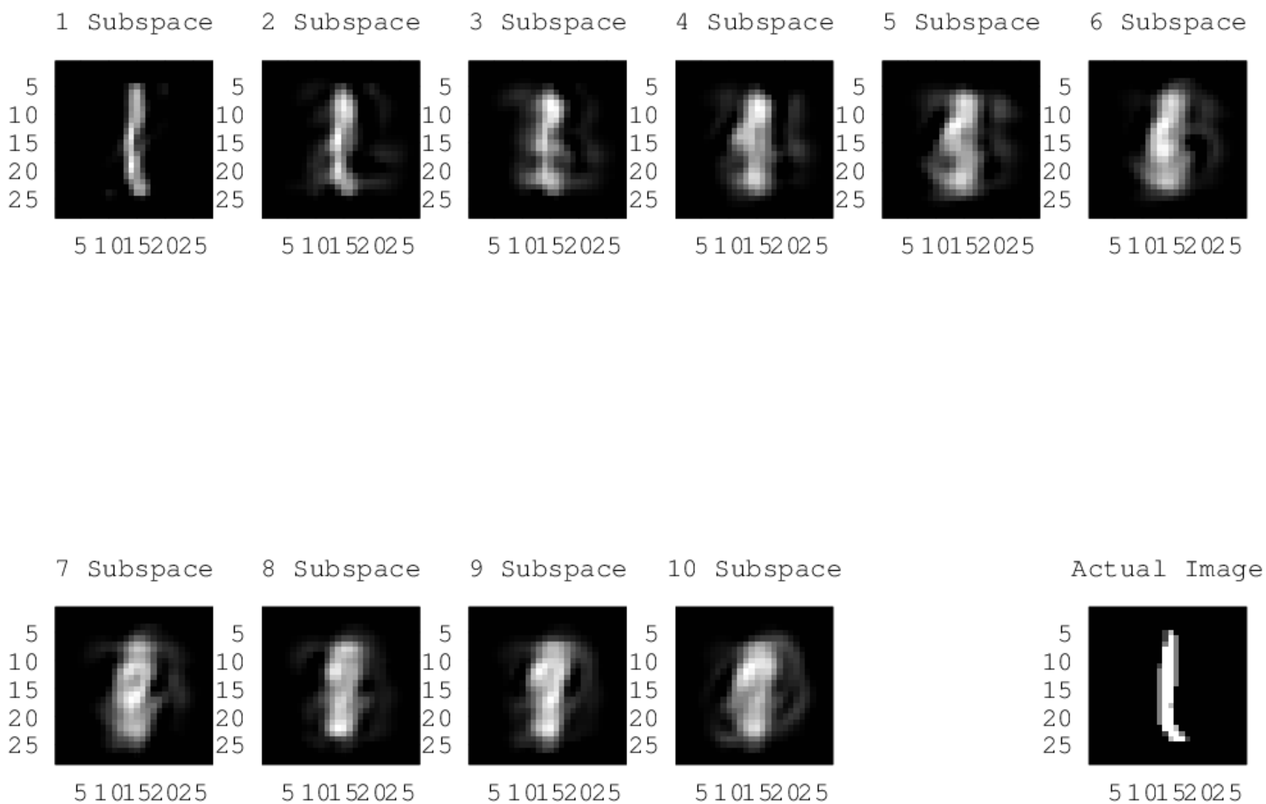
Application: Classification



Application: Classification

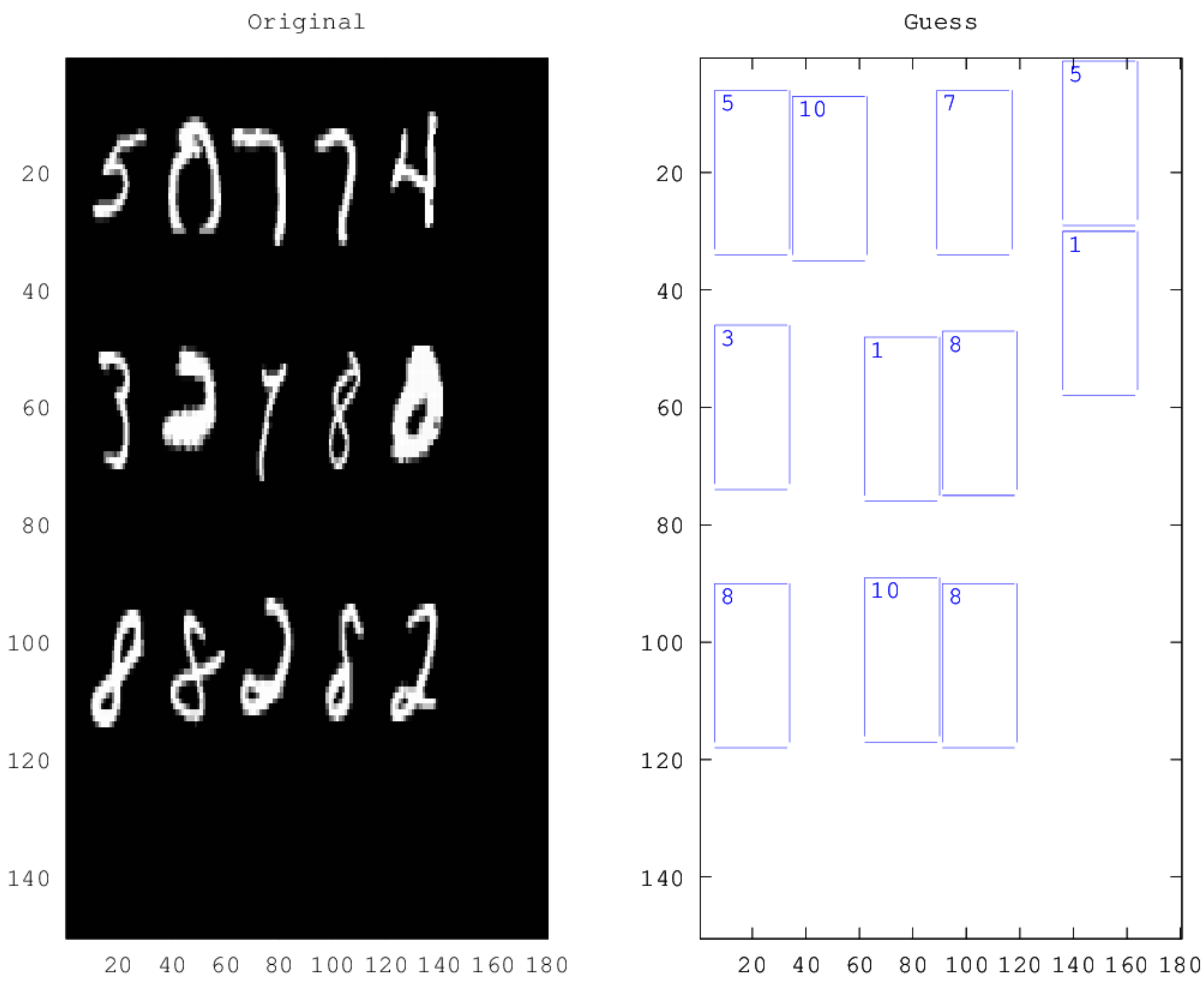


Application: Classification



Application: Classification

Application: Classification



Future Exploration

- Different inner products
- Removal of outliers in the training set
- Better classification algorithms (Support Vector Machine, Relevance Vector Machine, etc.)
- Application to weather data
- Winning at the internet by collecting all available images of cats.

An Introduction to Machine Learning Using Principal Component Analysis

Christina Branson

Toanl Nguyen