innovate achieve lead

BITS Pilani

Machine Learning (IS ZC464) Session 6 :

Feature Engineering

Discussion

- Feature engineering
- Dimensionality Reduction
- Feature Extraction techniques
  - Bag of words (TF-IDF, n-grams, stemming etc.)
  - Image features
  - Transform based features
- Feature selection algorithms
  - Principal Component Analysis (PCA)
  - Forward selection and backward elimination algorithm

Feature engineering

- This refers to the practice of using mathematical transformations of raw input data to create new features to be used in a machine learning model.
- Examples:
  - Counting the occurrence of a particular word across a text document.
  - Computing statistical summaries such as mean, standard deviation, median etc. of any raw data.

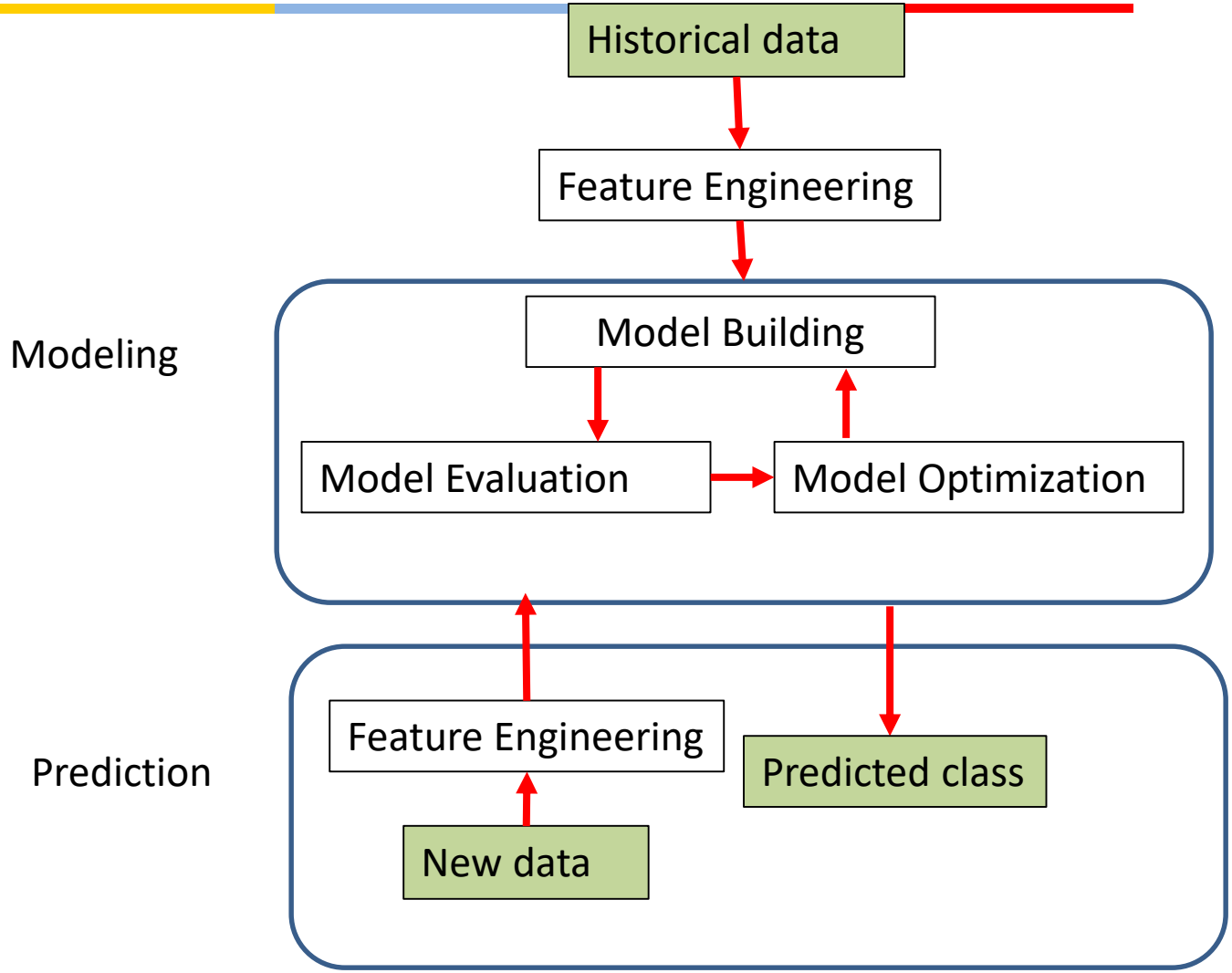
Why feature engineering?

- The features derived from the transformations may be more closely related to the target classes. (population density per unit area works better than the population data alone)
- This allows to bring in external data into the ML model (geographic locations of internet users along with their access patterns)
- Unstructured data can be used for machine learning applications (text, videos, images etc.)

Advantages of feature engineering

- Only the most informative and discriminative features are used in building the machine learning model.
- The selected good features in very less number can produce the best accuracy of classification at minimum computational cost.
- Real time response time of the system
- Capability to handle more robust scenarios
- Memory requirements are less.

Machine learning model



Data for an application

- Face recognition – face images
- Speaker recognition – speech signals
- Brain Computer Interface – EEG signals
- Stock index prediction – numeric data
- Sentiment analysis – text data
- Crime prediction – text and numeric data
- Email spam detection - text

Data describing objects or events

- Univariate data
  - Prediction of weight of a person using one value as height
  - Prediction of rain based on humidity alone
- Multivariate data
  - Face recognition using multiple features of face such as nose, eyes, jawline, color of eyes, shape of face etc.
  - Stock index prediction - term structure of interest rates (TS), short term interest rate (ST), long term interest rate (LT), consumer price index (CPI), industrial production (IP), government consumption (GC), private consumption (PC), gross national product (GNP), gross domestic product (GDP) etc.

Difference between data and features

- Data is a collection of raw facts collected over a period of time for the given application
- Feature is a value obtained by processing some or all of data values.
- Example : Face recognition application
  - Data – Intensity values of pixels
  - Features – mean, standard deviation of all values,

Feature extraction algorithms

- Images – 2 dimensional signals
  - Mean, standard deviation, histogram, moments
  - Transform based features – Fourier transform, Discrete Cosine Transform, Wavelet Transform etc.
- Speech – 1 dimensional signals
  - Energy of speech frame, zero-crossing rates, Mel-Frequency Cepstral Coefficients (MFCCs)
- Text
  - Bag of N-gram model, Term Frequency-Inverse Document Frequency (TF-IDF) etc.

Dimensionality Reduction

- Dimensionality refers to the number of features used in the classification or regression problem solving.
- Example : If an image is of size 100x100, and
  - the features are taken as pixel intensity, then the dimensionality is 10000.
  - If we take features as its mean and standard deviation, then dimensionality is 2
  - If we take the transform of the image and select only coefficients at the upper left corner 30 in number, then the dimensionality is 30

Why is dimensionality reduction necessary?

- To reduce space and time complexity of the algorithm.
- To reduce overfitting
- To increase accuracy

Working with text features

- Example text data
  - News items for automated summary in Natural Language Processing (NLP) domain
  - Tweets and messages for analysis
  - Email text messages
  - Facebook comments etc.
- Type of words as data
  - Significant words such as meeting, project, birthday etc.
  - Stop words such as is, and, the etc.

Document classification analogy

???

John Terry scored on a header to lift Chelsea to a 1-0 victory over Manchester United and extend the Blues' Premier League lead to 5 points. Chelsea had been frustrated by Manchester United for 76 minutes, but took advantage of a free kick awarded when Darren Fletcher fouled Ashley Cole.
Brian Ching scored six minutes into overtime and the Houston Dynamo advanced to Major League Soccer's Western

???

In the Senate, where proposals differ substantially from the House-passed measure on issues like a government-run plan and how to pay for coverage, the bill is stalled while budget analysts assess its overall costs. The slim margin in the House — the bill passed with just two votes to spare, and 39 Democrats opposed it — suggests even greater challenges in the Senate, where the majority leader, ...

Classify each document as sports or politics

Slide Resource : http://people.csail.mit.edu/tomasz/presentations/visiontutorial.ppt

Bag of words

- Count the occurrence of words and add that as a new feature
- Example text

Data set 1: John likes to watch movies. Mary likes movies too.

Data set 2: John also likes to watch football games

BoW1 ={"John":1,"likes":2,"to":1,"watch":1,"movies":2,"Mary":1,"too":1};

BoW2 ={"John":1,"also":1,"likes":1,"to":1,"watch":1,"football":1,"games":1};

Bag of words

- The Bag-of-words model is mainly used as a tool of feature generation.
- After transforming the text into a "bag of words", we can calculate various measures to characterize the text.
- The most common type of characteristics, or features calculated from the Bag-of-words model is term frequency, namely, the number of times a term appears in the text.

Example

BoW1 ={"John":1,"likes":2,"to":1,"watch":1,"movies":2,"Mary":1,"too":1};
BoW2 ={"John":1,"also":1,"likes":1,"to":1,"watch":1,"football":1,"games":1};

	John	likes	to	watch	movies	Mary	too	also	football	games
data1	1	2	1	1	2	1	1	0	0	0
data2	1	1	1	1	0	0	0	1	1	1

These occurrences are known as term frequencies.

The features in ML should be homogeneous for all observations, i.e. the number of features is same for all instances.

The bag of words is usually sparse (more zeros in it)

Bag of words model

- The splitting of the text into pieces is known as tokenization.
- Group of words in a split is called n-grams.
- Two or three words combinations are called as bigram and trigram e.g. ‘also likes’ and ‘likes to watch’ respectively.
- The extracted words can be transformed only to lower case letters. For example Likes and likes should be treated as the same feature.

Stemming

- It is a powerful transformation which removes word suffixes.
- Example *study, studied, studying, studies* etc. are converted to study and then the word count is computed

Feature vector using bag of words

	John	likes	to	watch	movies	Mary	too	also	football	games
data1	1	2	1	1	2	1	1	0	0	0
data2	1	1	1	1	0	0	0	1	1	1

- Feature vector corresponds to the collection of values for each instance

  Data 1: <1,2,1,1,2,1,1,0,0,0>

  Data 2: <1,1,1,1,0,0,0,1,1,1>
- Feature vector sizes are same for all instances
- Two similar text documents are likely to display similarity in their feature vector except only at one or two places
- It is advisable to remove stop words from the features.

Term frequency- inverse document frequency (TF-IDF)

- Term frequency(tf)
  - Number of times a word occurs
  - 1 if a word occurs or 0 if it does not (binary)
  - 1+log(tf) (logarithmic)
- Document frequency (df)
  - Number of documents containing specific word
- Term frequency-Inverse document frequency (TF_IDF)
  - TF_IDF = tf * (total_docs/df)

Term frequency- inverse document frequency (TF-IDF)

- This is a powerful technique to generate features from any corpus of text.
- Using this feature extraction technique, relatively uncommon words attain more value.

Image features

- An image is a two dimensional signal.
- Pixel intensities make the raw data.
- Raw pixel values are very large in number and do not make useful features (not discriminative)
- Image data need to be processed for more information

Digital image

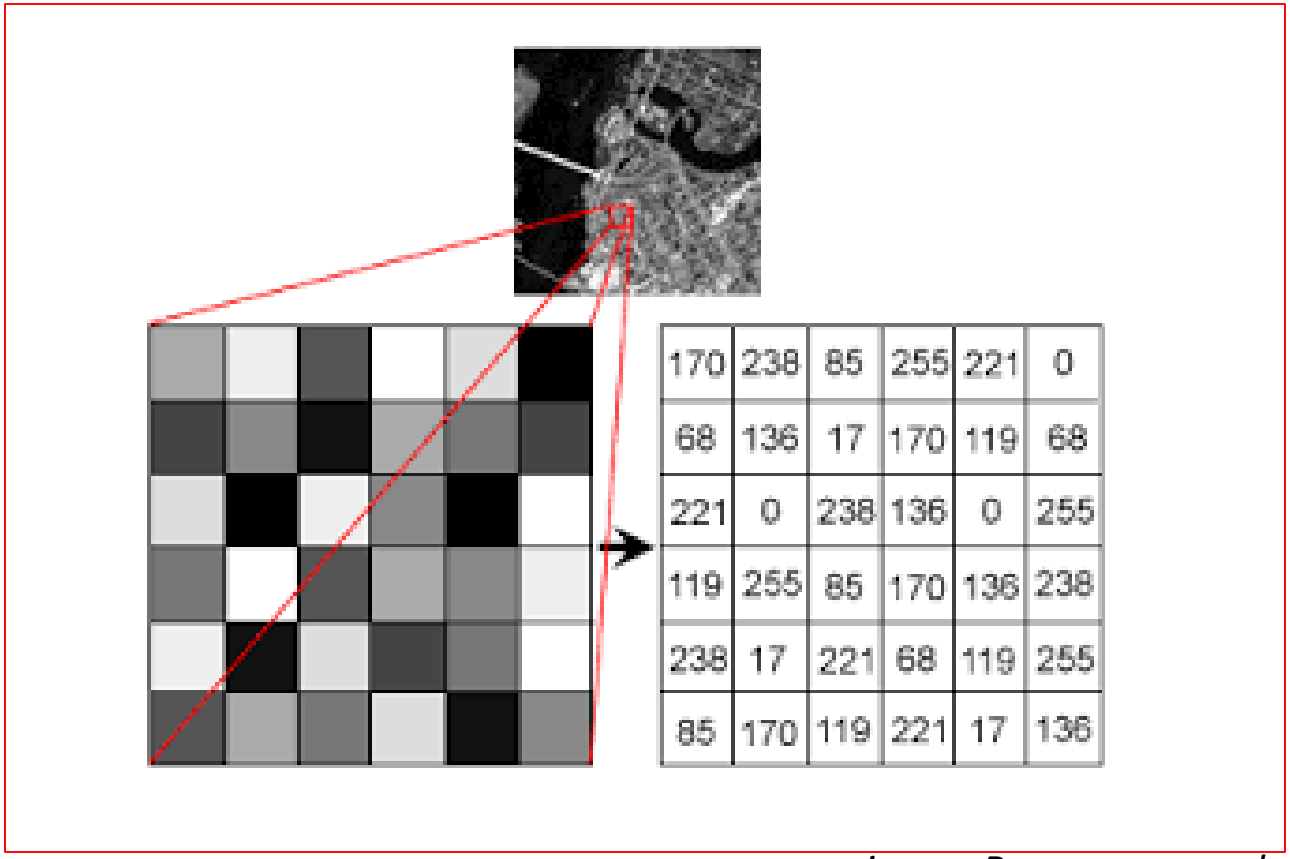


image Resource : google

Feature extraction from images: Mean and standard deviation features

- Normal pixel values range from 0 to 255 (8-bit images)
- Compute the sum of all pixel intensities.
- Find the average intensity of each image.
- Find the difference of each pixel value with the average intensity- call it deviation from mean
- Find the sum of squares of deviations
- Divide by the image size (compute standard deviation)

Mean and standard deviation can add to the feature vector obtained using other algorithms but rarely make good features alone.

Feature extraction from images: Color features

- Each color image consists of pixel values in three different color ranges – Red, Green and Blue
- Each color image of size nxn is equivalent to three nxn images (channels)
- Each channel image is separately processed for the feature extraction.
- If mean, median, standard deviation, mode etc. are the *k* features, then there are total *3k* features for a color image

Feature extraction from images: Shape features

- HOG features: Histogram of oriented gradients (HOG) describes the shapes and objects in image regions that are not too sensitive to small changes in scale and orientation.
- Steps-
  - Calculate the gradient image
  - Divide the image into small blocks (cells)
  - Calculate the orientation of the gradients inside those cells
  - Calculate the histogram of those orientations in individual cells

Feature extraction using transforms

- Fourier transform
- Discrete Cosine Transform
- Wavelet transform

(details to be studied in a separate course on Digital image processing)

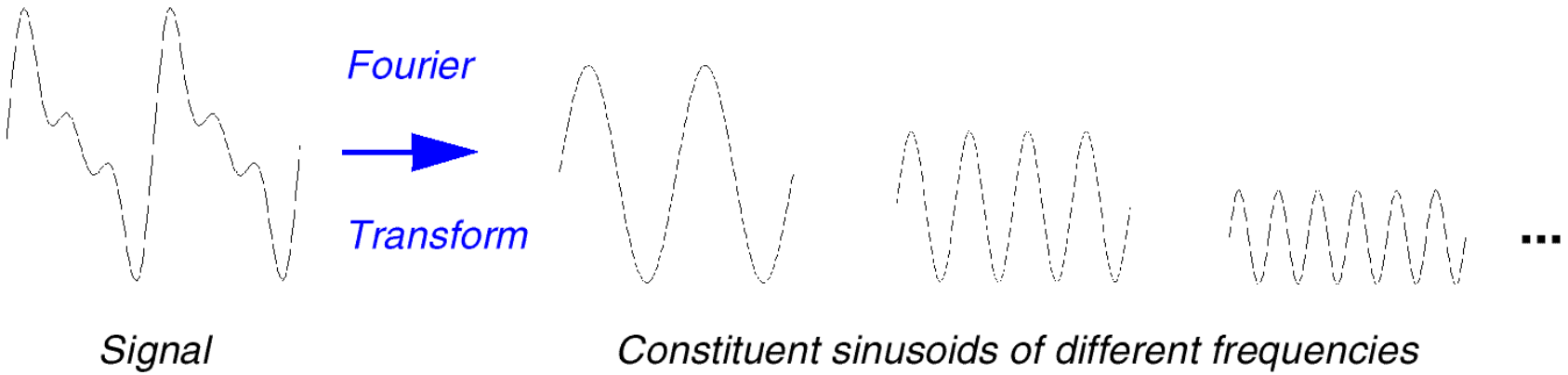
Fourier transform

- It transforms the original data into frequency domain
- Defined as the sum over all time of the signal *f(t)* multiplied by a complex exponential, and the result is the Fourier coefficients F

$$F(w) = \int_{-\infty}^{\infty} f(t) e^{-iwt} dt$$

- Loses time information

Fourier transform



Slides 30-37 Resource : google

Discrete Cosine Transform

$$S(k_1,k_2)=\sqrt{\frac{4}{N^2}}C(k_1)C(k_2)\sum_{n_1=0}^{N-1}\sum_{n_2=0}^{N-1}s(n_1,n_2)\cos\left(\frac{\pi(2n_1+1)k_1}{2N}\right)\cos\left(\frac{\pi(2n_2+1)k_2}{2N}\right)$$

where $k_1,k_2,n_1,n_2=0,1,\cdots\cdots N-1$, and

$$C(k)=\begin{cases}1/\sqrt{2} & \text{for } k=0\\ 1 & \text{otherwise}\end{cases}$$

Original image

166	162	162	160	155	163	160	155
166	162	162	160	155	163	160	155
166	162	162	160	155	163	160	155
166	162	162	160	155	163	160	155
166	162	162	160	155	163	160	155
161	160	155	159	154	154	156	154
159	163	158	163	155	155	156	152
159	162	162	160	153	153	153	151

transformed image

248	19	3	4	-7	9	1	-7
11	-2	3	6	-3	2	5	0
-4	2	-2	-3	0	-1	-1	0
-1	-1	1	1	2	0	-1	0
2	1	0	0	-2	0	3	0
0	0	-1	0	0	0	-1	-1
-3	0	1	0	1	0	0	0
3	0	0	0	-1	0	0	0

Wavelet transform

- Short time localized waves with zero integral value.

$$\Psi_{a,b}(x) = \frac{1}{\sqrt{a}} \Psi\left(\frac{x-b}{a}\right)$$

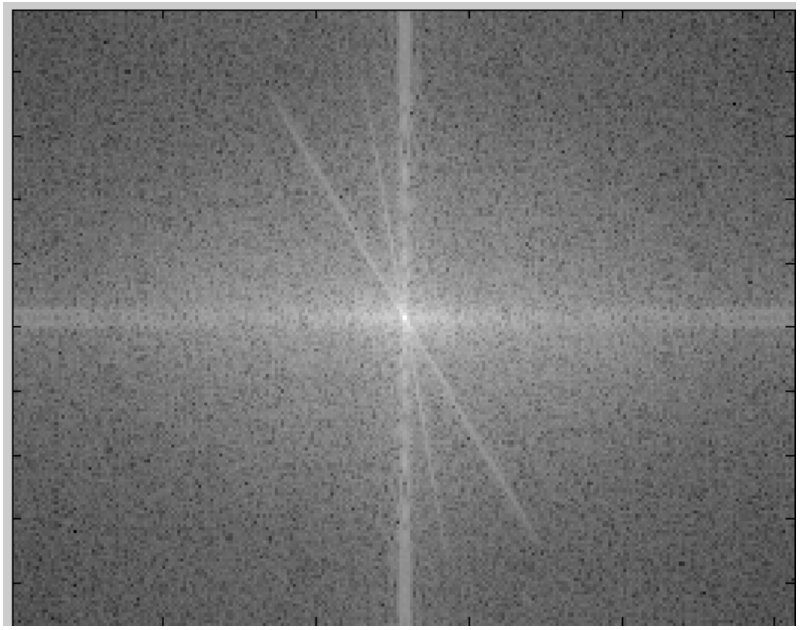
- b – shift coefficient
- a – scale coefficient

$$\Psi_{a,b_x,b_y}(x,y) = \frac{1}{|a|} \Psi\left(\frac{x-b_x}{a}, \frac{y-b_y}{a}\right)$$

- 2D function

Image Processing in the Fourier Domain

Magnitude of the FT

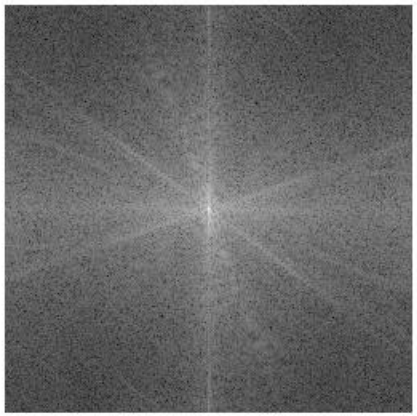


Does not look anything like what we have seen

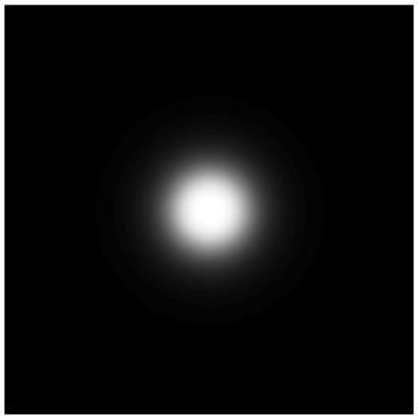
Low-pass Filtering

Original image

FFT of original image



Low-pass filter



Let the low frequencies pass and eliminating the high frequencies.

Low-pass image

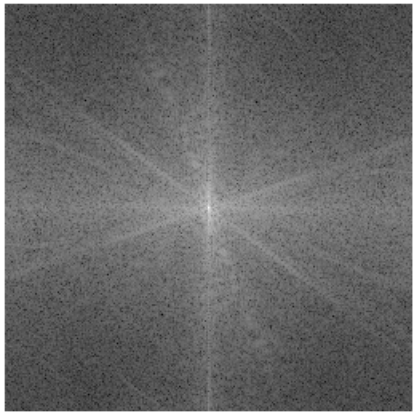
FFT of low-pass image

Generates image with overall shading, but not much detail

High-pass Filtering

Original image

FFT of original image

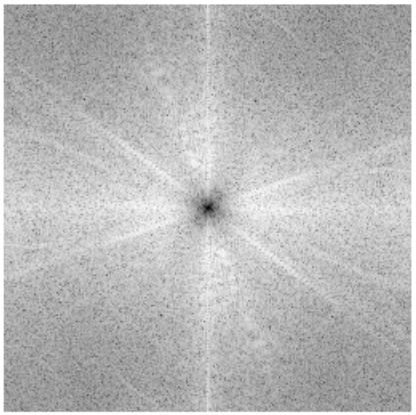


High–pass filter

Lets through the high frequencies (the detail), but eliminates the low frequencies (the overall shape). It acts like an edge enhancer.

High–pass image

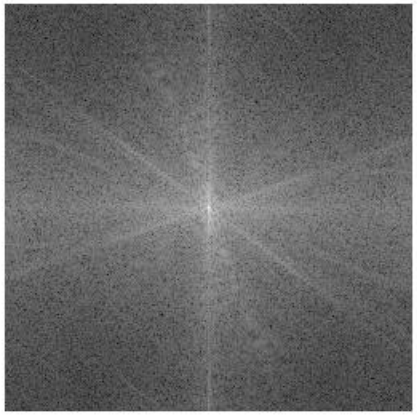
FFT of high–pass image



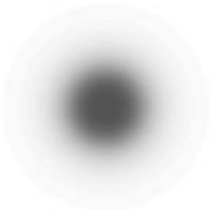
Boosting High Frequencies

Original image

FFT of original image

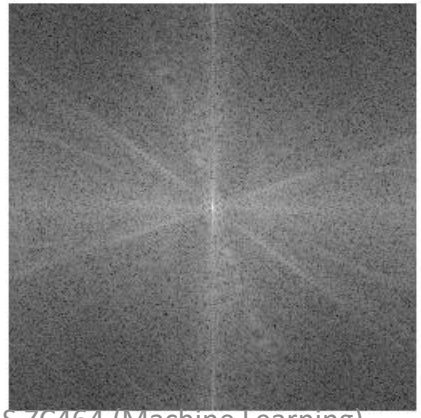


High–boost filter

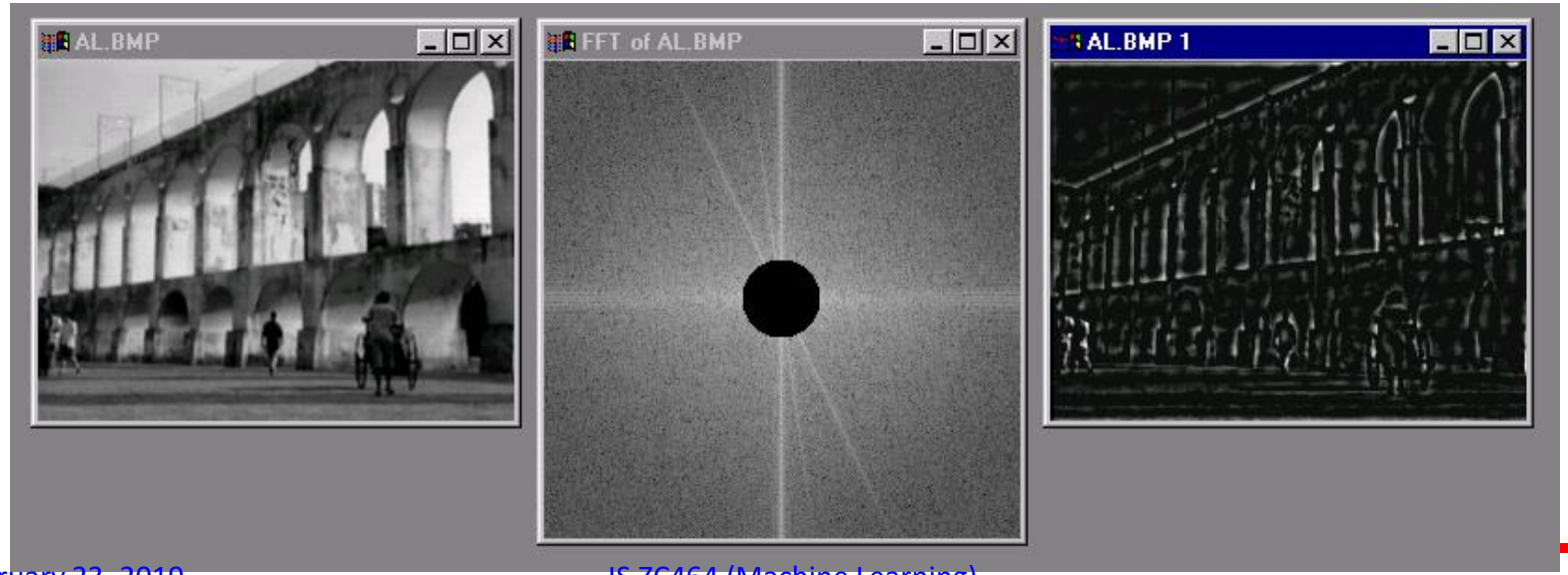


High boosted image

FFT of high boosted image



Most information at low frequencies!



Feature Selection

- Process of choosing the most predictive subset of features out of a larger set
- Example
  - Bag of words – sort and select only few words with frequencies (or if-idf) greater than a set threshold
  - Discrete Cosine Transform – select the upper left corner coefficients in zig zag way
  - Wavelets transform – only a particular level of resolution

Dimensionality Reduction

- Principle Component Analysis

  Principal component analysis** (**PCA**) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called **principal components (Wikipedia)

- Forward selection and backward elimination

  Iterative selection or removal of features