Smarter education for smarter planet

IBM Innovation Centre for Education

Pattern and Anomaly Detection Introduction

Unit objectives

After completing this unit, you should be able to:

- Understand the concept of pattern recognition and anomaly detection
- Gain knowledge on example of polynomial curve fitting
- Learn about probability theory architecture and working model
- Understand Information theory

What is pattern?

- Pattern is all about it in this digital age.
- A pattern can be either visually identified or mathematically detected via the implementation of algorithms.

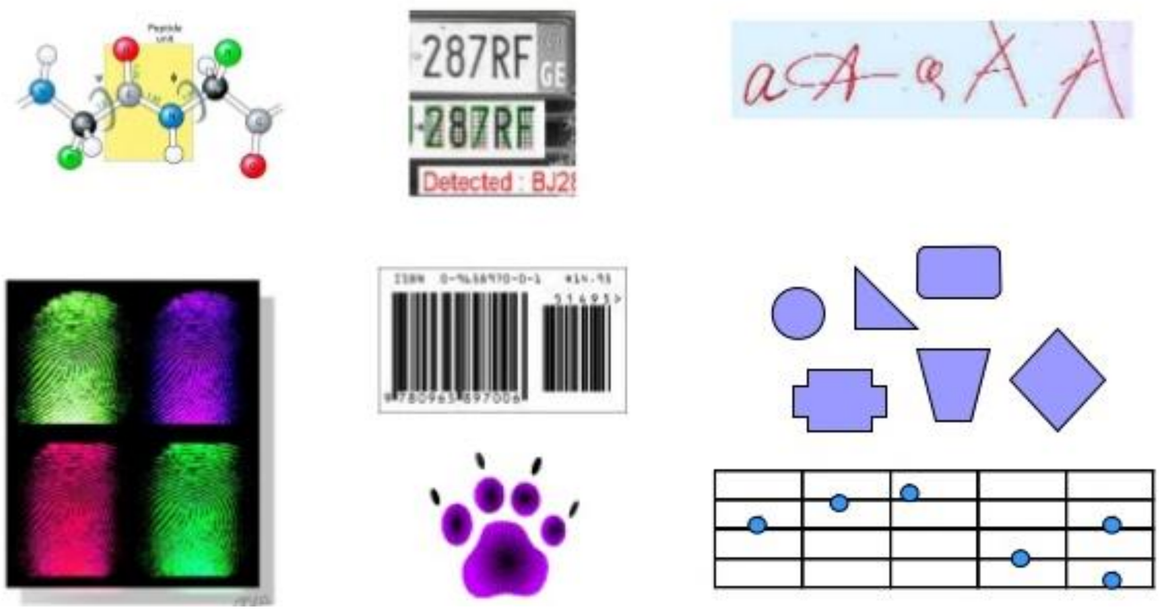


Figure: Pattern definition

Source: https://images.app.goo.gl/NmRdihnyFymA23uRA

What is pattern recognition?

- As per Wikipedia, pattern recognition is the automated recognition of patterns and regularities in data.

- It has applications in statistical data analysis, signal processing, image analysis, information retrieval, bioinformatics, data compression, computer graphics and machine learning.

Pattern recognition techniques

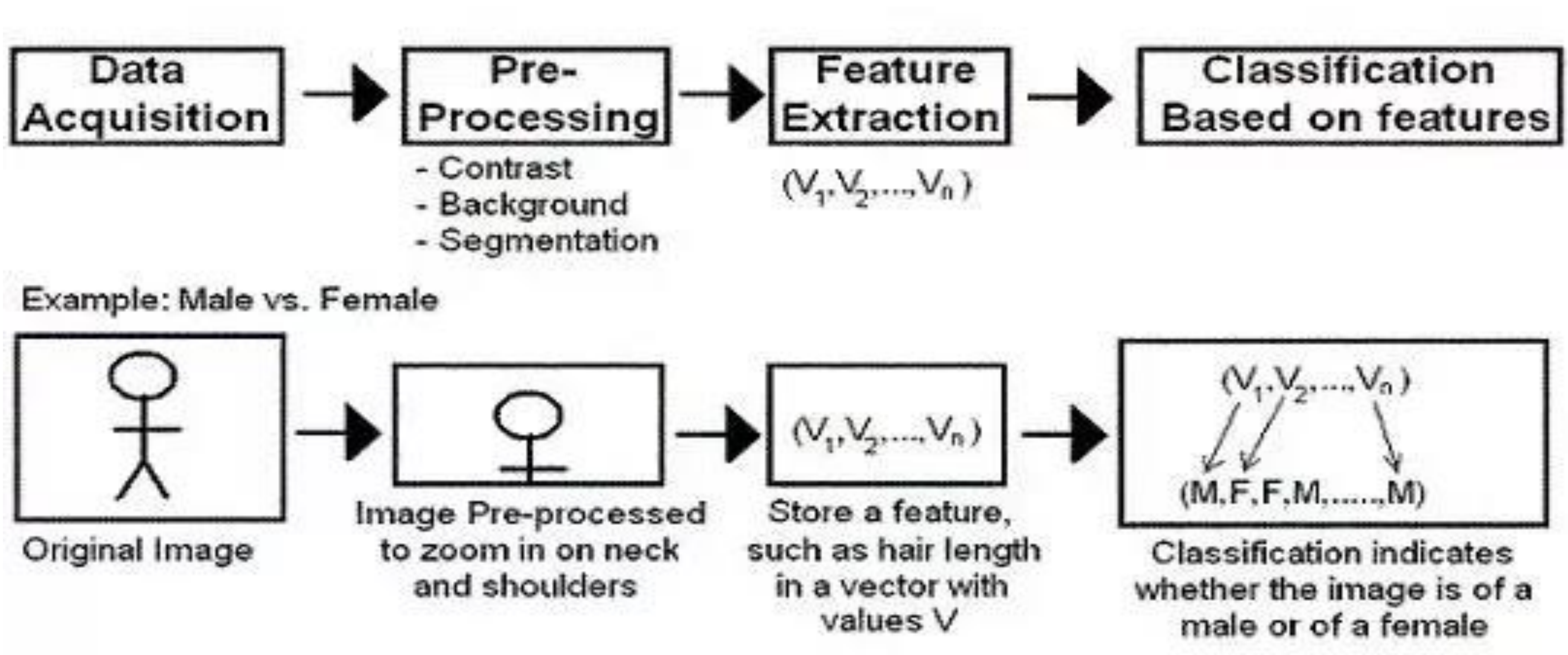


Figure: Pattern recognition process

Source: https://images.app.goo.gl/3x6gZVVao9u3vV8w7

Training and learning in pattern recognition

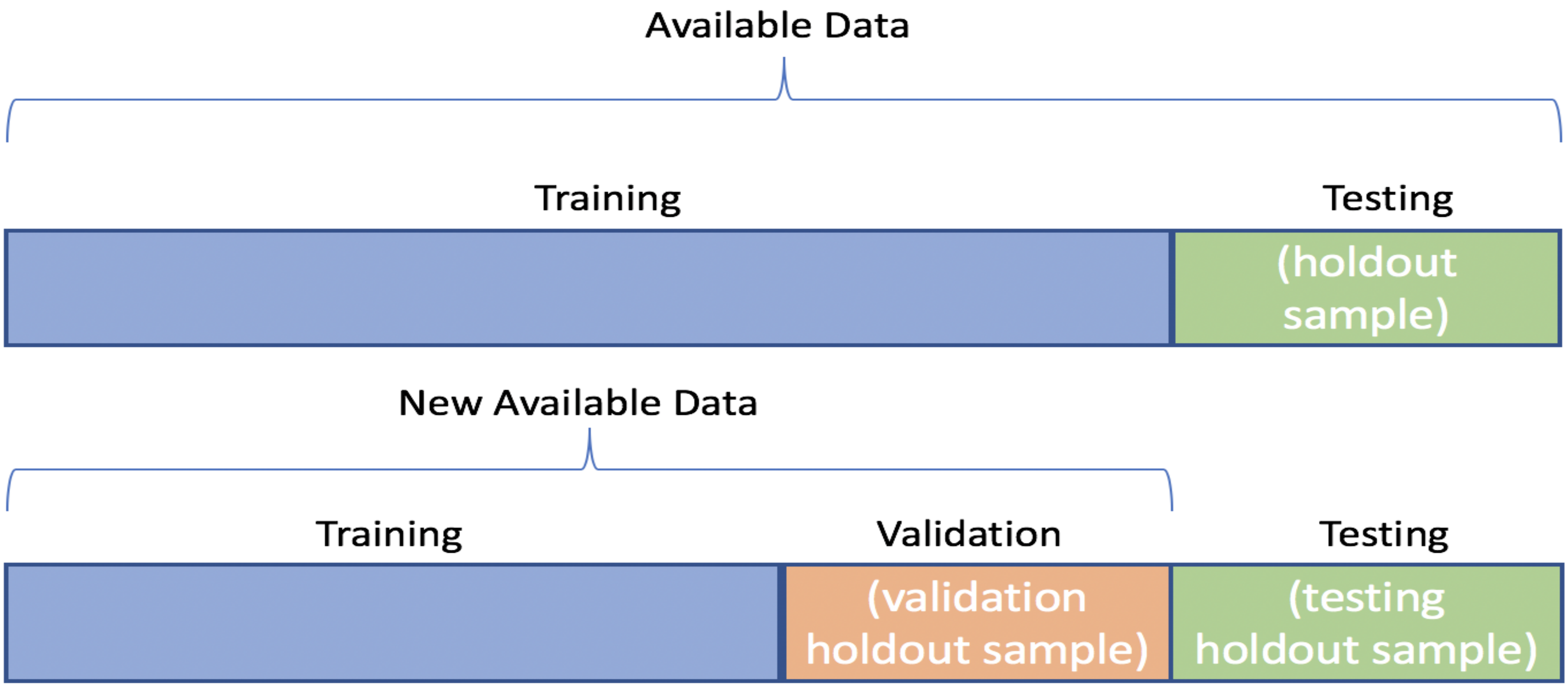


Figure: Training and testing dataset

Source: https://images.app.goo.gl/vetePfKYS2t7vmX99

Pattern recognition applications

- A model is an ideal concrete entity, or theoretical idea. The definition of the animal is an illustration when thinking regarding animal groups.
- The definition of the ball is a trend while speaking about various styles about balls.
- The groups can be baseball, cricket game, table tennis match, in sample case balls.
- Before approaching a new species, the species class has to be established.
- Choosing attributes and describing patterns is a very critical phase in classifying the layout.
- Effective presentation requires the use of non-discriminatory attributes, and the sample classification computational pressure.

Pattern recognition use cases

- Customer research and stock market analysis.
- Chat bots, NLP with text generation, text analysis, text translation.
- Optical Character Recognition (OCR), document classification and signature verification.
- Image recognition, visual search, face recognition.
- Voice recognition and ai assistants.
- Recommendation sentiment analysis, audience research.

What is anomaly detection?

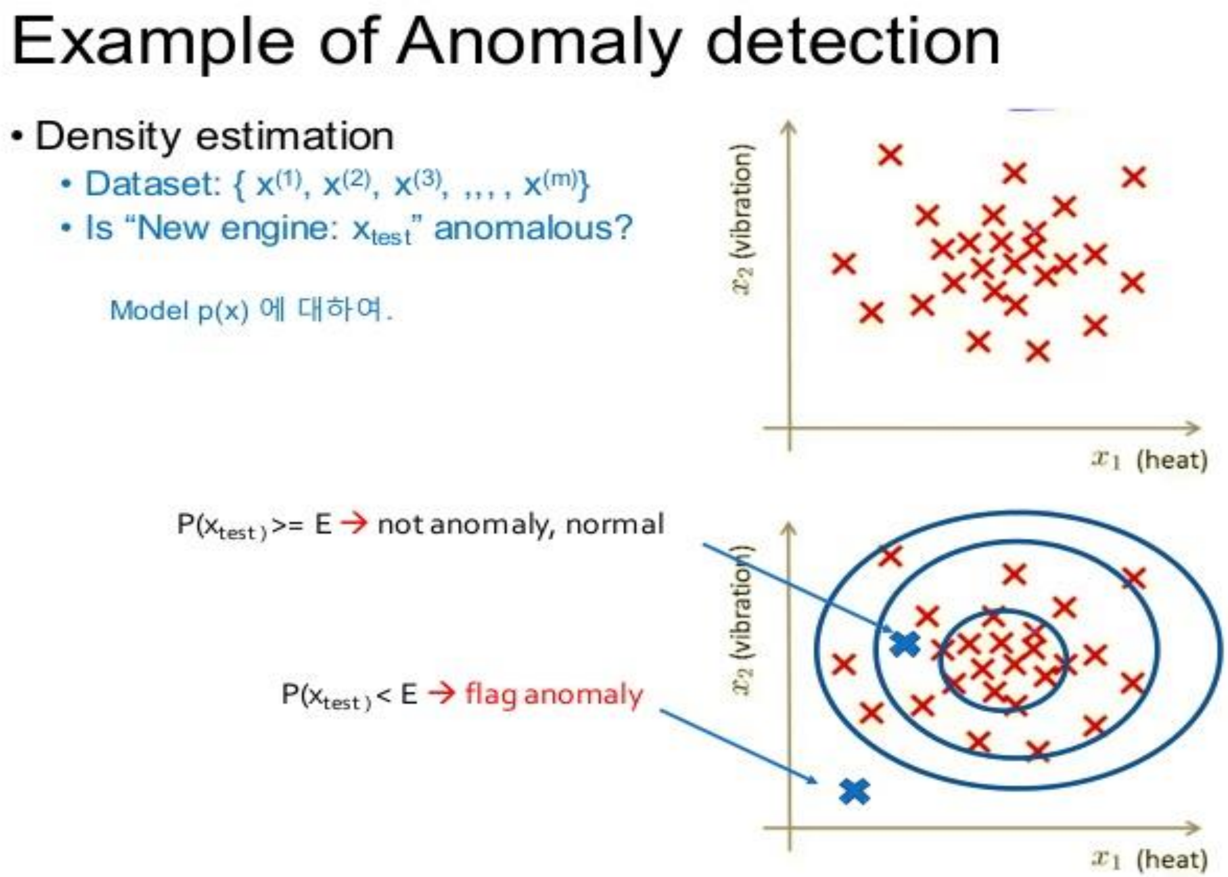


Figure: Anomaly detection example

Source: https://images.app.goo.gl/WpR4Rk1Xth1Fj4rt6

What are some other practical uses for anomaly detection?

- Traffic dropped or spiked.
- Transactions or revenue dropped.
- Traffic from social media increased or decreased.
- Traffic from organic search increased or decreased.

How is anomaly detection calculated over time?

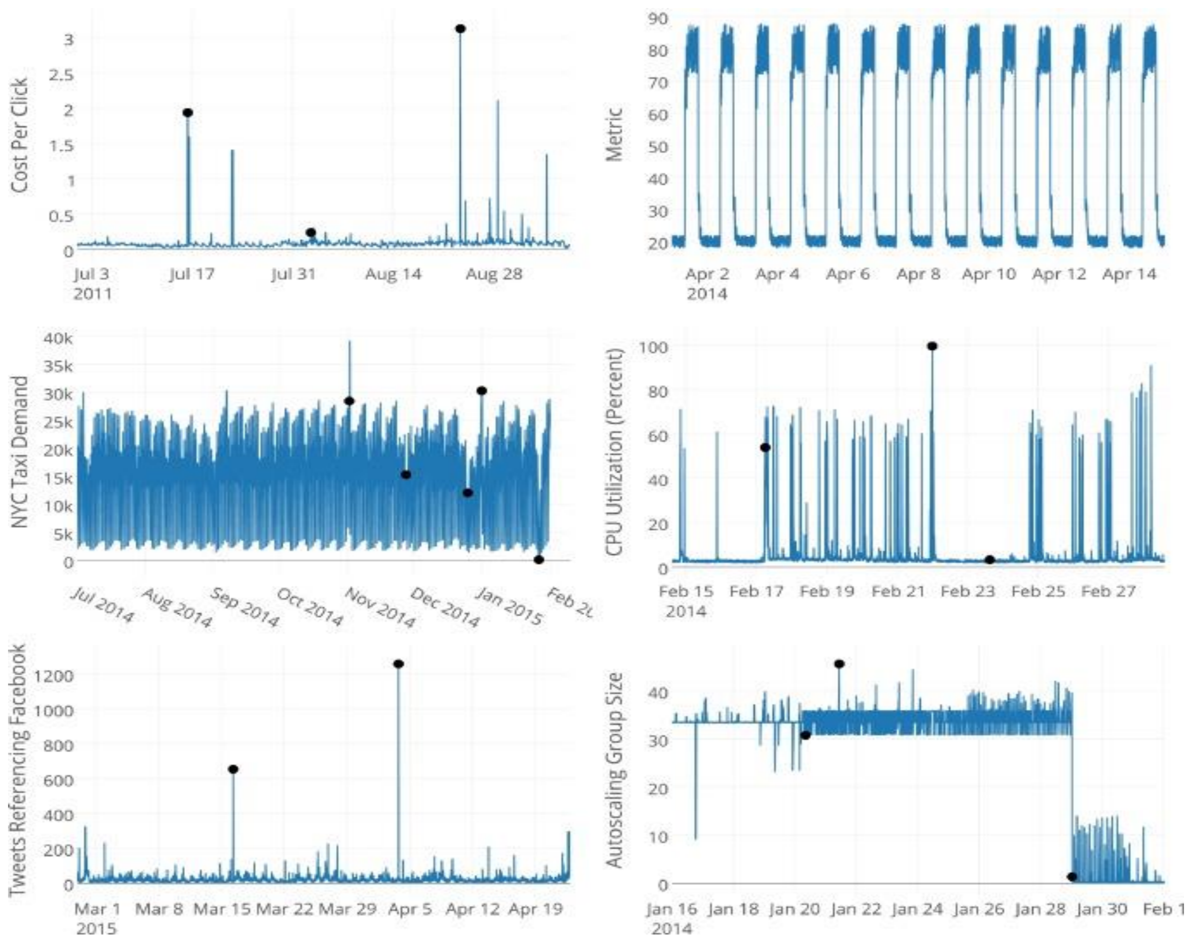


Figure: CPU time frame for anomaly detection

Source: https://images.app.goo.gl/a5wk76ZUmKZLj63v6

Self evaluation: Exercise 1

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 1: Polynomial curve fitting.

Key point for AI and ML anomaly detection

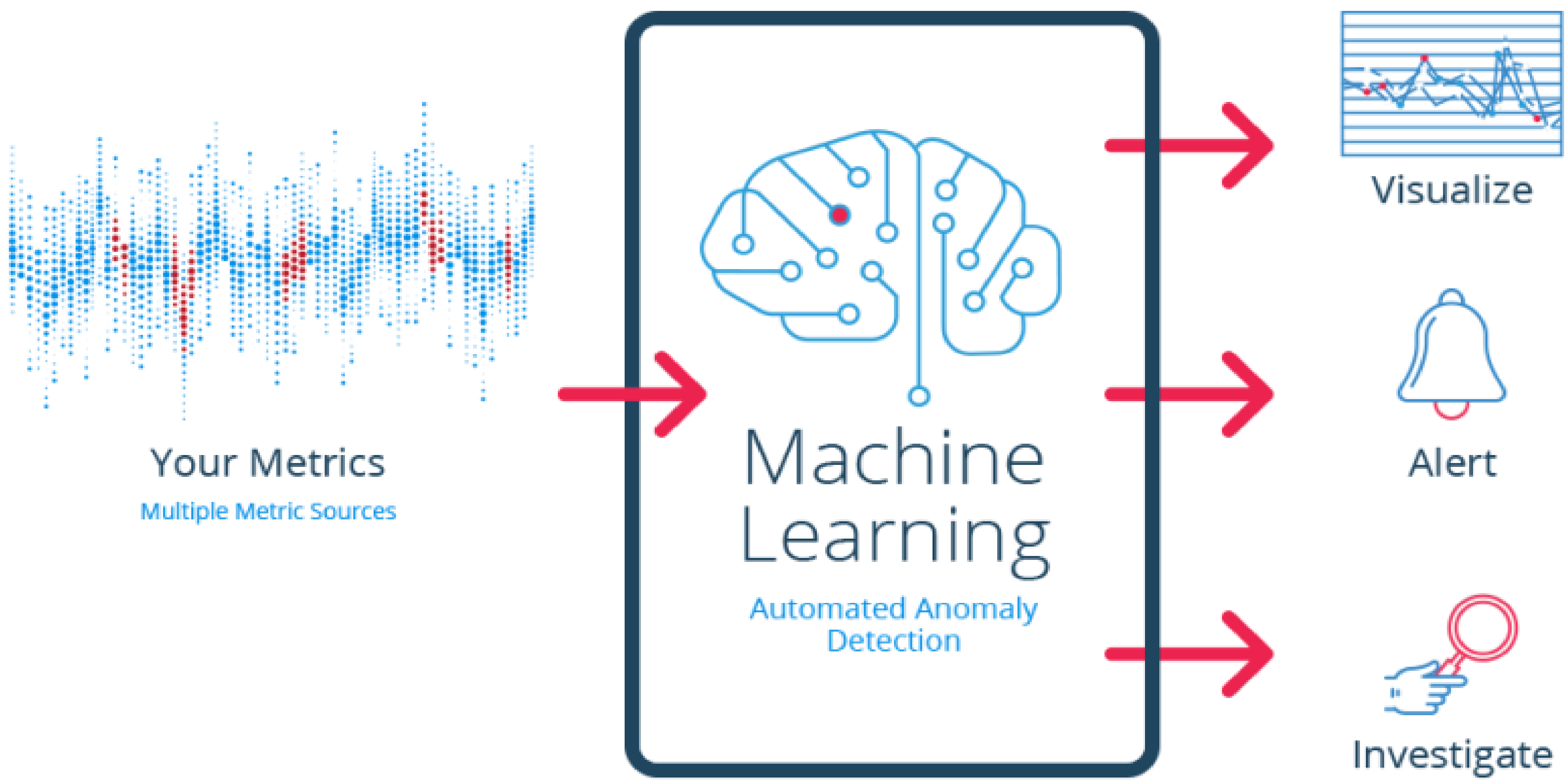


Figure: Anomaly detection flow with Machine learning

Source: https://images.app.goo.gl/ewiV18h8cpAkEyCL9

Tasks for artificial intelligence

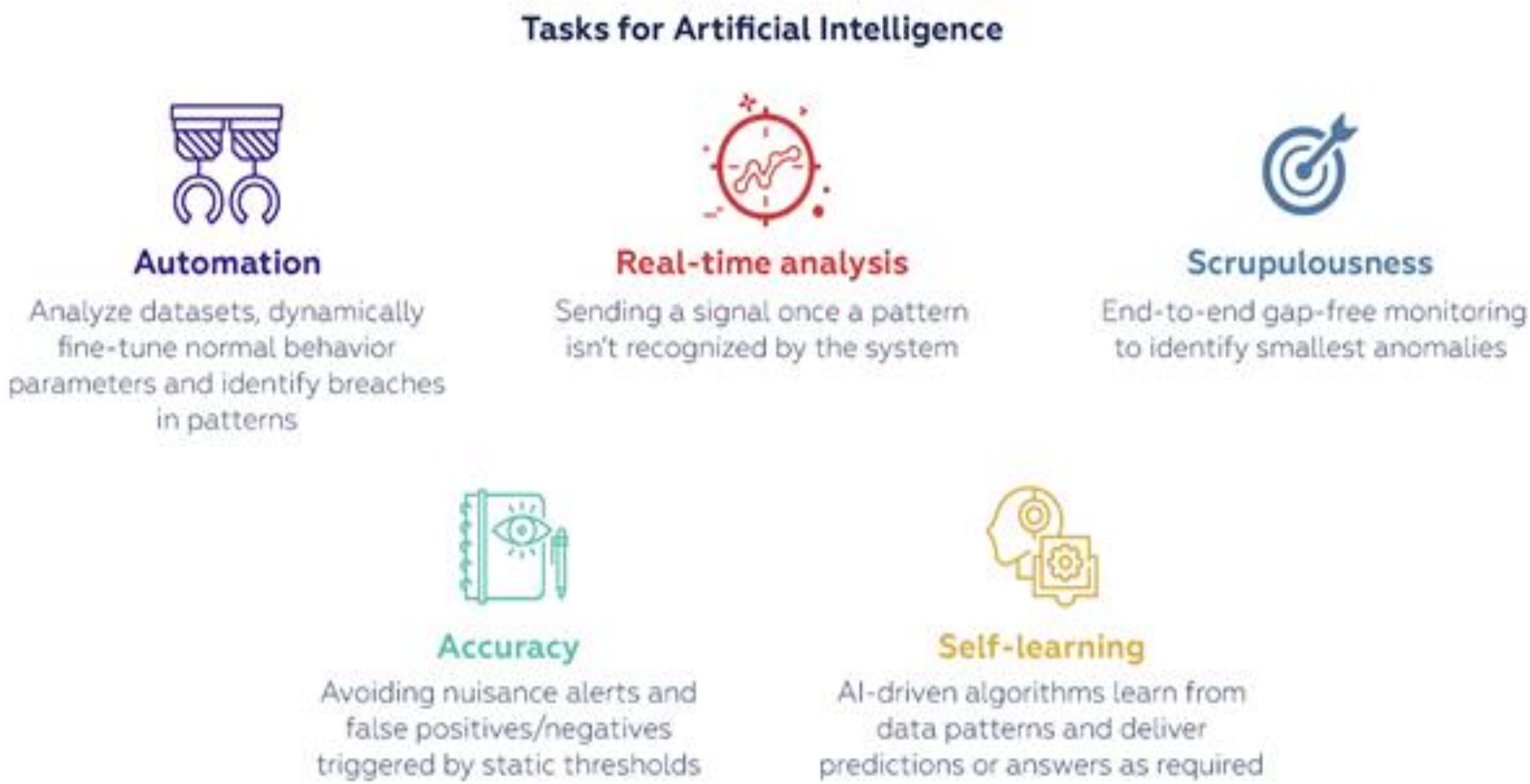


Figure: AI Task

Source: https://images.app.goo.gl/F9Z664sSWR53yGqLA

AI system learning process (1 of 2)

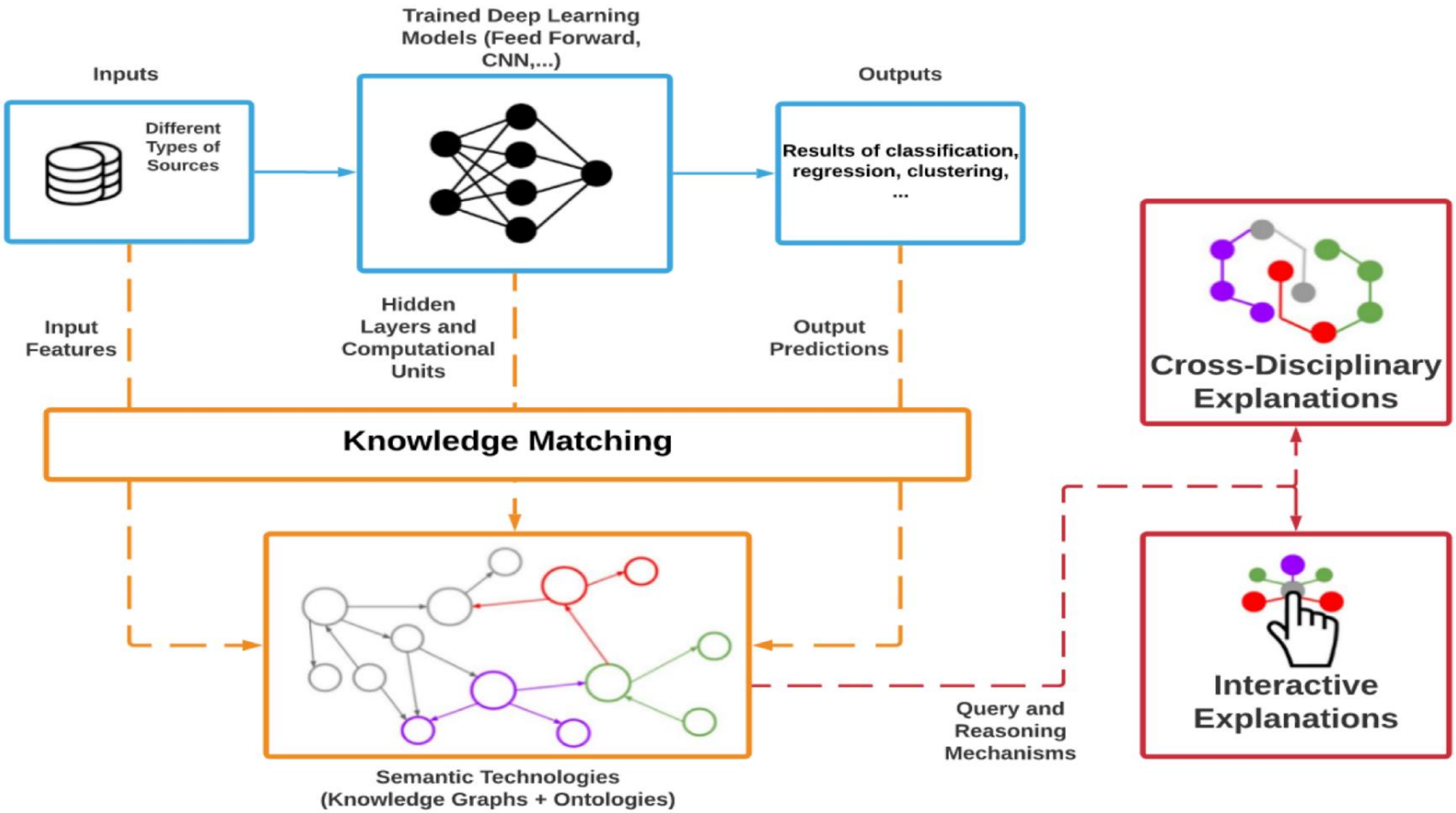


Figure: AI Learning Process

Source: https://images.app.goo.gl/dBtnk7CPRtqe7xM59

AI system learning process (2 of 2)

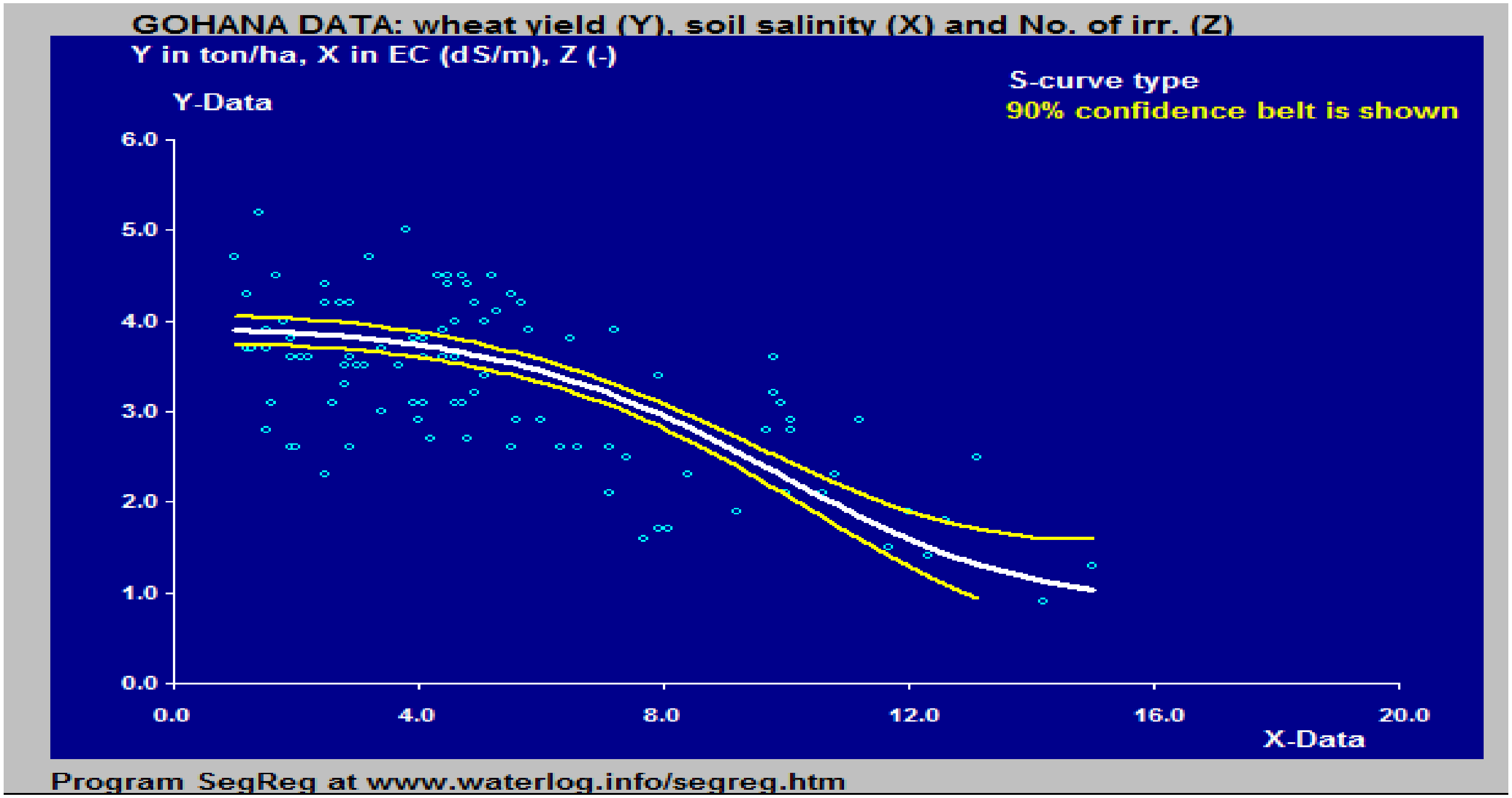


Figure: Relation between wheat yield and soil salinity

Source: https://upload.wikimedia.org/wikipedia/commons/thumb/4/46/Gohana_inverted_S-curve.png/560px-Gohana_inverted_S-curve.png

Self evaluation: Exercise 2

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 2: Probability and distribution.

Test to geometric requirements for curves algebraic

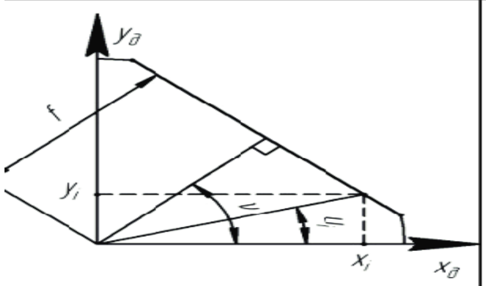
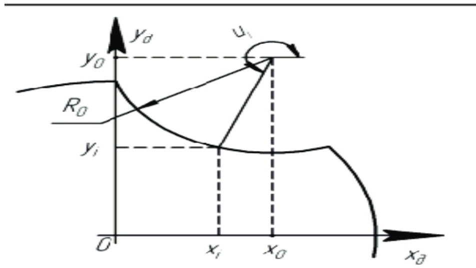
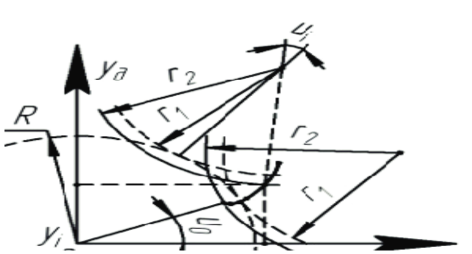
Scheme of curve plotting	Equation of a curve in a parametrical look
1	2
	Straight line: $$x_i = \frac{f}{\cos(u_i - \nu)} \cdot \cos(u_i)$$ $$y_i = \frac{f}{\cos(u_i - \nu)} \cdot \sin(u_i)$$
	Circle arch: $$x_i = x_0 + R_0 \times \cos(u_i)$$ $$y_i = y_0 + R_0 \times \sin(u_i)$$
	Epicycloid: $$x_i = (R + r_1)\cos(u_i + u_0) -$$ $$-r_2 \cos\left[\left(\frac{R + r_1}{r_1}\right)(u_i + u_0)\right]$$ $$y_i = (R + r_1)\sin(u_i + u_0) -$$ $$-r_2 \sin\left[\left(\frac{R + r_1}{r_1}\right)(u_i + u_0)\right]$$

Figure: Algebraic curves in a parametrical form used for creation of the forming line in the edge section of different rated surfaces

Source: https://images.app.goo.gl/Rn75dzFLCv9y6nu5A

Curves matched to data points (1 of 2)

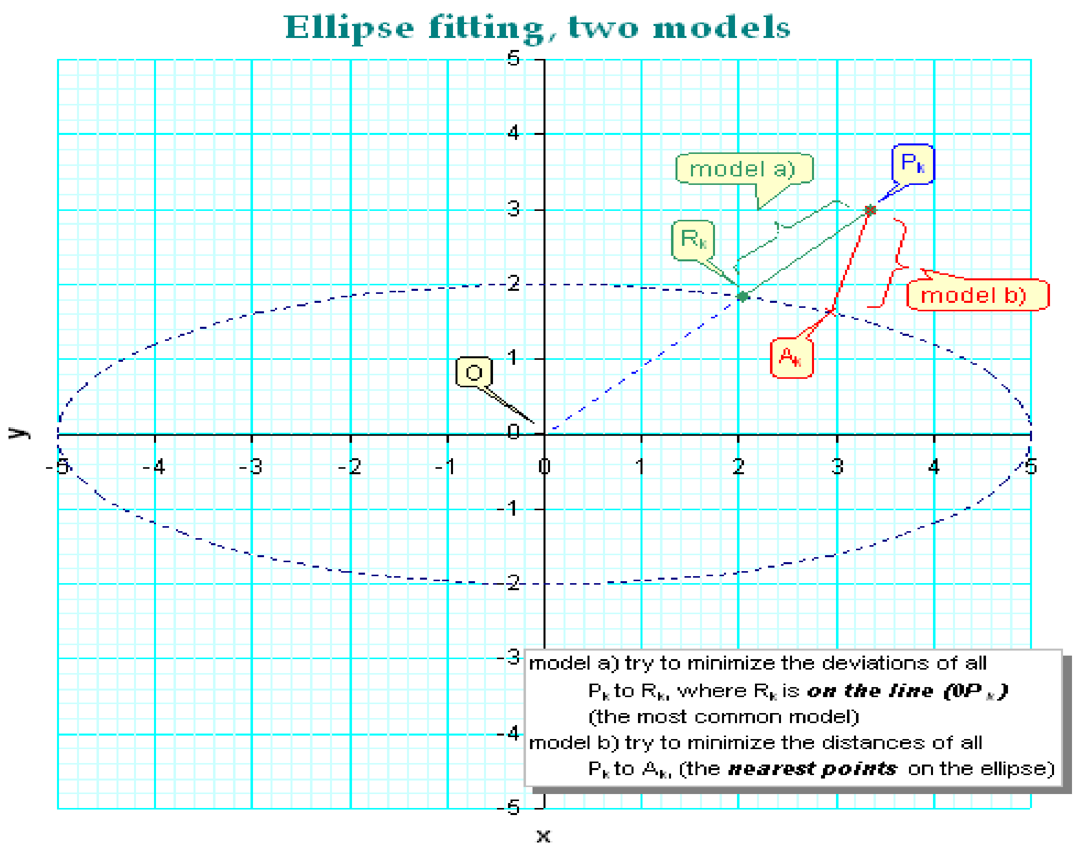


Figure: Different models of ellipse fitting

Source: https://lh3.googleusercontent.com/HkRug5Yd6SlGy0AkSgLZ9FYwrq3Os5jeSoEiHqg5ft1se9C8uSUcXjY9p3yfYfhg13eyUA=s86

Curves matched to data points (2 of 2)

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.977411977					
R Square	0.955334173					
Adjusted R Square	0.952025593					
Standard Error	13.21744878					
Observations	30					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	2	100887.8743	50443.93714	288.7444885	5.95206E-19	
Residual	27	4716.925714	174.7009524			
Total	29	105604.8				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	21.92	10.57395903	2.073017301	0.047839767	0.224028187	43.61597181
Month	-24.5485714	6.91777293	-3.54862348	0.001441196	-38.742669	-10.3544738
MonSq	8.057142857	0.967418193	8.328500456	6.14519E-09	6.072164688	10.04212103

Figure: Linear regression output

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.91683485					
R Square	0.840586142					
Adjusted R Square	0.83489279					
Standard Error	24.52030396					
Observations	30					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	88769.93143	88769.93143	147.6434502	1.1088E-12	
Residual	28	16834.86857	601.2453061			
Total	29	105604.8				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-53.28	10.20861661	-5.21912048	1.52374E-05	-74.1914032	-32.3685968
Month	31.85142857	2.621330755	12.15086212	1.1088E-12	26.48187593	37.22098121

Figure: Quadratic regression output

Case study: Anomaly detection with IBM Watson

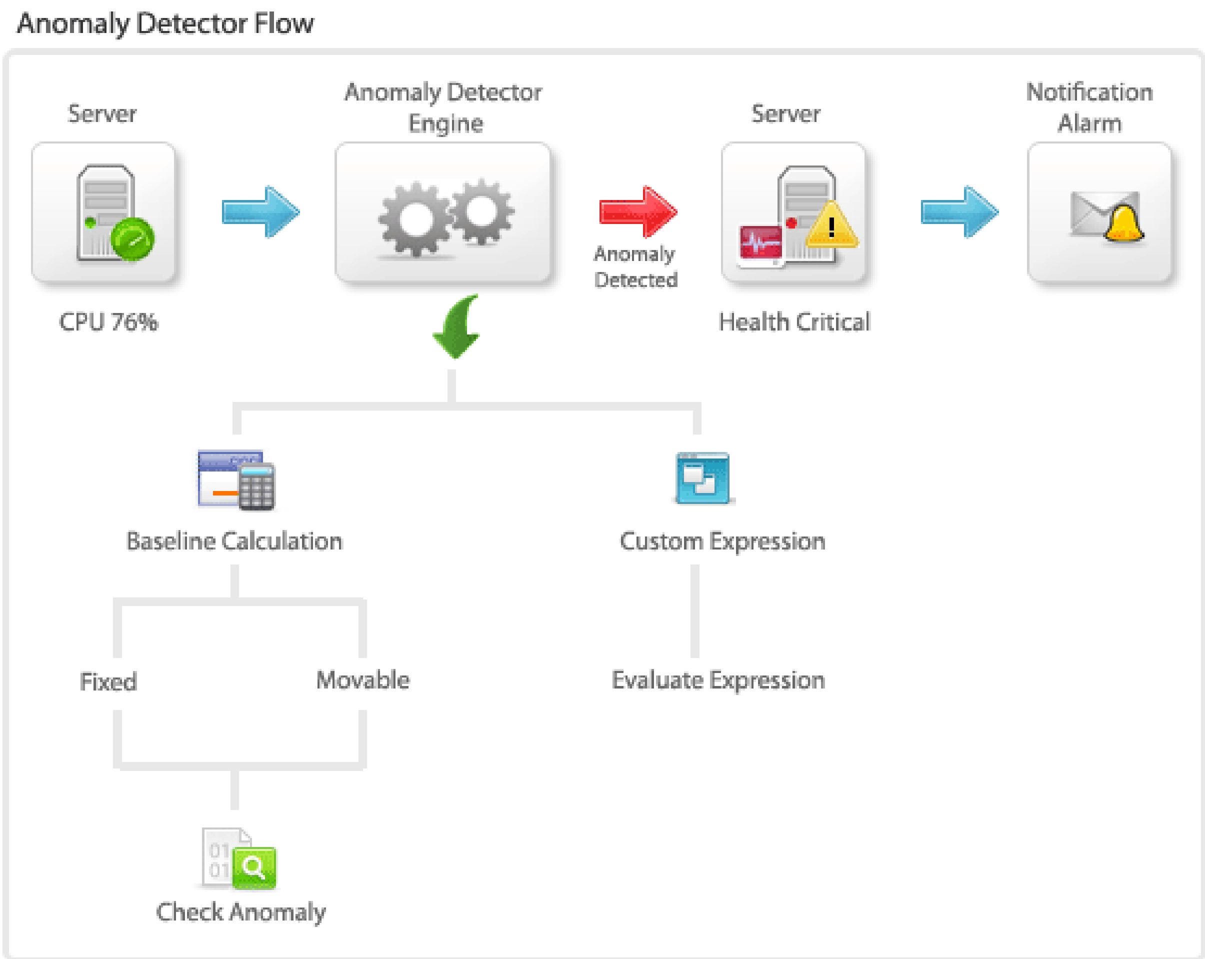


Figure: Anomaly detection workflow engine

Source: https://images.app.goo.gl/ukrNQnHbjnP5XXKf6

Self evaluation: Exercise 3

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 3: Simple linear regression.

Probability theory (1 of 2)

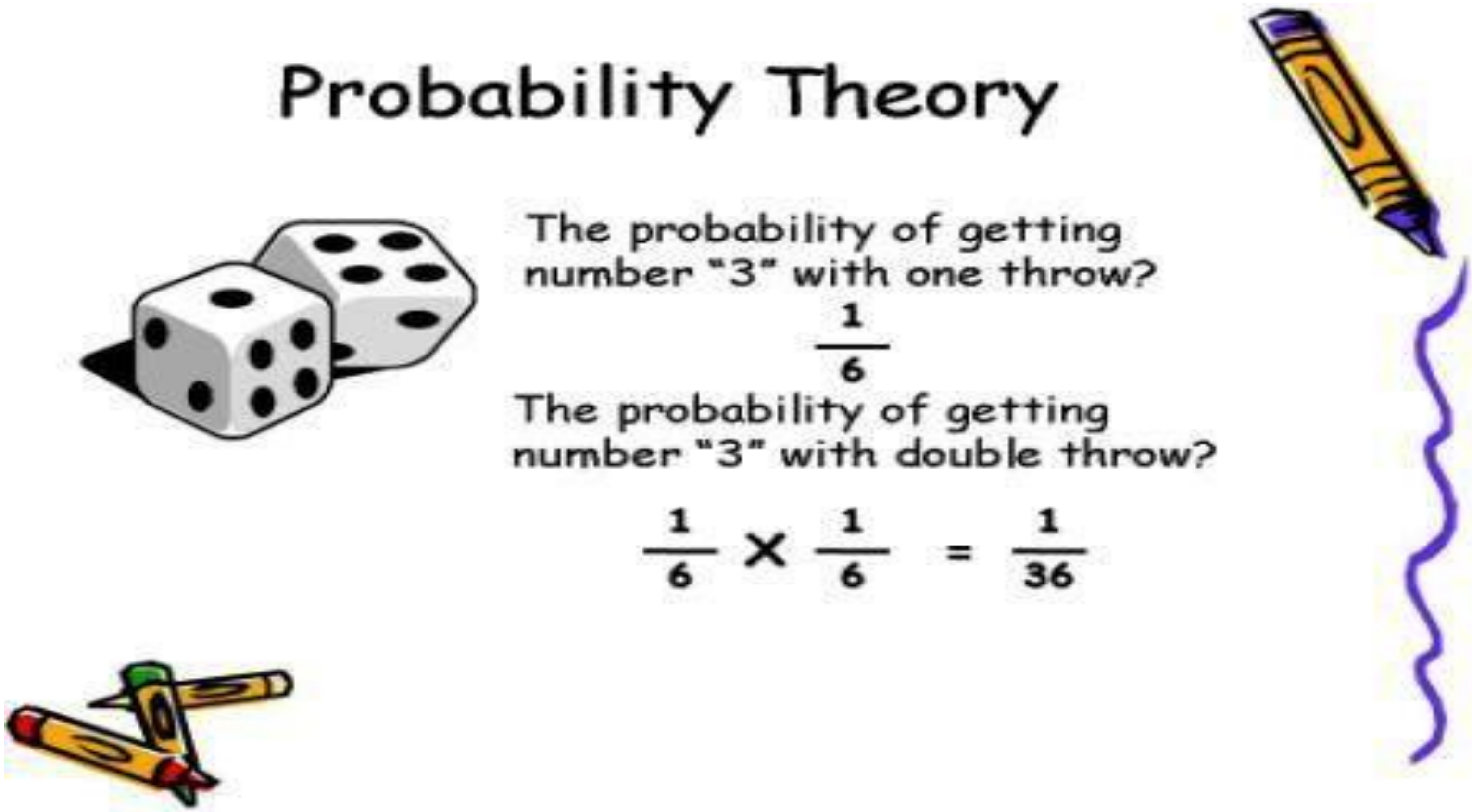


Figure: Probability Theory

Source: https://images.app.goo.gl/DKudJzQZCPZEQyvt6

Probability theory (2 of 2)

- Sample Space: 12 There are 12 marbles total (4+5+1+2 = 12)

Probability= Total Possible outcome

- P (black) = 2/12 = 1/6 There are 2 black marbles in the bag, 12 is your sample space.

.

- P (blue) = 4/12 = 1/3 There are 4 blue marbles in the bag , 12 is your sample space.
- P (blue or black) = 6/12= 1/2 4 blue + 2 black = 6 , 12 is your sample space.
- P (not green) = 11/12 There's 1 green. So 12-1 = 11 that aren't green,12 is your sample space.
- P (not purple) = 1
- I will select a marble that is not purple because there are no purple marbles in the bag. Whenever the chance of something occurring is definite, the probability is i.

Maximum likelihood theory and estimation (1 of 2)

- The estimate of density is the issue of evaluating the distribution of likelihood for a sub-set of a sample in a question domain.

- Two Important concepts:
  - Probability density estimation problem.
  - Maximum likelihood calculation.

Maximum likelihood theory and estimation (2 of 2)

- Suppose that we are given a sequence (x1….xn) of IID random variables and a priori distribution of it is given by Us wish to find the MAP estimate of it. Note that the normal distribution is its own conjugate prior, so we will be able to find a closed-form solution analytically.

- The function to be maximized is then given by:

$$f(\mu)f(x \mid \mu) = \pi(\mu)L(\mu) = \frac{1}{\sqrt{2\pi}\sigma_m} \exp\left(-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_m}\right)^2\right) \prod_{j=1}^{n} \frac{1}{\sqrt{2\pi}\sigma_v} \exp\left(-\frac{1}{2}\left(\frac{x_j-\mu}{\sigma_v}\right)^2\right),$$

- Which is equivalent to minimizing the following function of it:

$$\sum_{j=1}^{n}\left(\frac{x_j-\mu}{\sigma_v}\right)^2 + \left(\frac{\mu-\mu_0}{\sigma_m}\right)^2.$$

- Thus, we see that the MAP estimator for p is given by:

$$\hat{\mu}_{\mathrm{MAP}} = \frac{\sigma_m^2\, n}{\sigma_m^2\, n + \sigma_v^2}\left(\frac{1}{n}\sum_{j=1}^{n} x_j\right) + \frac{\sigma_v^2}{\sigma_m^2\, n + \sigma_v^2}\,\mu_0 = \frac{\sigma_m^2\left(\sum_{j=1}^{n} x_j\right) + \sigma_v^2\,\mu_0}{\sigma_m^2\, n + \sigma_v^2}.$$

- Which turns out to be a linear interpolation between the prior mean and the sample mean weighted by their respective covariance's. The case of is called a non-informative prior and leads to an ill-defined a priori probability distribution.

Self evaluation: Exercise 4

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.
- You are instructed to write the following activities using python code.
- Exercise 4: Multiple linear regression.

Model selection (1 of 2)

- The MDL (Minimum Description Length) statistic is calculated as follows:

$$MDL = L(h) + L(D \mid h)$$

- Where h is the model, D is the predictions made by the model, L(h) is the number of bits required to represent the model, and L(D | h) is the number of bits required to represent the predictions from the model on the training dataset.

- The score as defined above is minimized, e.g., the model with the lowest MDL is selected.

- The number of bits required to encode (D | h) and the number of bits required to encode (h) can be calculated as the negative log-likelihood. For example:

$$MDL = -\log(P(theta)) - \log(P(y \mid X, theta))$$

- Or the negative log-likelihood of the model parameters (theta) and the negative log-likelihood of the target values (y) given the input values (X) and the model parameters (theta).

Model selection (2 of 2)

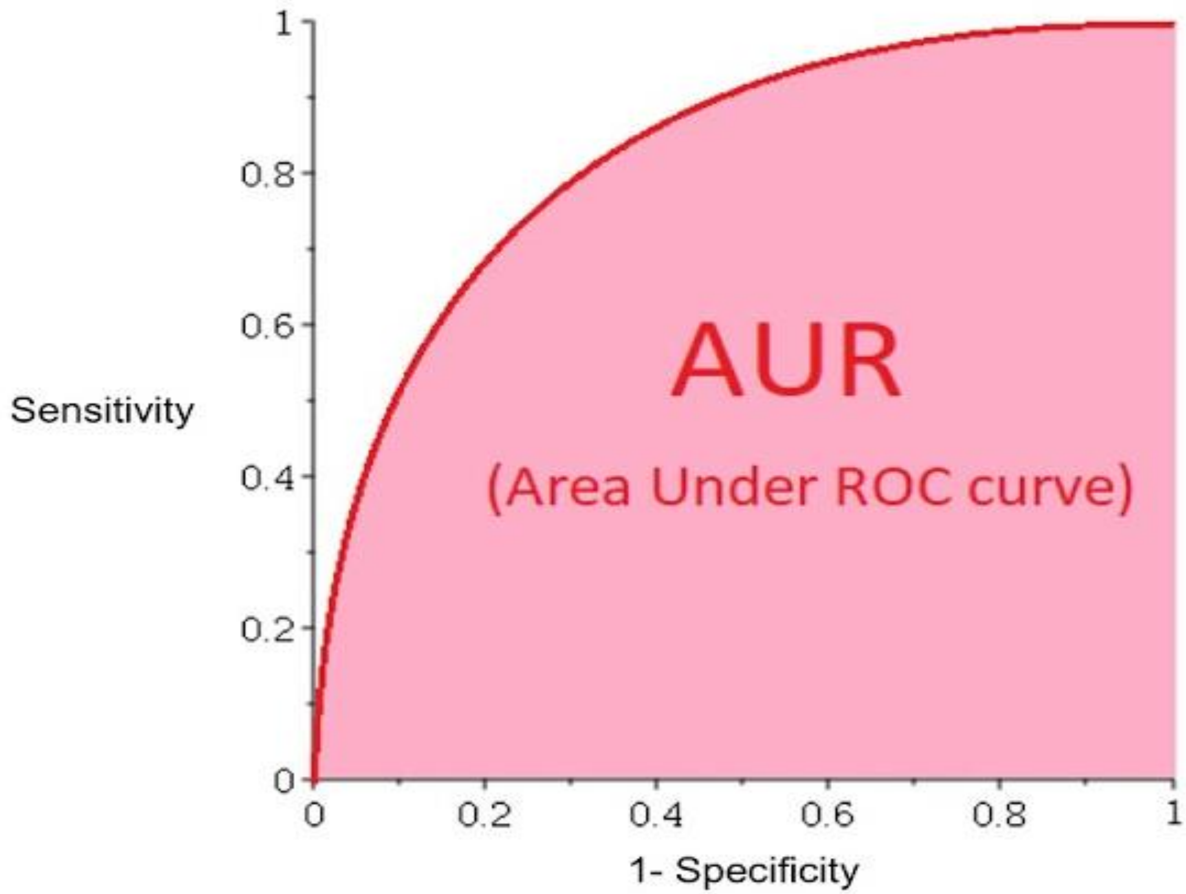


Figure: AU ROC

Source: https://images.app.goo.gl/iMXwj7jLYADko1uCA

Matrices of uncertainty (confusion matrices)

- A few other metrics are computed from these values:

- Accuracy: How often is the classifier correct? $\left(\frac{TP+TN}{total}\right)$
- Misclassification rate (or "error rate"): How often is the classifier wrong? $\left(\frac{FP+FN}{total} = 1 - accuracy\right)$
- Recall (or "sensitivity" or "true positive rate"): How often are positive-labeled samples predicted as positive? $\left(\frac{TP}{\text{num positive-labeled examples}}\right)$
- False positive rate: How often are negative-labeled samples predicted as positive? $\frac{FP}{\text{num negative-labeled examples}})$
- Specificity (or "true negative rate"): How often are negative-labeled samples predicted as negative? $\frac{TN}{\text{num negative-labeled examples}})$
- Precision: How many of the predicted positive samples are correctly predicted? $\left(\frac{TP}{TP+FP}\right)$
- Prevalence: How many labeled-positive samples are there in the data?

$$\left(\frac{\text{num positive-labeled examples}}{\text{num examples}}\right)$$

Loss of logging (log-loss)

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i))$$

Binary Cross-Entropy / Log Loss

Figure: Log Loss formula

Source: https://images.app.goo.gl/UpFaWENnNrSm935R9

Rate for F1 (F1 score)

- The F1 score is the weighted average accuracy and warning, even the balanced F-score or F-measure.

$$F_1 = 2\frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

Figure: F1 score

Metric selection

- Metric selection is more complex for biased groups (or strongly predetermined bias data).
- For example, you have a dataset with only 0.5% of the data in category 1.
- You run your experiment and remember that 99.5 percent of the tests are correctly graded.

Hyperparameter selection (1 of 2)

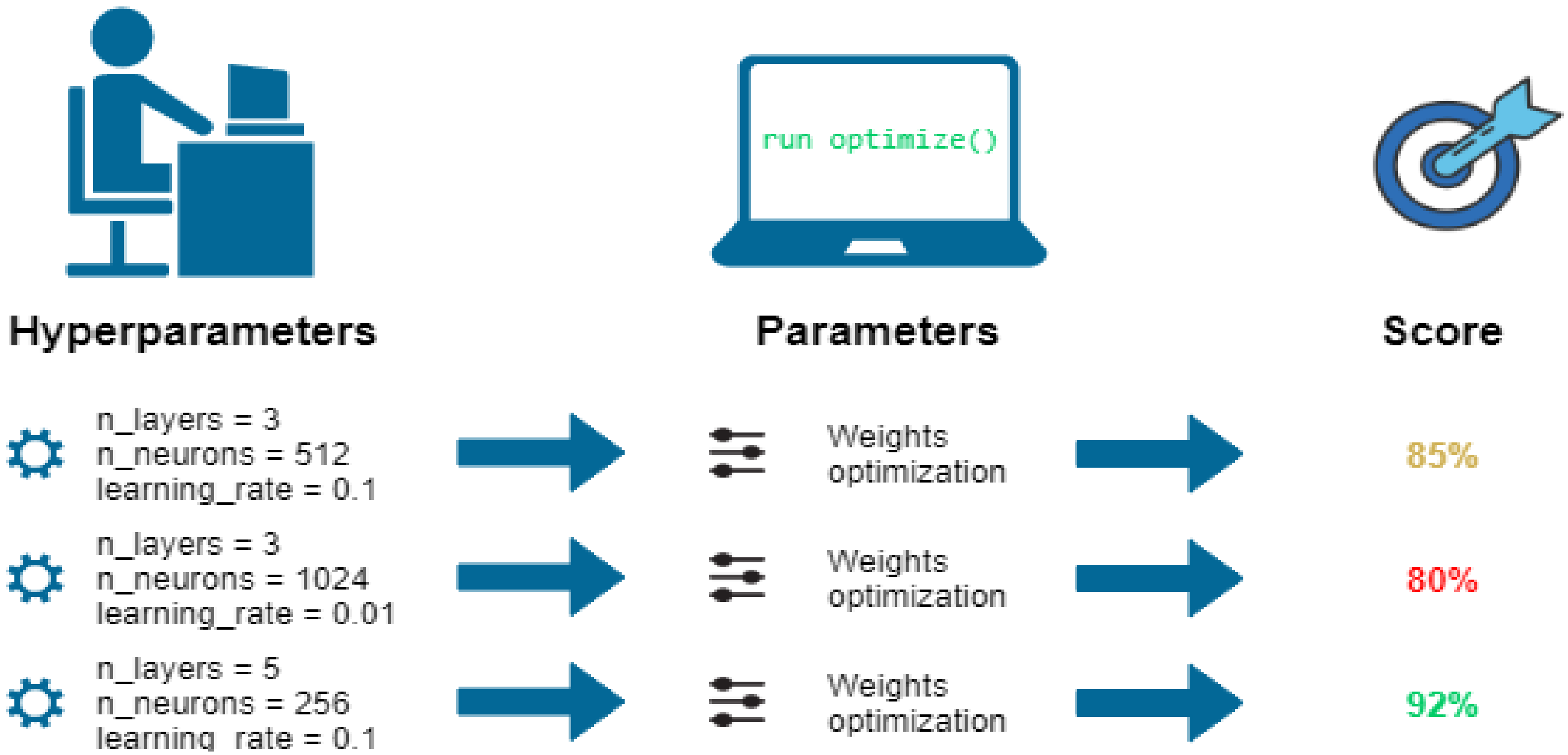


Figure: Hyperparameter selection

Source: https://images.app.goo.gl/o3DaxpbGLPVRxKnPA

Hyperparameter selection (2 of 2)

- Optimization of Bayesian hyperparameter.

- There are two parts:
  - Exploration: Test the feature with the most uncertain outcome in collection of hyperparameters.
  - Operating: Test this function on a set of high-value hyperparameters.

The problem with high dimensionality

- The dimension of a problem refers to the number of input variables (actually, degrees of freedom).

- The exponential increase in data required to densely populate space as the dimension increases.

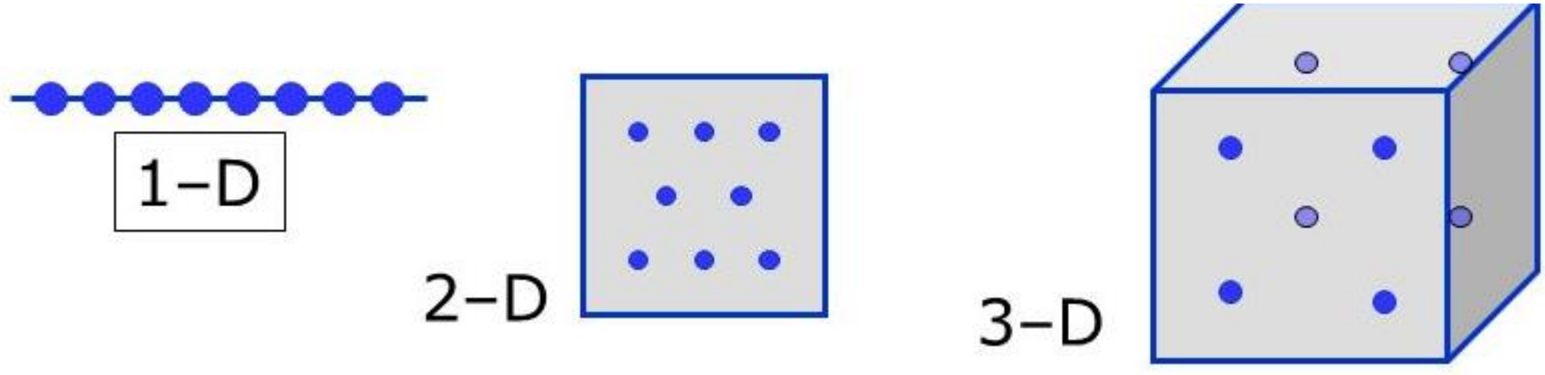


Figure: The Problem with High Dimensionality

Source: https://images.app.goo.gl/T6QLvXses34dSeXXA

Information theory

Information:

$$I(x) = -\log P(x). \tag{3.48}$$

Entropy:

$$H(\mathrm{x}) = \mathbb{E}_{\mathrm{x}\sim P}[I(x)] = -\mathbb{E}_{\mathrm{x}\sim P}[\log P(x)]. \tag{3.49}$$

KL divergence:

$$D_{\mathrm{KL}}(P\|Q) = \mathbb{E}_{\mathrm{x}\sim P}\left[\log \frac{P(x)}{Q(x)}\right] = \mathbb{E}_{\mathrm{x}\sim P}\left[\log P(x) - \log Q(x)\right]. \tag{3.50}$$

Figure: Information theory Formulas

Source: https://images.app.goo.gl/DnbEL1fyXk23ory49

Self evaluation: Exercise 5

- To continue with the training, after learning the various steps involved in pattern recognition and anomaly detection, it is instructed to utilize the concepts to perform the following activity.

- You are instructed to write the following activities using python code.

- Exercise 5: Logistic regression model.

Checkpoint (1 of 2)

Multiple choice questions:

1. The recalled output in pattern association problem depends on?
   a) Nature of input-output
   b) Design of network
   c) Both input & design
   d) None of the mentioned

2. What is the objective of feature maps?
   a) To capture the features in space of input patterns
   b) To capture just the input patterns
   c) Update weights
   d) To capture output patterns

3. Use of nonlinear units in the feedback layer of competitive network leads to concept of?
   a) Feature mapping
   b) Pattern storage
   c) Pattern classification
   d) None of the mentioned

Checkpoint solutions (1 of 2)

Multiple choice questions:

1. The recalled output in pattern association problem depends on?
   a) Nature of input-output
   b) Design of network
   c) Both input & design
   d) None of the mentioned

2. What is the objective of feature maps?
   a) To capture the features in space of input patterns
   b) To capture just the input patterns
   c) Update weights
   d) To capture output patterns

3. Use of nonlinear units in the feedback layer of competitive network leads to concept of?
   a) Feature mapping
   b) Pattern storage
   c) Pattern classification
   d) None of the mentioned

Checkpoint (2 of 2)

Fill in the blanks:

1. ___________learning is involved in pattern clustering task.
2. If the weight matrix stores the given patterns, then the network becomes __________.
3. Activation models are ________.
4. Information theory is using in ___________ detection.

True or False:

1. From given input-output pairs pattern recognition model should capture characteristics of the system? True/False
2. Can system be both interpolative & accretive at same time? True/False
3. Does pattern classification belong to category of non-supervised learning? True/False

Checkpoint solutions (2 of 2)

Fill in the blanks:

1. **Unsupervised** learning is involved in pattern clustering task.
2. If the weight matrix stores the given patterns, then the network becomes **auto associative memory.**
3. Activation models are **deterministic.**
4. Information theory is using in **pattern** detection.

True or False:

1. From given input-output pairs pattern recognition model should capture characteristics of the system? **True**
2. Can system be both interpolative & accretive at same time? **False**
3. Does pattern classification belong to category of non-supervised learning? **False**

Question bank

Two mark questions:

1. What is pattern detection?
2. What is information theory?
3. What is linear regression model?
4. What is the math formula for curve designing?

Four mark questions:

1. What is the difference between pattern and anomaly detection?
2. What is polynomial curve fitting?
3. Describe high dimensionality problems.
4. Describe information theory components.

Eight mark questions:

1. Explain model selection techniques.
2. Explain probability theory in details.

Unit summary

After completing this unit, you should be able to:

- Understand the concept of pattern recognition and anomaly detection
- Gain knowledge on example of polynomial curve fitting
- Learn about probability theory architecture and working model
- Understand Information theory