Review of Material for Mid-Term

- All assessment points should be posted to moodle now (except latest quiz)
- Mid-Term Examination to be held on 11 March
- Coverage will be cumulative since the beginning of the semester
- Reviewing now because of disrupted schedule next week

Supervised and Unsupervised Learning

- Supervised Learning
  - LDA
  - Support Vector Machines
  - Nearest Neighbor
  - Regression
  - Regularization
- Unsupervised Learning
  - PCA
  - other matrix factorizations
  - Hierarchical and *k*-means clustering
- Cross-Validation
- Probability and Linear Algebra

Covariance and Sample Covariance

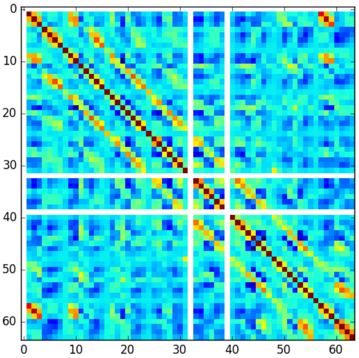
- Many simple data analysis problems involve finding the eigenvectors of the sample covariance matrix given the data
- Test for relationships between variables/features
- PCA: Gives a natural basis in which to express the data

$$\mathbf{R} = E[\mathbf{x}^T\mathbf{x}] \tag{1}$$

$$R = \frac{1}{n-1}\sum_{i=1}^{N}(x_i - \mu)^T(x_i - \mu) \tag{2}$$

$$\frac{1}{n-1}\mathbf{X}^T\mathbf{X} \tag{3}$$

Visualizing the Correlation or Covariance

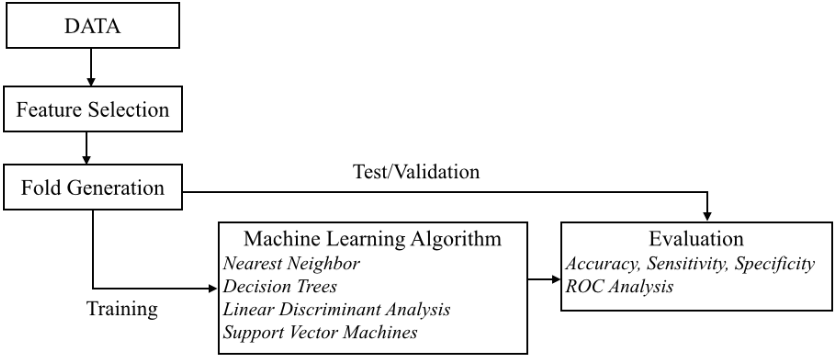


- Rendering the Correlation as an image immediately reveals which variables are related

Overview of Typical Machine Learning Workflow

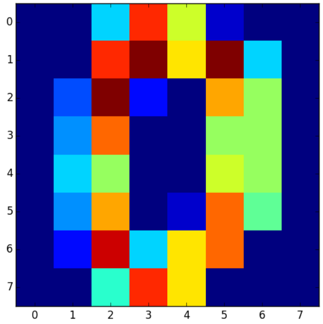
Box 1

A (supervised) ML workflow will minimally contain: *data*,*feature selection process*, *cross-validation*,*an ML algorithm*, and a means to *evaluate* performance



Working with the zipcode "digits" dataset

- Dataset of 8×8 images of hand-written characters



Linear Discriminant Analysis

- Linear Discriminant Analysis (LDA) is a supervised learning technique that attempts to provide a one-dimensional projection of the data such that discrimination between classes is maximized
- As in previous examples $\{x_1, x_2, \ldots, x_n\}$ we have a set of observations where each observation belongs to one of a finite number of classes

LDA

- Attempt to find some *linear* function of the values of feature values such that we can use the output to classify new data instances
- Which coefficients w maximize separability?

$$\phi = w^T x \tag{4}$$

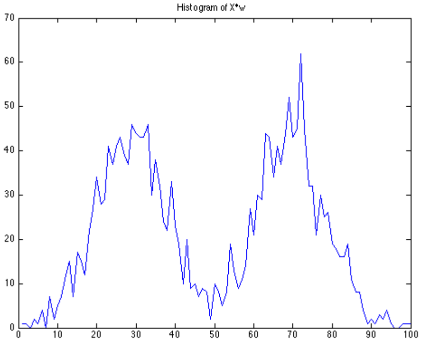
Performance Metrics

- In any binary classification task two types of error can occur, false positives and false negatives
- In general we need to keep track of both of these errors to understand how well our classifier is performing
- The *Receiver Operating Characteristic* (ROC) keeps track of both of these error rates as we vary the threshold

$$\text{threshold} > w^T x \tag{5}$$

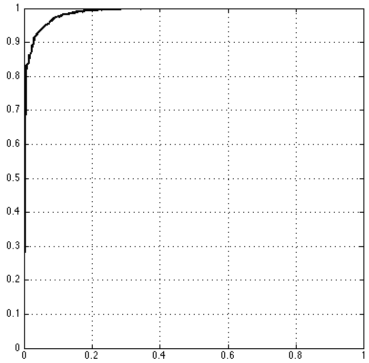
Receiver Operating Characteristic

- Imagine we look at a histogram of our test statistic
- Our threshold will fall somewhere along the x-axis and will determine the two error rates



Example

- Each point along this curve corresponds to a single value of the threshold parameter



LDA recipe

- The LDA solution can be found explicitly in terms of the data by differentiating the objective with respect to w and setting the result equal to 0

$$\frac{d}{dw}J(w) = 0 \tag{6}$$

- When we work this out (homework) the solution we arrive at is equal to the following

$$\hat{w} = S_W^{-1}(\mu_1 - \mu_2) \tag{7}$$

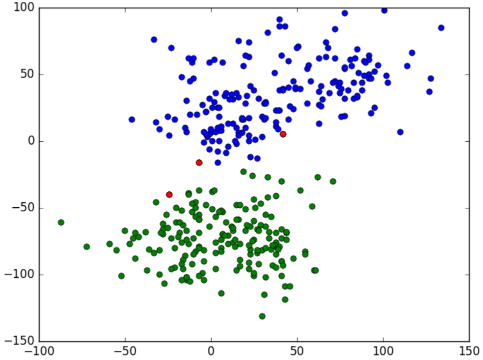
Defining a Symmetry Feature

```
def symlr(t):
    s = np.fliplr(t)
    y = t - s
    val = np.sum(y[:, 0:3])
    return val


def symud(t):
    s = np.flipud(t)
    y = t - s
    val = np.sum(y[0:3, :])
    return val
```

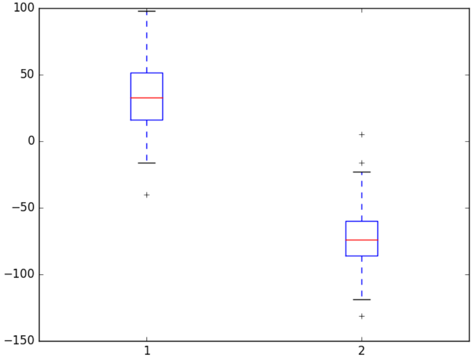
- The above functions calculate a symmetry measure on our 8 by 8 images

Symmetry features for digits 5 and 6



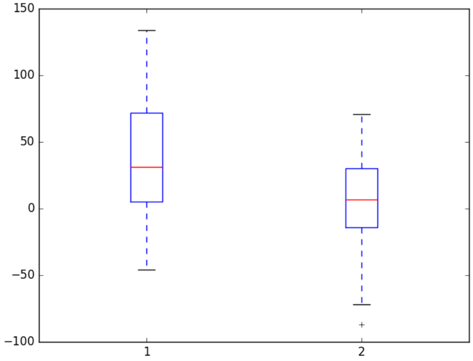
- Are both features equally discriminative?

Discriminative Feature



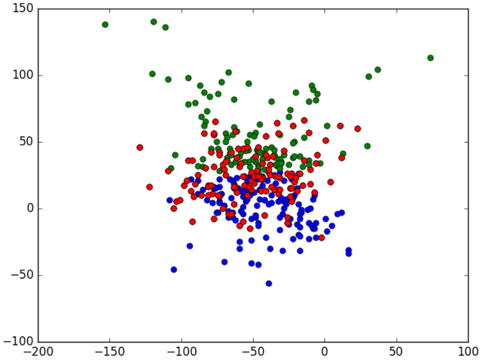
- Box-Plot Shows clear difference between the two classes

Non-discriminative Feature



- Still significant difference, but classes are overlapping

Same result for digits 3 and 9



- Less discriminant power in these features for a different digits

Solution

- Solve constrained optimization with lagrange multipliers

$$L_P = (1/2)||\beta||^2 - \sum \alpha_i[y_i(x_i^T\beta + \beta_0) - 1] \tag{8}$$

$$\beta = \sum \alpha_i y_i x_i \tag{9}$$

$$0 = \sum \alpha_i y_i \tag{10}$$

$$f(x) = \sum_{i=1}^{N} \alpha_i y_i (x_i, x) + \beta_0 \tag{11}$$

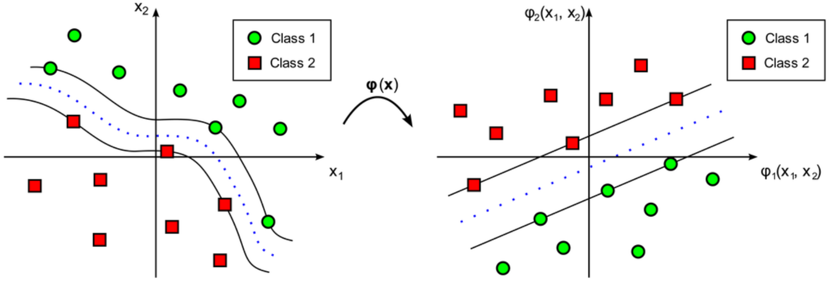
The Classification Rule

- Usage of the kernel trick results in the following classification rule
- The kernel function is selected in advance the algorithm finds the weights as well as the support vectors that maximize the margin

$$\sum_{s \in \text{support vectors}} K(x_{test}, x_s) + \text{bias term} > 0 \tag{12}$$

Non-linear Mapping to a new Feature Space

- Non-linear boundary in the original feature space becomes non-linear in the new feature space



(a) (b)

Common Kernels

- The polynomial, radial basis function, linear and neural net are the most common kernels

$$K(x, y) = (xy^T + 1)^p \tag{13}$$

$$K(x, y) = e^{-||x-y||^2/2\sigma^2} \tag{14}$$

$$K(x, y) = \tanh(kxy^T - \delta) \tag{15}$$

Higher Dimensional Feature Space, Polynomial

- Higher dimensional feature space based on polynomials would look something like the following

$$\mathbf{z} = [1, x_1, x_2, x_1 x_2, x_1^2, x_2^2] \tag{16}$$

- Vectors in the transformed feature space are then 6 dimensional $\mathbf{z} = [z_1, \ldots, z_6]$
- The advantage of SVM is that it allows to work in the higher dimensional feature space without having to actually compute the *vectors* in the higher dimensional space
- What is the kernel function for the z space in terms of x?

Non-linear Support Vector Machines

- Describe the problem and solution in the z space

$$L = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j z_i^T z_j \tag{17}$$

$$f(x) = \operatorname{sgn}(w^T z + b) \tag{18}$$

$$w = \sum_{z_i \in \mathrm{SV}} \alpha_i y_i z_i \tag{19}$$

$$y_i(w^T z_i + b) = 1 \tag{20}$$

- All the elements of the problem and solution can be stated in terms of the inner product alone

Polynomial Kernel

- Recall the Polynomial Kernel Given above, lets expand it for $d = 2$ and a two dimensional input space

$$K(x, y) = (1 + x^T y)^d$$

$$K(x, y) = (1 + x^T y)^2$$

$$K(x, y) = (1 + x_1 y_1 + x_2 y_2)^2$$

$$K(x, y) = 1 + 2x_1 y_1 + 2x_2 y_2 + (x_1 y_1)^2 + (x_2 y_2)^2 + 2x_1 y_1 x_2 y_2$$

Establishing the Existence of the Transformed Space

- Recall that the support vector machine relied on an implicit transformation to a new space
- Since we do not actually compute the feature vectors in the new space we can seemingly choose *any* function of two arguments for the kernel

$$K(x, y) = (\phi(x), \phi(y))$$

- In certain cases the existence of the transformed space is guaranteed by Mercer's Theorem

The Gram Matrix

- The Gram Matrix is a matrix of inner products of all the different pairs of data points

$$K = \begin{bmatrix} K(x_1, x_1) & K(x_1, x_2) & \cdots \\ K(x_2, x_1) & K(x_2, x_2) & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \tag{21}$$

- Can also be expressed in terms of the data matrix

$$K = XX^T$$

Mercer's Theorem

- Mercer's Theorem says that the gram matrix should be *positive semidefinite*
- Assumption: $K(x, y)$ is a symmetric function
- Theorem: $K(x, y)$ can be expressed as an inner product $K(x, y) = (\phi(x), \phi(y))$ if the gram matrix is positive semidefinite for any possible collection of vectors $\{x_1, x_2, \cdots, x_N\}$

Characteristic Polynomial

- The Characteristic Polynomial of a matrix is computed by means of a special determinant
- The roots of characteristic polynomial are the *eigenvalues*
- Eigenvectors are vectors satisfying the expression

$$\Delta(\lambda) = |(\lambda I - A)| \tag{22}$$

$$Rv = \lambda v \tag{23}$$

$$0 = (\lambda I - R)v \tag{24}$$

Diagonal Factorization, Spectral Decomposition

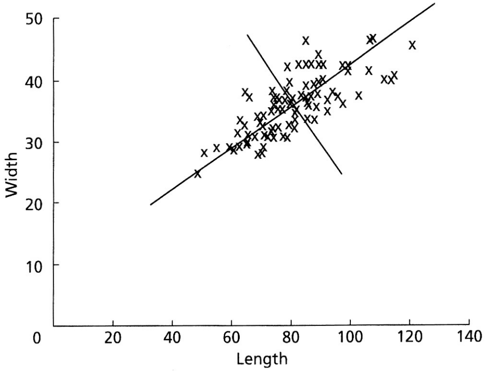
- Reversing the process we can write a matrix as a similarity transformation of a diagonal matrix
- If A is real-symmetric then the roots of its characteristic polynomial are guaranteed to be real.
- Eigenvectors belonging to distinct eigenvalues are orthogonal to one another $u \cdot v = 0$

$$A = PDP^{-1} \tag{25}$$

Principal Component Analysis in Two Dimensions

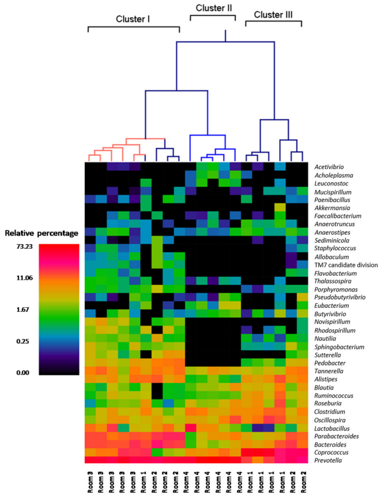
Box 1

A (supervised) ML workflow will minimally contain: *data*,*feature selection process*, *cross-validation*,*an ML algorithm*, and a means to *evaluate* performance



Clustering

- Methods: Hierarchical and k-means
- Look at the example again and what conclusions should we draw from this experiment?
- Important to observe that the dendrogram effectively defines a permutation of the features

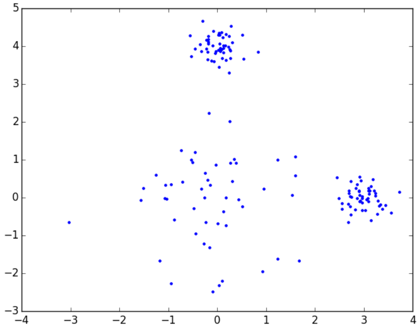


Clustering

- Divide an (unlabeled) dataset into a fixed number of clusters (or groups) such that the instances within each cluster are similar to one another and different from the instances in other clusters
- Purpose: Discover the underlying structure in some dataset, or to find an optimal reduced representation
- k-means
- Gaussian Mixture Models with E-M algorithm
- Hierarchical Clustering

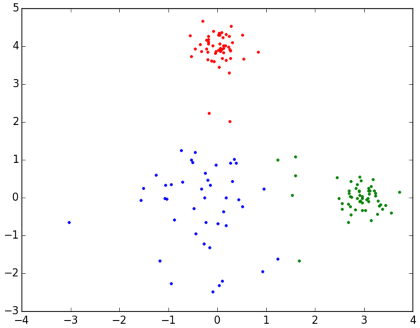
Example

- Easily visualized for two-dimensional data



Distinction between clustering and classification

- In classification problems we know already some set of classes to which each of our data instances belongs
- In clustering, conversely we have a completely flat dataset it is up to us to discover the structure



Clustering Function

- If we write our data instances $a_1, \ldots, a_n$ then we need some metric to quantify how well our data is clustered
- We can use the euclidean distance for this task: $d(a_i, a_j) = ||a_i - a_j||_2^2$
- Assume that we know in advance the number of clusters we want to fit to our data K
- A clustering is a function that assigns a label to each data instance: $f : \text{data} \rightarrow \{1, \ldots, K\}$

Objective for Clustering

- The clustering problem can also be formulated as an optimization
- Here f is the classification and n_k is the number of instances assigned to class k

$$J = \frac{1}{2} \sum_{k=1}^{K} \frac{1}{n_k} \sum_{f(i)=k, f(j)=k} d(a_i, a_j) \tag{26}$$

Clustering for Euclidean Distance

- It can be shown that the objective reduces as follows for the case of the Euclidean Distance, where A_k is the average data instance in group k

$$J = \sum_{k=1}^{K} \sum_{f(i)=k} ||a_i - A_k||_2^2 \tag{27}$$

- This suggests a more general approach where we iteratively minimize first over the groupings then the cluster centers

K-Means algorithm

$$J = \sum_{k=1}^{K} \sum_{f(i)=k} ||a_i - F_k||_2^2 \tag{28}$$

- Fix the number of clusters and pick an initial (random) guess of the cluster centers F_k
- Follow two steps until convergence:
- Assign each point to the nearest cluster center
- Recalculate the cluster centers as the average of all of the points in the cluster

Aspects of the K-Means Algorithm

- The cluster centers become a representative point for *all* of the points in a given cluster
- The minimum distance criterion partitions the space into a set of convex polyhedra
- Although the algorithm is guaranteed to converge the solution can be highly dependent upon the initial conditions, often desirable to run the algorithm multiples times and choose the result with the best (smallest) within cluster variance
- Another name for k-means is *vector quantization*
- Combinatorial dependence on the size of the dataset: infeasible to check all possible solutions

K-Means in scipy

- Scipy provides two options for implementing the k-means algorithm: `scipy.cluster.vq.kmeans` and `sklearn.cluster.KMeans`

Clustering Recap

- Last lecture: introduced clustering and a specific clustering algorithm called k-means
- *Clustering*: Process of dividing a dataset into clusters or groups such that all the instances in each cluster are similar to one another
- Properties of k-means: 1) number of clusters specified in advance (k) 2) Assignment dependent upon random initialization algorithm
- Algorithm: a) Cluster centers compute assignment b) Assignment compute new cluster centers

k-means versus Hierarchical Clustering

- Today we will introduce a new type of clustering algorithm called hierarchical clustering
- Unlike k-means the assignment will not depend upon initialization of the algorithm
- Also the number of clusters is not specified in advance
- Input data: measurement of distances between all pairs of data points, $(x_1, x_2, \ldots, x_n)$, use the matrix of distances, $d(x_i, x_j)$

Top-down and Bottom-up Clustering

- There are two types of hierarchical clustering algorithms: *agglomerative* and *divisive*
- Agglomerative: Begin with all data instances in separate clusters and repeatedly merge clusters until we have a single group (merge two 'nearest' clusters)
- Divisive: Opposite of agglomerative, begin with single large group and split until all clusters are singleton
- We will look agglomerative algorithms today
- From the definition we need a way to measure distances between two clusters (not only between data instances): this is known as the *linkage*

Linkages

- Given two groups of data instances $A = \{x_1, x_2, x_3\}$ and $B = \{x_4, x_5, x_6\}$
- Compare A and B using the distances their elements
- Distance here can be any metric we choose including the euclidean distance
- Goal: some function d that takes clusters (groups) as arguments and returns a number representing the dissimilarity of those clusters, $d(A, B)$

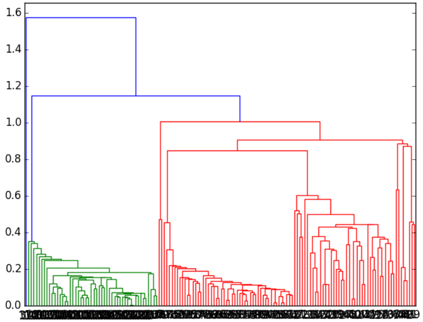
Single Linkage

- In single linkage we return the *smallest* distance between two groups
- Two groups are ‘close’ together if any of their points are near to each other

$$d(A, B) = \min_{i \in A, j \in B} d(x_i, x_j) \tag{29}$$

Visualizing Hierarchical Clustering with Dendrograms

- A *dendrogram* is a natural visual representation of the hierarchical clustering process (single linkage example)
- Cutting a dendrogram a fixed height gives a clustering of the data



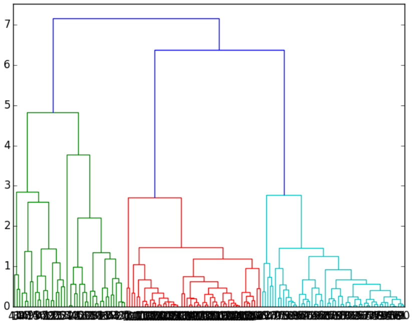
Dendrograms

- Tree where each node represents a cluster
- Leaf nodes represent singletons
- Root node contains all the data instances
- The height at which we draw a node is proportional to the dissimilarity of the two nodes
- Also the order of data instances is usually permuted so that the tree is more visually appealing (closer nodes placed near each other)

Complete Linkage

- In complete linkage we return the *maximum* distance between two groups
- Distance of the farthest pair

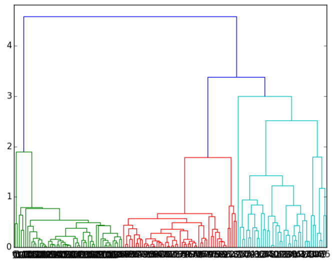
$$d(A, B) = \max_{i \in A, j \in B} d(x_i, x_j) \tag{30}$$



Average Linkage

- In complete linkage we return the *average* distance between all pairs of points in the two groups
- Distance of the farthest pair

$$d(A,B) = \frac{1}{|A||B|} \sum_{i \in A, j \in B} d(x_i, x_j) \tag{31}$$



Properties of hierarchical clustering

- As we run the algorithm (i.e. as we merge more nodes together) the distances between the clusters we are merging is always increasing
- For examples it is convenient to use the euclidean distance $||x - y||_2$ but in general we have many different distance metrics to choose from

Properties of hierarchical clustering

- As we run the algorithm (i.e. as we merge more nodes together) the distances between the clusters we are merging is always increasing
- For examples it is convenient to use the euclidean distance $||x - y||_2$ but in general we have many different distance metrics to choose from

Defining different metrics

- Different ways to compute distance between data instances

$$d(x, y) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2}, \text{ Euclidean} \tag{32}$$

$$d(x, y) = \sum_{i=1}^{N} |x_i - y_i|, \text{ Taxicab} \tag{33}$$

$$d(x, y) = \max_i |x_i - y_i|, \text{ Chebyshev} \tag{34}$$

$$d(x, y) = \left(\sum_{i=1}^{N} (x_i - y_i)^p \right)^{1/p}, \text{ Minkowski} \tag{35}$$

Problems with Single and Complete Linkages

- Single linkage can result in clusters that are too spread out while complete linkage suffers from chaining
- Average linkage is a compromise between single and complete although it also has its own difficulties for example hard to interpret dendrogram

Problem Statement for Images

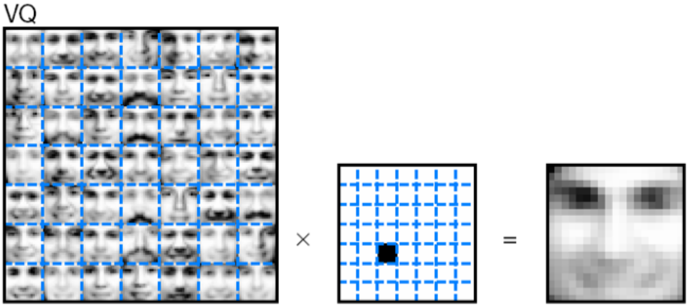
- Given a set of images we want to be able to do the following
- Create a set of basis images from which we can create new images by linear combination
- Find weights that produce any input image from the basis image
    - One set of weights for each input image

slide content adapted from materials from Marshall Tappen

Some possible solutions

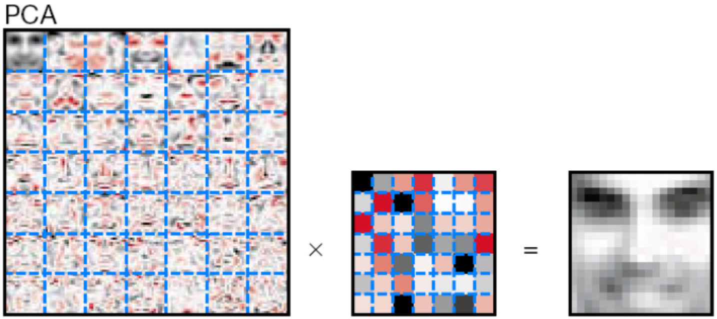
- Vector Quantization
- Principal Components Analysis
- Non-negative Matrix Factorization

Vector Quantization



- Find closest match
- Similar to nearest-neighbor classification
- *Limitations*: scales with number of basis images, provides no analysis

Principal Components Analysis

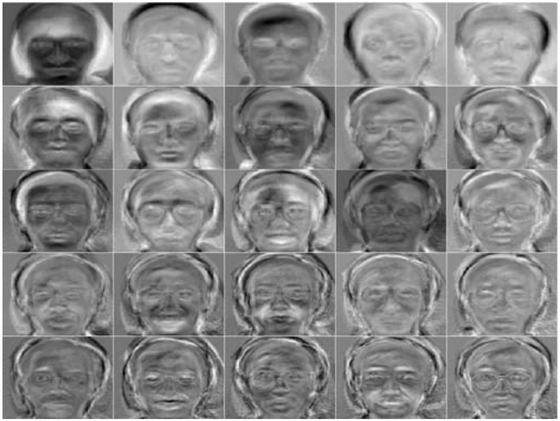


- Basis images are, by construction, orthogonal
- Reconstruction by linear combination

Aspects of PCA

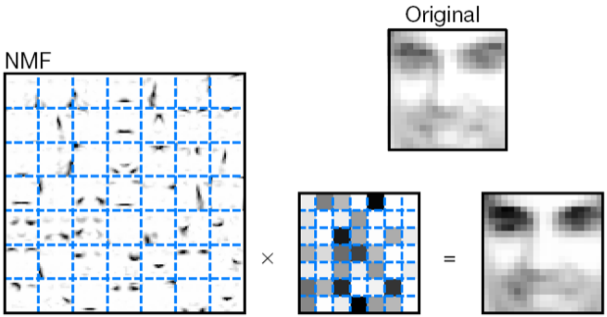
- PCA involves arbitrary linear combinations, we add some and se subtract some
- The basis images obtained do not necessarily correspond with our intuition
- In some contexts subtraction is not a sensible operation
- How is a face subtracted, what is subtraction in the context of document classification.

Eigenfaces Example



- Probably not how faces are represented in the brain

Non-negative matrix factorization

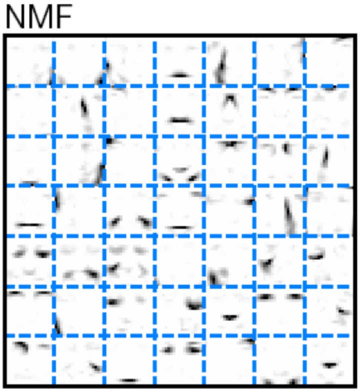


- Similar to PCA, but the coefficients cannot be non-negative

NMF Basis Image Properties

- Only allowing addition makes more intuitive sense in certain contexts and has some correspondence with how neurons operate
- Constraining the reconstruction coefficients to be positive often leads to nice basis images
    - Basis images represent different *parts* of the objects being studied

Non-negative matrix factorization

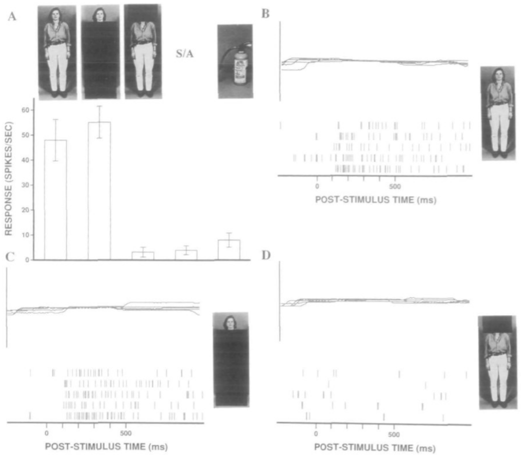


- The factorization has naturally found what we might think of as parts of the faces

Comparison of PCA and NMF

- PCA
  - Produces an optimal set of basis images
  - But this optimality might not be useful for your application
- NMF
  - Produces coefficients with a constraint
  - Can naturally produce a nicer basis, not constrained to be orthogonal

Evidence from Neuroscience



- Different visual stimuli presented to macaque monkey

Wachsmuth et al 1994

- Of 53 neurons,
- 32 percent responded to head only
- 9 percent responded to body only
- 41 percent responded to both the head and the body in isolation
- 17 percent responded to the whole body only
- Suggestive a parts based encoding of figures in the brain

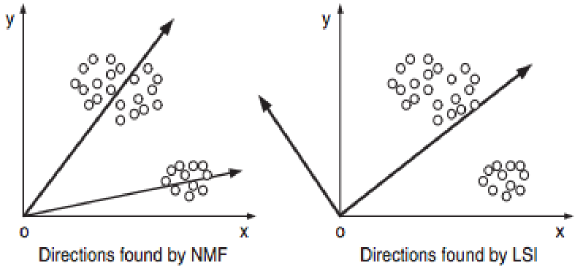
Non-negative Matrix Factorization

- Non-negative Matrix Factorization provides a similar decomposition of the dt matrix
- Naturally additive model since the values are all constrained to be non-negative

$$\arg\min_{W,H} ||X - WH||^2 = \sum_{i,j} X_{ij} - WH_{ij}$$

Document Clustering

- *Document Clustering*: Partitioning a corpus into a predefined number of clusters related to a coherent topic
- NMF applied to text analysis by Xu, Liu, Gong “Document Clustering Base on Non-negative Matrix Factorization”



NMF Document Clustering Algorithm

- *Document Clustering Algorithm*:
    - Construct the term-document matrix $\mathbf{X}$ from the given corpus
    - Find an NMF decomposition of $\mathbf{X}$
    - Normalize the factors U and V
    - Examine each column of V and look for the component with the largest value and assign the corresponding document to cluster k
- Standard Datasets for Document Clustering: NIST Topic Detection and Tracking (TDT2), Reuters dataset

Performing Non-Negative Matrix Factorization in Python

```
import numpy as np
from sklearn.decomposition import NMF
<...>
X = <...>
<...>
model = NMF(n_components=2, init='random', random_state=0)
W = model(fit(X))
H = model.components_
```

- Demonstration of the singular value decomposition routine in the numpy linear algebra package
- This can be used to experiment with LSA on a small document corpus

Task

- Write a script to perform NMF decomposition on a few simple matrices
- Characterize the error between the reconstruction and the original data as you change the rank
- Create a plot of error vs. rank