Solution to Sample Mid-Semester Examination Question Paper T. E. (Mechanical Engineering) Subject: Artificial Intelligent and Machine Learning

Abhishek D. Patange
ABB
115 PUBLICATIONS **753** CITATIONS

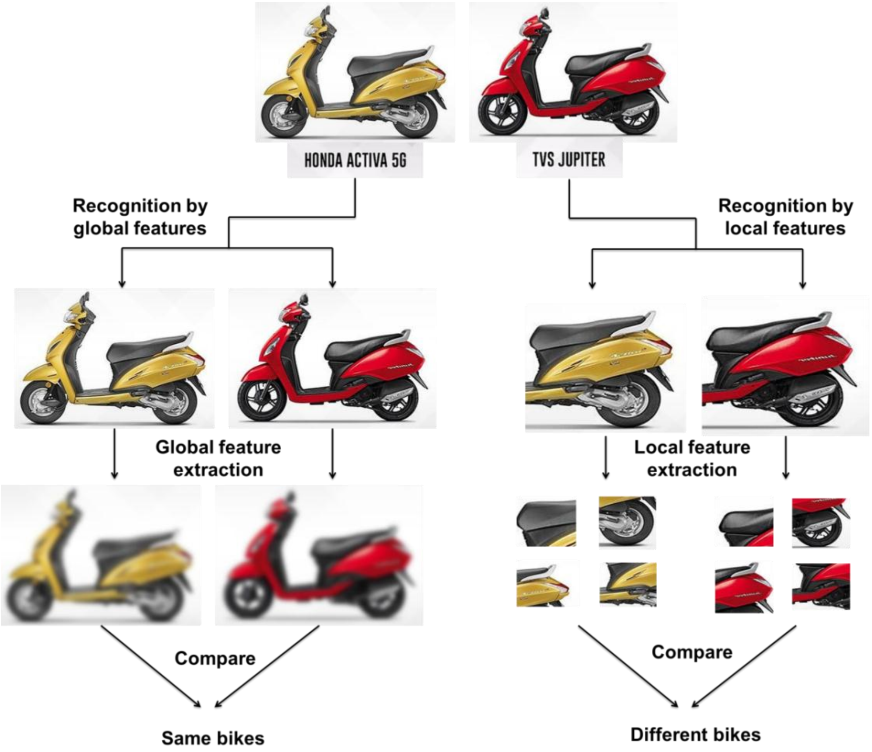
Mid-Semester Examination T. E. (Mechanical Engineering)

Subject: Artificial Intelligent and Machine Learning

Time: 1.30 hours **Marks: 30**

1 Define and explain following terms with respect to example given below. 2

a. Global feature b. Local feature



First of all definition of Global and Local feature is expected:

- Global features describe the visual content of the entire image by a single vector. They represent the texture, color, shape information which are the most popular for image representation.
- Local features aim to detect the interest points (IPs) in an image and describe them by a set of vectors.
- Simply speaking, global features describe the entire image, whereas local features describe the image patches (small group of pixels).

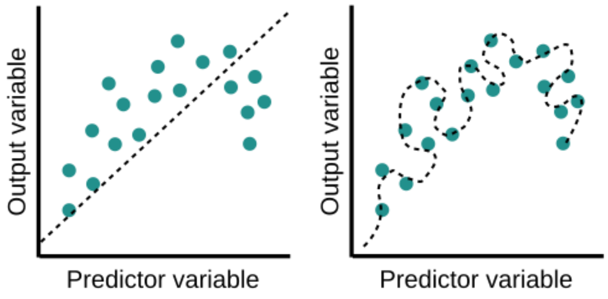
Explanation with respect to example:

- Here in the comparison of 2 images of Honda active 5G and TVS Jupiter, global features describe the texture, shape information of the entire image of Honda active 5G and TVS Jupiter by a single vector.
- So overall both bikes seems to be same. For this example, using global features representation may create confusion.
- On the other hand local features i.e. image patches (side view of rear end of

Honda active 5G and TVS Jupiter) in terms of small group of pixels tries to detect the interest points (IPs).

- This would help in identifying difference between two bikes. Thus local features seem to be useful to develop a classification model.

2 Represent over-fitting and under-fitting in regression problem pictorially. State one real-life example. 4

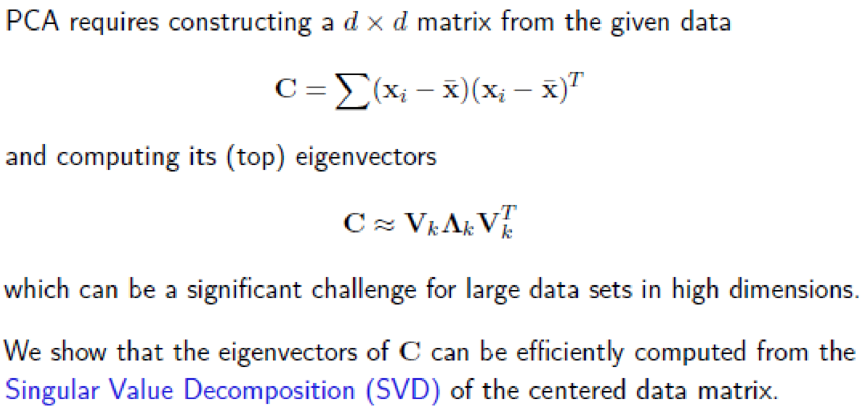


(a) Under-fitting (b) Over-fitting

Example: predict who will be an “A” student in college.

- Let’s say “Peter Evans Hope” is an “A” student in your dataset.
- If your model says anyone named “Peter Evans Hope” is an “A” student, the model will correctly predict this specific student but it is over-fitted.
- Because in the general population, there probably isn’t another “Peter Evans Hope” (and if there is, probably not an “A” student).
- If your model says anyone who graduated from [insert top high school] will be an “A” student in college, this is under-fitting (too general).
- Within the graduates from that high school, there will be a range of college GPAs - what else can explain who gets the “A”?

3 Explain mathematics behind PCA (Principal Components Analysis). 4



PCA requires constructing a $d \times d$ matrix from the given data

$$\mathbf{C} = \sum(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

and computing its (top) eigenvectors

$$\mathbf{C} \approx \mathbf{V}_k \mathbf{\Lambda}_k \mathbf{V}_k^T$$

which can be a significant challenge for large data sets in high dimensions.

We show that the eigenvectors of $\mathbf{C}$ can be efficiently computed from the Singular Value Decomposition (SVD) of the centered data matrix.

Let $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} \in \mathbb{R}^{n \times d}$ and $\tilde{\mathbf{X}} = \begin{bmatrix} \tilde{\mathbf{x}}_1^T \\ \vdots \\ \tilde{\mathbf{x}}_n^T \end{bmatrix} \in \mathbb{R}^{n \times d}$ (where $\tilde{\mathbf{x}}_i = \mathbf{x}_i - \bar{\mathbf{x}}$) be the original and centered data matrices (rows are data points).

Then

$$\mathbf{C} = \sum \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^T = [\tilde{\mathbf{x}}_1 \ldots \tilde{\mathbf{x}}_n] \cdot \begin{bmatrix} \tilde{\mathbf{x}}_1^T \\ \vdots \\ \tilde{\mathbf{x}}_n^T \end{bmatrix} = \tilde{\mathbf{X}}^T \tilde{\mathbf{X}}.$$

Again, this shows that $\mathbf{C}$ is square, symmetric and positive semidefinite and thus only has nonnegative eigenvalues.

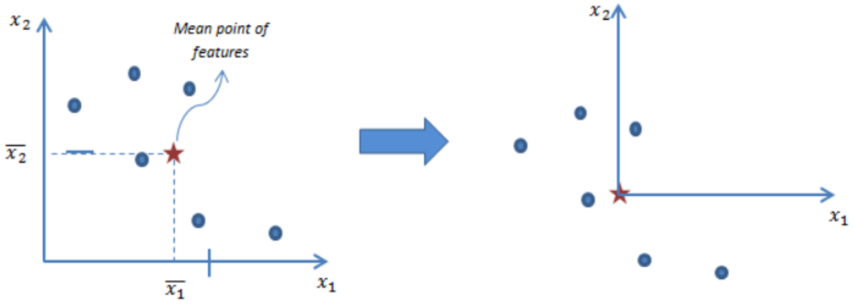


Fig 1 : Finding the Mean point of Features and Shifting the data points

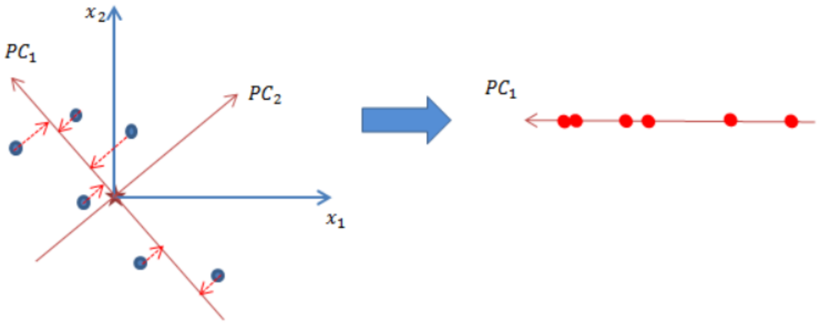


Fig 2 : Finding the directions of variance of data points and Projecting data points on PC1

4 Suppose you have a several pictures of Nuts, Bolts, Washers and Locating Pins with different orientations. You need to develop an intelligent classification model. Which approach of machine learning will you select – supervised or unsupervised? How? 4

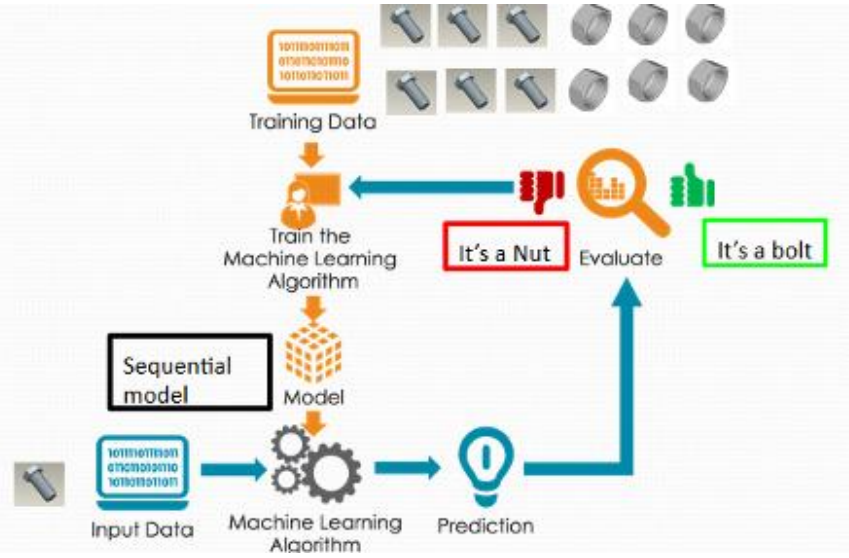
Since a several pictures of Nuts, Bolts, Washers and Locating Pins with different orientations are available and labels are known, I would select a supervised approach of machine learning for developing an intelligent classification model.

A methodology would be as follows:

1. **Data Collection**: Data for each class was collected from standard part libraries.
2. **Data Preparation:** Isometric views can be taken from each image and reduced to

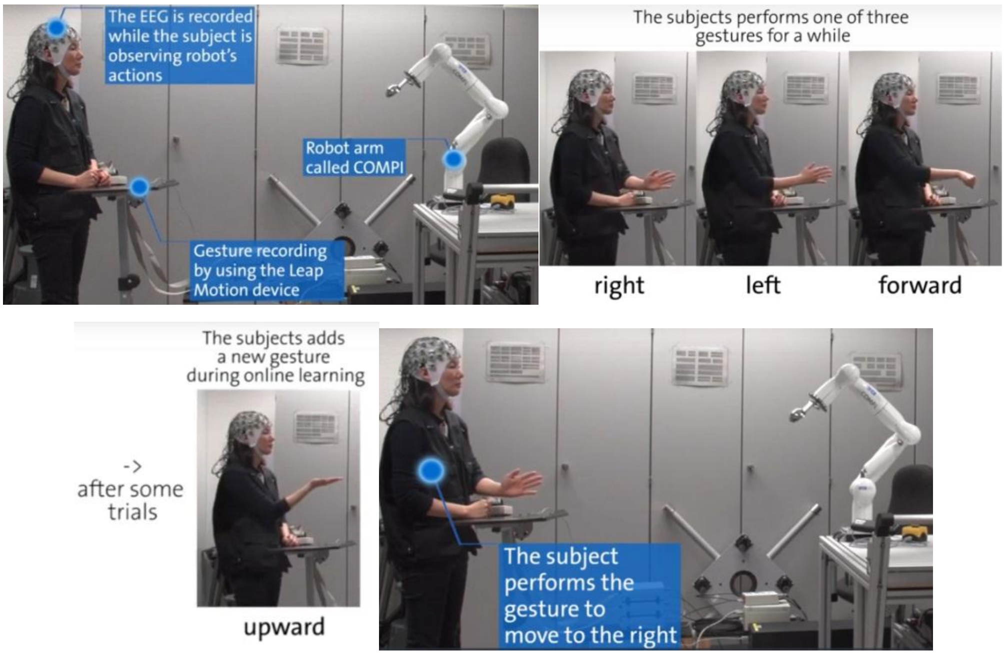
standard pixels.

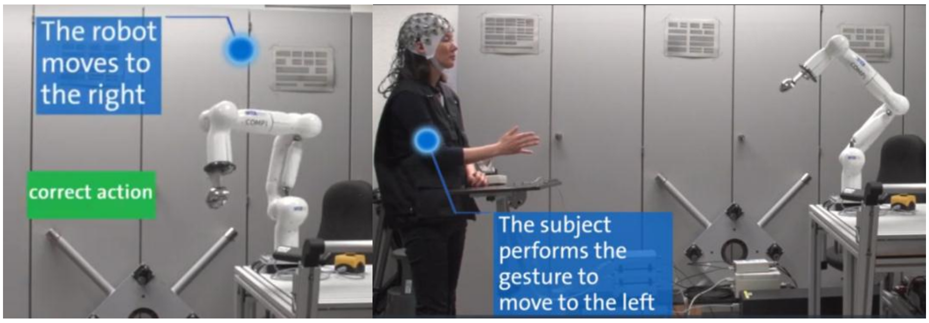
3. **Model Selection :** A Machine Learning model was selected as it was simple and good for image classification
4. **Train the Model:** Consider a suitable train-test split
5. **Evaluate the Model:** The results of the model were evaluated. How well it predicted the classes?
6. **Hyperparameter Tuning:** This process is done to tune the hyperparameters to get better results.
7. **Make Predictions:** Check how well it predicts the real world data.

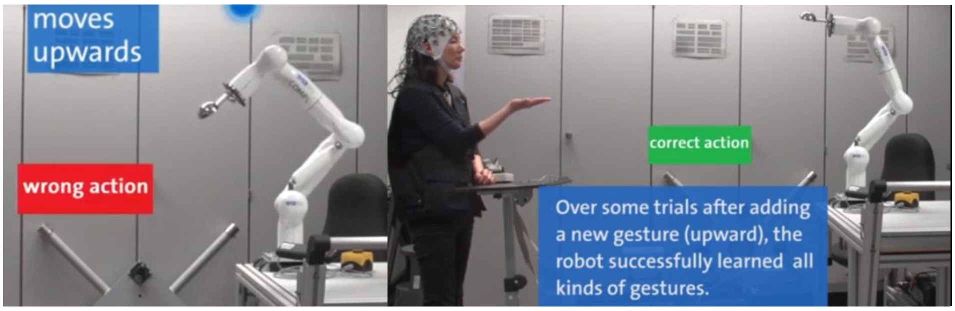


OR

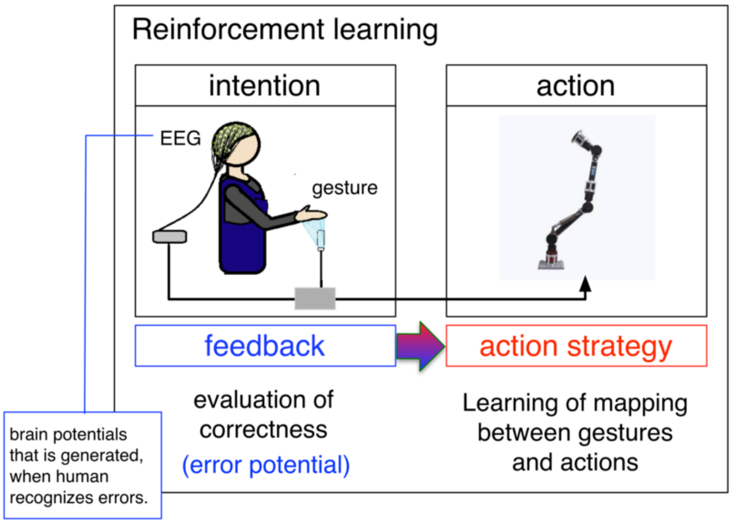
4 Explain role of reinforcement learning in following example. Identify environment, agent, different actions, reward, punishment etc. Draw its block diagram. 4

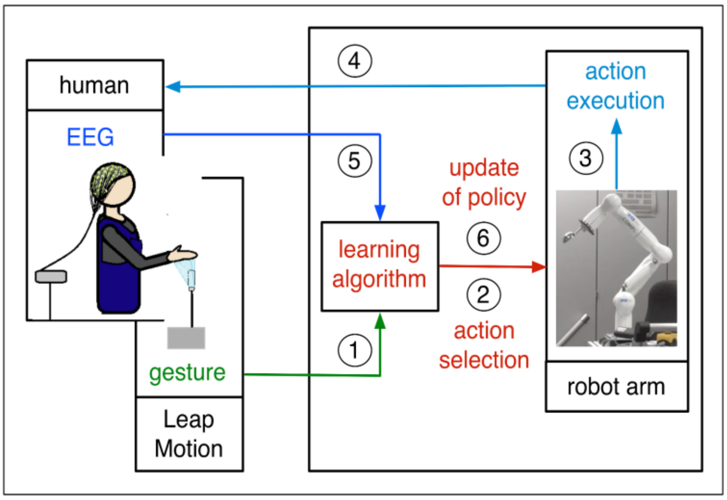

The EEG is recorded while the subject is observing robot's actions
Robot arm called COMPI
Gesture recording by using the Leap Motion device
The subjects performs one of three gestures for a while
right
left
forward
-> after some trials
The subjects adds a new gesture during online learning
upward
The subject performs the gesture to move to the right


The robot moves to the right
correct action
The subject performs the gesture to move to the left


moves upwards
wrong action
correct action
Over some trials after adding a new gesture (upward), the robot successfully learned all kinds of gestures.

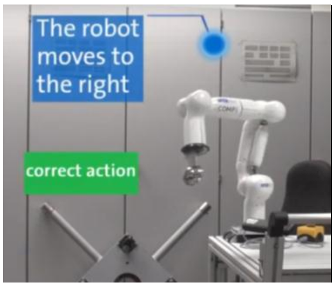
The reinforcement learning provides the means for robots to learn complex behavior from interaction on the basis of generalizable behavioural primitives. From the human negative feedback, the robot learns from its own misconduct.


Reinforcement learning
intention
action
EEG
gesture
feedback
action strategy
evaluation of correctness
(error potential)
Learning of mapping between gestures and actions
brain potentials that is generated, when human recognizes errors.



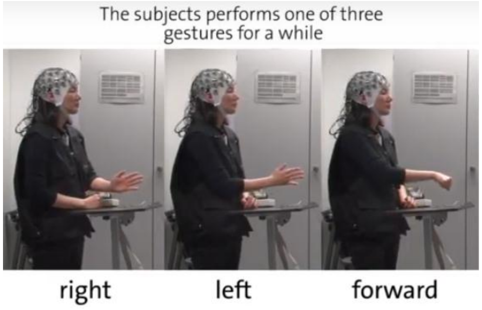
Agent

Reward

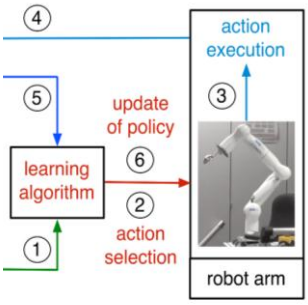


Punishment

Different actions

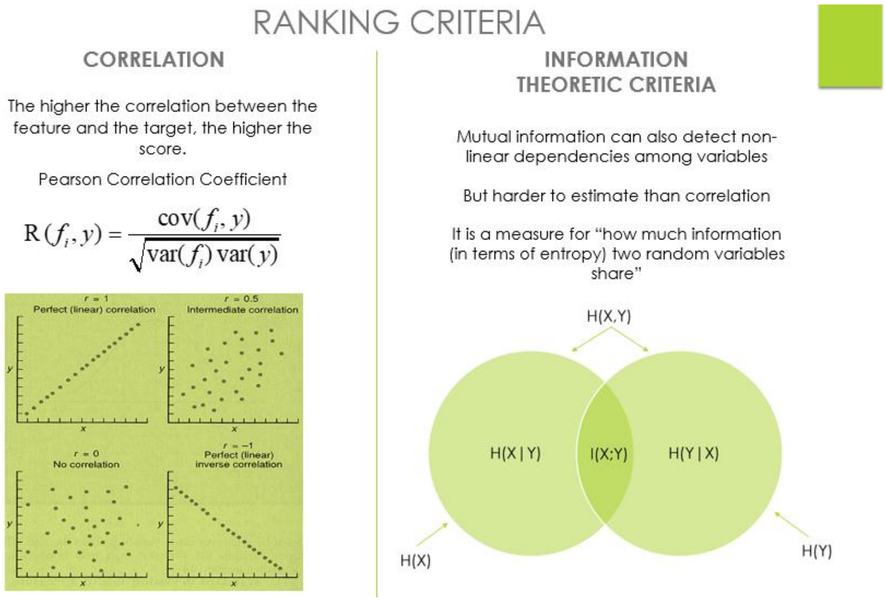


Environment



5 Explain feature ranking as a feature selection method. 4

Ranking is the process of ordering the features by the value of some scoring function, which usually measures feature-relevance. Resulting set: The score *S(fi)* is computed from the training data, measuring some criteria of feature *fi*. By convention a high score is indicative for a valuable (relevant) feature. A simple method for *feature selection* using **ranking** is to select the *k* highest ranked features according to *S*. This is usually not optimal, but often preferable to other, more complicated methods. It is computationally efficient — only calculation and sorting of *n* scores.



RANKING CRITERIA

CORRELATION

The higher the correlation between the feature and the target, the higher the score.

Pearson Correlation Coefficient

$$R(f_i, y) = \frac{\mathrm{cov}(f_i, y)}{\sqrt{\mathrm{var}(f_i)\,\mathrm{var}(y)}}$$

r = 1
Perfect (linear) correlation
r = 0.5
Intermediate correlation
r = 0
No correlation
r = −1
Perfect (linear) inverse correlation
x
y

INFORMATION THEORETIC CRITERIA

Mutual information can also detect non-linear dependencies among variables

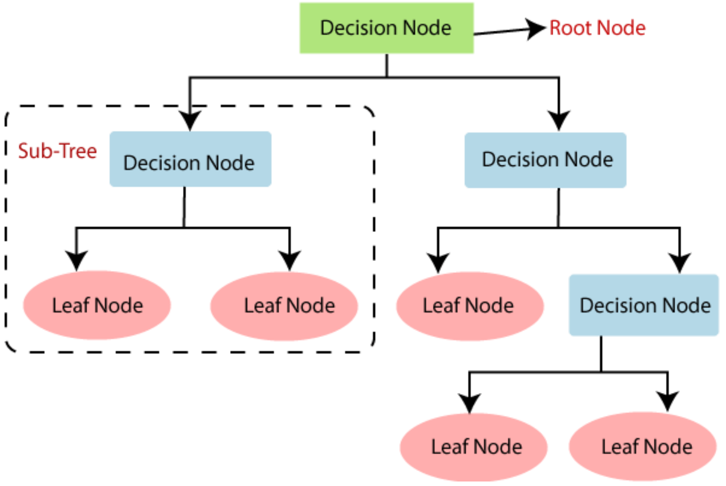
But harder to estimate than correlation

It is a measure for "how much information (in terms of entropy) two random variables share"

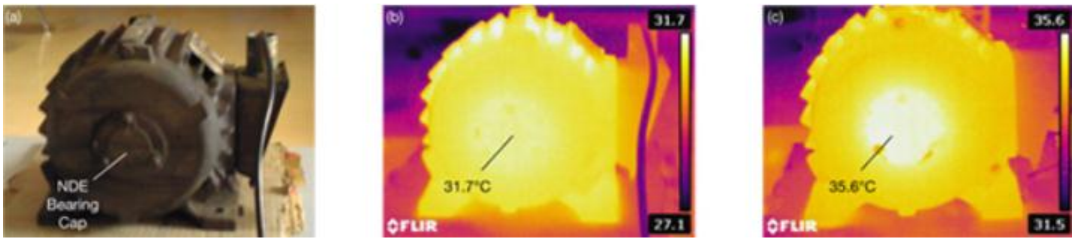
OR

5 What does 'Root Node', 'Leaf Node', and 'Branch node' represent in decision tree model? Represent them pictorially. 4

- **Root Node:** Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.
- **Branch Node:** A node formed by splitting the tree using different tests.
- **Leaf Node:** Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.



6 Following picture represent an external temperature rise at the non-drive end (NDE) bearing cap of a motor which is detected by comparing normal and abnormal thermographic images. 2



(a) (b) (c)

(a) motor

(b) normal thermographic image

(c) abnormal thermographic image

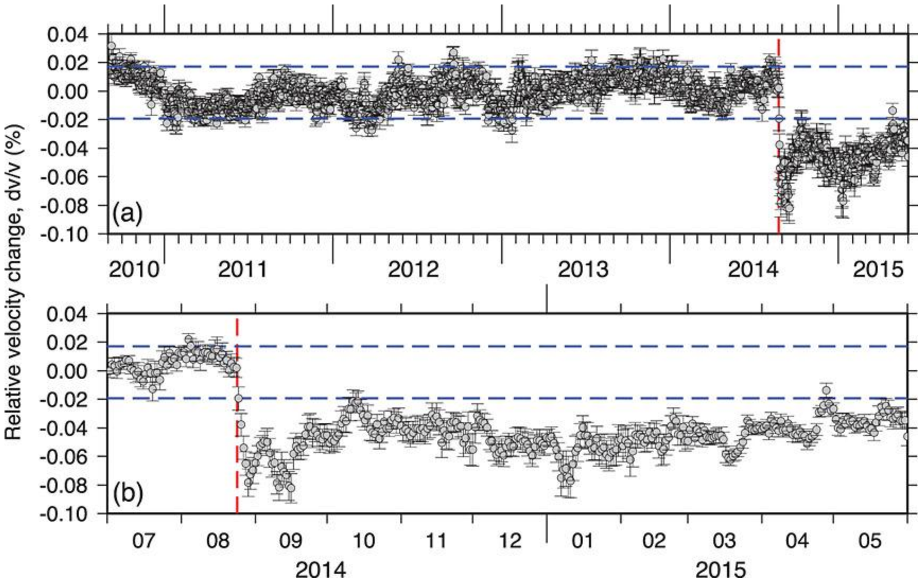
In order to develop machine learning based classification model, which features will you extract from these images so as to depict difference between two.

- Machine Learning algorithms see any images in the form of a matrix of numbers. The size of this matrix actually depends on the number of pixels of the input image.

- The Pixel Values for each of the pixels stands for or describes how bright that pixel is, and what colour it should be. So In the simplest case of the binary images, the pixel value is a 1-bit number indicating either foreground or background.
- So pixels are the numbers or the pixel values which denote the intensity or brightness of the pixel.
- Smaller numbers that are closer to zero helps to represent black, and the larger numbers which are closer to 255 denote white.
- So this is the concept of pixels and how the machine sees the images without eyes through the numbers.
- For this case of a coloured image, we have three Matrices or the channels: Red, Green and Blue. So in these three matrices, each of the matrices has values between 0-255 which represents the intensity of the colour of that pixel.

OR

6 Following figure represent time history of relative velocity change d*v*/*v* with two sigma 2
standard deviations for stack of 5 days in the time intervals (a) September 2010 through May 2015 and (b) July 2014 through May 2015. Red dashed line is the occurrence time of the 2014 M_w 6.0 South Napa earthquake. Blue dashed lines indicate the 95th percentile range of d*v*/*v* distribution obtained from the time interval 1 September 2010 through 23 August 2014.



In order to develop a machine learning-based prediction model, how do standard deviation and 95th percentile range help?

- A standard deviation is a measure of how dispersed the data is in relation to the mean. Low standard deviation means data are clustered around the mean, and high standard deviation indicates data are more spread out. It can be observed that two sigma standard deviations for stack of 5 days in the time intervals (a) September 2010 through May 2015 and (b) July 2014 through May 2015 is able to identify change significantly.

- A 95th percentile says that 95% of the time data points are below that value and 5% of the time they are above that value. In this example, the 95th percentile range of dv/v distribution obtained from the time interval 1 September 2010 through 23 August 2014 able to find clear variation between two distributions.

5 Consider the two-dimensional fluid flow patterns 6

(2, 1), (3, 5), (4, 3), (5, 6), (6, 7), (7, 8)

Compute the principal component using PCA Algorithm.

The given feature vectors are,

$$\begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 3 \\ 5 \end{bmatrix} \begin{bmatrix} 4 \\ 3 \end{bmatrix} \begin{bmatrix} 5 \\ 6 \end{bmatrix} \begin{bmatrix} 6 \\ 7 \end{bmatrix} \begin{bmatrix} 7 \\ 8 \end{bmatrix}$$

Step-02:

Calculate the mean vector (μ).

Mean vector (μ)

= ((2 + 3 + 4 + 5 + 6 + 7) / 6, (1 + 5 + 3 + 6 + 7 + 8) / 6)

= (4.5, 5)

Thus,

$$\text{Mean vector } (\mu) = \begin{bmatrix} 4.5 \\ 5 \end{bmatrix}$$

Step-03:

Subtract mean vector (μ) from the given feature vectors.

- $x_1 - \mu = (2 - 4.5, 1 - 5) = (-2.5, -4)$
- $x_2 - \mu = (3 - 4.5, 5 - 5) = (-1.5, 0)$
- $x_3 - \mu = (4 - 4.5, 3 - 5) = (-0.5, -2)$
- $x_4 - \mu = (5 - 4.5, 6 - 5) = (0.5, 1)$
- $x_5 - \mu = (6 - 4.5, 7 - 5) = (1.5, 2)$
- $x_6 - \mu = (7 - 4.5, 8 - 5) = (2.5, 3)$

Feature vectors (x_i) after subtracting mean vector (μ) are-

$$\begin{bmatrix} -2.5 \\ -4 \end{bmatrix} \begin{bmatrix} -1.5 \\ 0 \end{bmatrix} \begin{bmatrix} -0.5 \\ -2 \end{bmatrix} \begin{bmatrix} 0.5 \\ 1 \end{bmatrix} \begin{bmatrix} 1.5 \\ 2 \end{bmatrix} \begin{bmatrix} 2.5 \\ 3 \end{bmatrix}$$

Step-04:

Calculate the covariance matrix.

Covariance matrix is given by-

$$\text{Covariance Matrix} = \frac{\Sigma\,(x_i - \mu)(x_i - \mu)^t}{n}$$

Now,

$$m_1 = (x_1 - \mu)(x_1 - \mu)^t = \begin{bmatrix} -2.5 \\ -4 \end{bmatrix} \begin{bmatrix} -2.5 & -4 \end{bmatrix} = \begin{bmatrix} 6.25 & 10 \\ 10 & 16 \end{bmatrix}$$

$$m_2 = (x_2 - \mu)(x_2 - \mu)^t = \begin{bmatrix} -1.5 \\ 0 \end{bmatrix} \begin{bmatrix} -1.5 & 0 \end{bmatrix} = \begin{bmatrix} 2.25 & 0 \\ 0 & 0 \end{bmatrix}$$

$$m_3 = (x_3 - \mu)(x_3 - \mu)^t = \begin{bmatrix} -0.5 \\ -2 \end{bmatrix} \begin{bmatrix} -0.5 & -2 \end{bmatrix} = \begin{bmatrix} 0.25 & 1 \\ 1 & 4 \end{bmatrix}$$

$$m_4 = (x_4 - \mu)(x_4 - \mu)^t = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix} \begin{bmatrix} 0.5 & 1 \end{bmatrix} = \begin{bmatrix} 0.25 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

$$m_5 = (x_5 - \mu)(x_5 - \mu)^t = \begin{bmatrix} 1.5 \\ 2 \end{bmatrix} \begin{bmatrix} 1.5 & 2 \end{bmatrix} = \begin{bmatrix} 2.25 & 3 \\ 3 & 4 \end{bmatrix}$$

$$m_6 = (x_6 - \mu)(x_6 - \mu)^t = \begin{bmatrix} 2.5 \\ 3 \end{bmatrix} \begin{bmatrix} 2.5 & 3 \end{bmatrix} = \begin{bmatrix} 6.25 & 7.5 \\ 7.5 & 9 \end{bmatrix}$$

Now,

Covariance matrix

$= (m_1 + m_2 + m_3 + m_4 + m_5 + m_6) / 6$

On adding the above matrices and dividing by 6, we get

$$\text{Covariance Matrix} = \frac{1}{6} \begin{bmatrix} 17.5 & 22 \\ 22 & 34 \end{bmatrix}$$

$$\text{Covariance Matrix} = \begin{bmatrix} 2.92 & 3.67 \\ 3.67 & 5.67 \end{bmatrix}$$

Step-05:

Calculate the eigen values and eigen vectors of the covariance matrix.

λ is an eigen value for a matrix M if it is a solution of the characteristic equation $|M - \lambda I|$

= 0. So, we have

$$\begin{vmatrix} 2.92 & 3.67 \\ 3.67 & 5.67 \end{vmatrix} - \begin{vmatrix} \lambda & 0 \\ 0 & \lambda \end{vmatrix} = 0$$

$$\begin{vmatrix} 2.92 - \lambda & 3.67 \\ 3.67 & 5.67 - \lambda \end{vmatrix} = 0$$

From here,

$(2.92 – \lambda)(5.67 – \lambda) – (3.67 \text{ x } 3.67) = 0$

$16.56 – 2.92\lambda – 5.67\lambda + \lambda^2 – 13.47 = 0$

$\lambda^2 – 8.59\lambda + 3.09 = 0$

Solving this quadratic equation, we get $\lambda = 8.22, 0.38$

Thus, two eigen values are $\lambda_1 = 8.22$ and $\lambda_2 = 0.38$.

Clearly, the second eigen value is very small compared to the first eigen value.

So, the second eigen vector can be left out.

Eigen vector corresponding to the greatest eigen value is the principal component for the given data set.

So. we find the eigen vector corresponding to eigen value λ_1.

We use the following equation to find the eigen vector-

$MX = \lambda X$

where

- M = Covariance Matrix
- X = Eigen vector
- λ = Eigen value

Substituting the values in the above equation, we get-

$$\begin{bmatrix} 2.92 & 3.67 \\ 3.67 & 5.67 \end{bmatrix} \begin{bmatrix} X1 \\ X2 \end{bmatrix} = 8.22 \begin{bmatrix} X1 \\ X2 \end{bmatrix}$$

Solving these, we get-

$2.92X_1 + 3.67X_2 = 8.22X_1$

$3.67X_1 + 5.67X_2 = 8.22X_2$

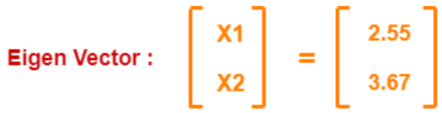
On simplification, we get-

$5.3X_1 = 3.67X_2$ ………(1)

$3.67X_1 = 2.55X_2$ ………(2)

From (1) and (2), $\mathbf{X_1 = 0.69X_2}$

From (2), the eigen vector is-



Eigen Vector : $\begin{bmatrix} X1 \\ X2 \end{bmatrix} = \begin{bmatrix} 2.55 \\ 3.67 \end{bmatrix}$

Thus, principal component for the given data set is-

Principal Component : $\begin{bmatrix} X1 \\ X2 \end{bmatrix} = \begin{bmatrix} 2.55 \\ 3.67 \end{bmatrix}$

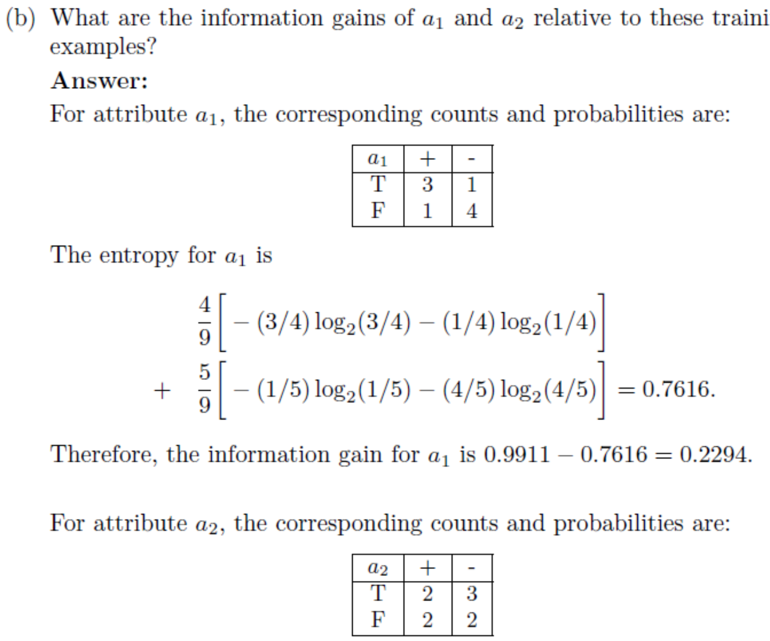
6 Consider the training examples shown in following table below for a binary classification problem. 6

Instances	α_1	α_2	α_3	Target class (condition of bearing)
1	T	T	1	Cage fault
2	T	T	6	Cage fault
3	T	F	5	Ball fault
4	F	F	4	Cage fault
5	F	T	7	Ball fault
6	F	T	3	Ball fault
7	F	F	8	Ball fault
8	T	F	7	Cage fault
9	F	T	5	Ball fault

a. What are the information gains of α_1 and α_2 relative to these training examples?

b. For α_3 which is a continuous attribute, compute the information gain for every possible split.

c. What is the best split (among α_1, α_2 and α_3) according to the information gain?



(b) What are the information gains of a_1 and a_2 relative to these training examples?

Answer:

For attribute a_1, the corresponding counts and probabilities are:

a_1	+	-
T	3	1
F	1	4

The entropy for a_1 is

$$\frac{4}{9}\Big[-(3/4)\log_2(3/4)-(1/4)\log_2(1/4)\Big] + \frac{5}{9}\Big[-(1/5)\log_2(1/5)-(4/5)\log_2(4/5)\Big] = 0.7616.$$

Therefore, the information gain for a_1 is $0.9911 - 0.7616 = 0.2294$.

For attribute a_2, the corresponding counts and probabilities are:

a_2	+	-
T	2	3
F	2	2

The entropy for a_2 is

$$\begin{aligned} & \frac{5}{9}\Big[-(2/5)\log_2(2/5) - (3/5)\log_2(3/5) \Big] \\ + & \frac{4}{9}\Big[-(2/4)\log_2(2/4) - (2/4)\log_2(2/4) \Big] = 0.9839. \end{aligned}$$

Therefore, the information gain for a_2 is $0.9911 - 0.9839 = 0.0072$.

(c) For a_3, which is a continuous attribute, compute the information gain for every possible split.

Answer:

a_3	Class label	Split point	Entropy	Info Gain
1.0	+	2.0	0.8484	0.1427
3.0	-	3.5	0.9885	0.0026
4.0	+	4.5	0.9183	0.0728
5.0	-			
5.0	-	5.5	0.9839	0.0072
6.0	+	6.5	0.9728	0.0183
7.0	+			
7.0	-	7.5	0.8889	0.1022

The best split for a_3 occurs at split point equals to 2.

(d) What is the best split (among a_1, a_2, and a_3) according to the information gain?

Answer:

According to information gain, a_1 produces the best split.

****