BAUERBERG KLEIN

TRAINING & CONSULTING

BAUERBERG KLEIN
TRAINING & CONSULTING

Harness Upstream Geophysical and Petrophysical Data with AI Workflows

INDEX

INDEX

Module 04

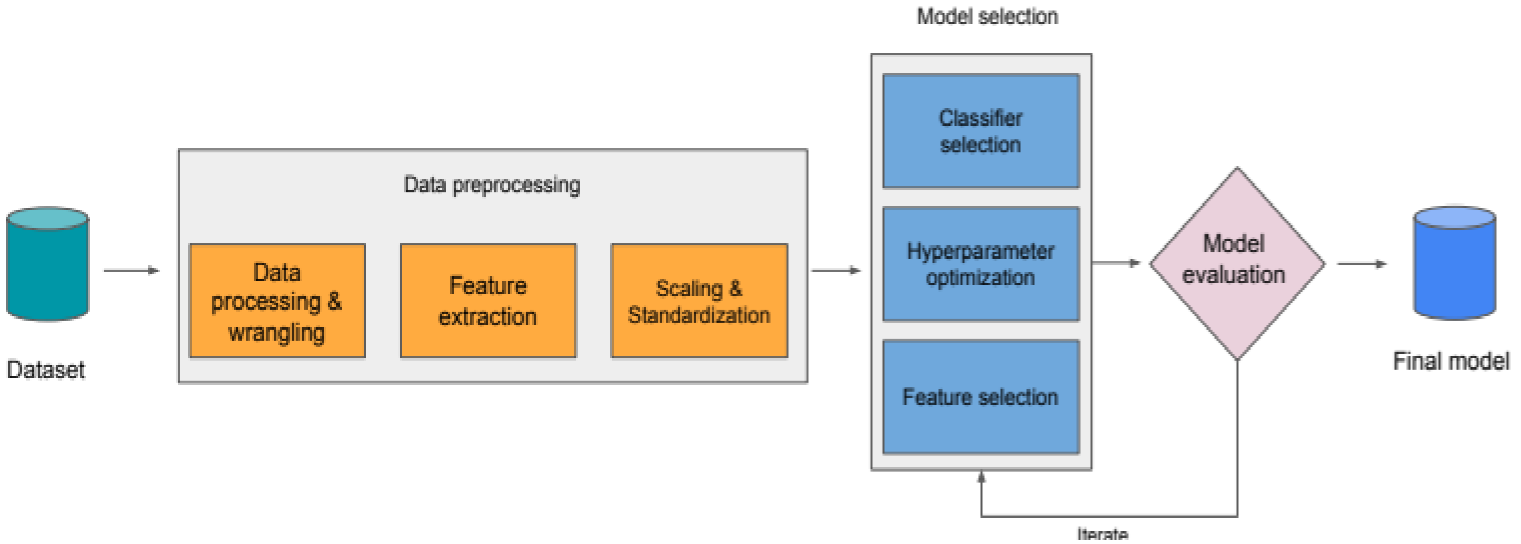
Machine Learning Techniques: Supervised and Unsupervised in E&P

LEARNING OBJECTIVES

- ➢ GOAL01: Machine Learning Fundamentals: Classification Clustering
- ➢ GOAL02: ML and DL Algorithms: Best Practices
- ➢ GOAL03: Modeling Limitations
- ➢ GOAL04: Model Selection Criteria

Machine Learning Techniques

Regression Classification Clustering

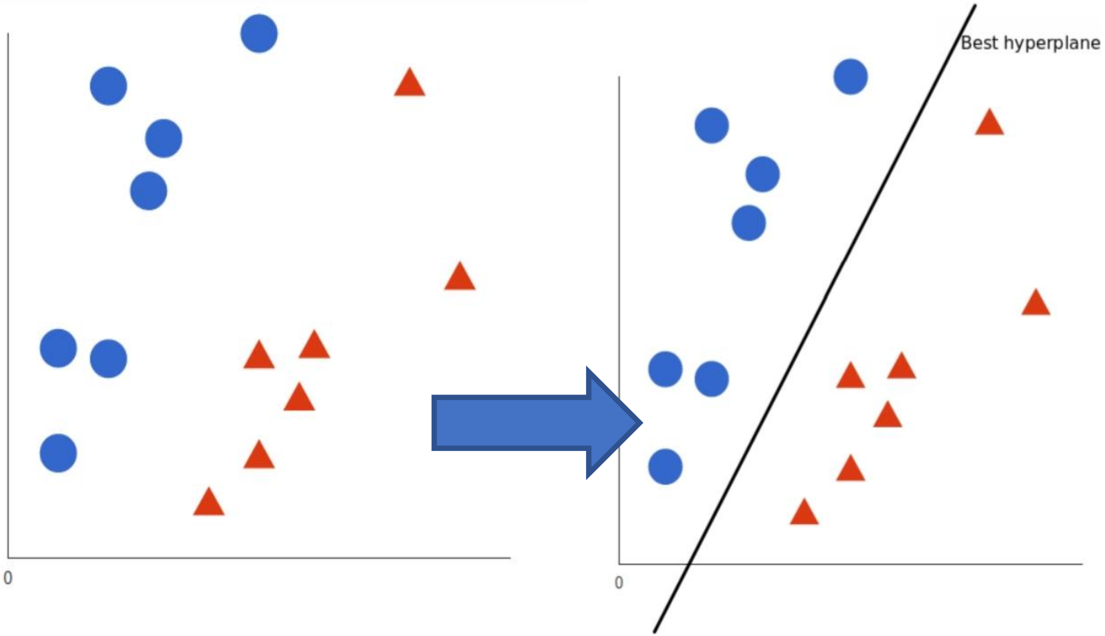


Machine Learning Techniques

Classification

Popular Classification Algorithms:

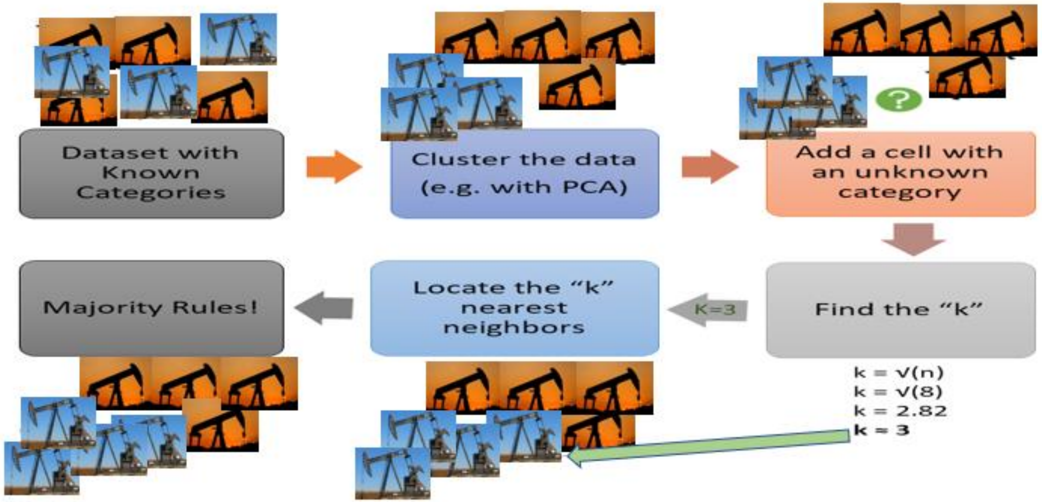
- Logistic Regression
- Naive Bayes
- **K-Nearest Neighbors**
- Decision Tree
- Support Vector Machines

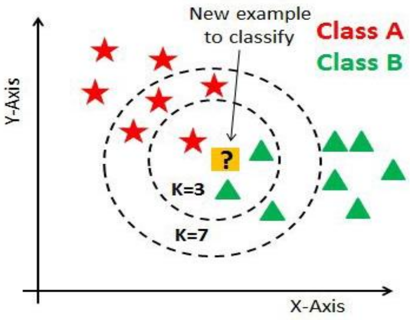


Machine Learning Techniques

Classification Clustering

K-Nearest Neighbor in one picture:

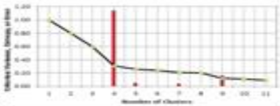


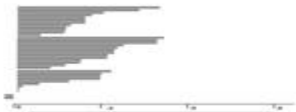


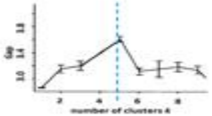
Machine Learning Techniques

Classification Clustering

Determining Optimum Number of Clusters for an Upstream Analysis



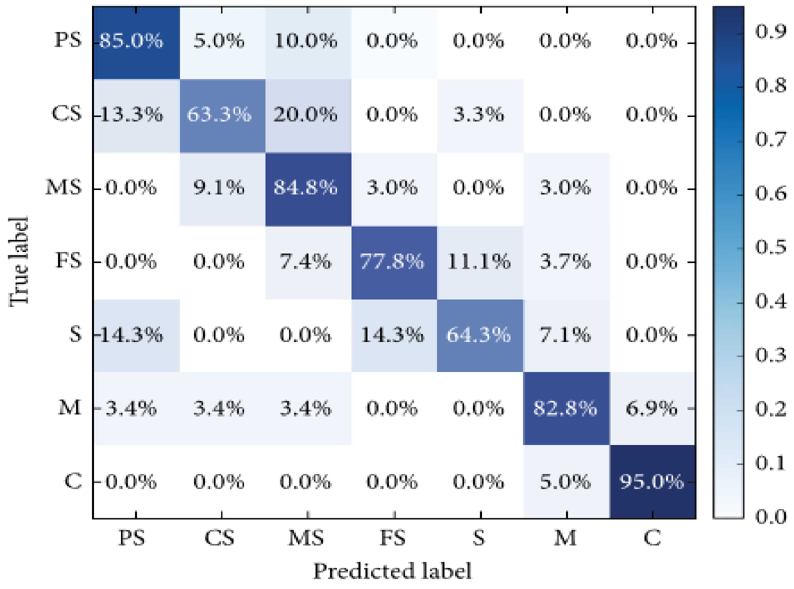




Technique	Elbow Rule	Silhouette Rule	Gap Statistic
How it works	Finds optimal (minimized) within cluster sum of squares (SSE) as a function of number of clusters	Measures how well objects are contained within clusters. Ranges from –1 to +1. High values mean well-matched to own cluster vs. other clusters.	Compares different "k" values with expected values from a dataset with no obvious clustering. The larger the "gap", the better.
Advantages	Ease of use and verification: Clear visual shows an "elbow" (red vertical line in the above image), beyond which more clusters add very little value	Easy to use. Popularity, which means there is plenty of available documentation.	Ease of use: optimal **k** value clear to see (a dashed blue line in the above image) May outperform other methods, according to authors.
Disadvantages	Sometimes ambiguous	More challenging to use and interpret; visual not too intuitive	Functions are poorly documented in R (Steorts, 2017)

Machine Learning Techniques

Classification



The diagonal of the matrix presents the percentage of lithology classes that are correctly classified

Machine Learning Techniques

ML and DL Algorithms used in O&G: Supervised & Unsupervised

Algorithms	Application	Advantages	Disadvantages
ANN Artificial neural networks MLP multi-layer perceptron FF feed forward RBF radial basis function CN convolutional FN functional PN probabilistic	Regression/ classification/ clustering	Learning algorithms are simple With available data it can superior any other model Does not depend on linearity of any function Can be used for problems which are hard or not practical to get a formula for	They are "black box" in nature so it is not easy to be understood or interpreted Lack the ability of generalization as they are exposed to overtraining and might memorize specific data For small datasets, the predictions are not acceptable
		ANN can be used for tasks that linear programs cannot handle Due to the parallel nature of the networks, they can proceed without problems even if an element fails They can learn from experience and avoid reprogramming Applicable in most problems	Neural networks need training to be used The architecture is different from problem to another For big networks, the training and processing time is high
		It is tolerant to faults Can learn from experience Effect of small changes is minor	Needs parallel processing abilities
		Handle nonlinear data Excellent in fitting applications	Exposed to overfitting Can be trapped in local optimum solution Consumes large time in training

Machine Learning Techniques

ML and DL Algorithms used in O&G: Supervised & Unsupervised

Algorithms	Application	Advantages	Disadvantages
FL Fuzzy logic	Classification/clustering	Quick, easy, strong and effect of environment changes is minor Gives a combination of numeric and symbolic picture of systems Can handle problems with strict conditions or even without exact solution Can be described with few data points or approximated datasets	If mathematical model is existing, FL is used only in case of low computational capabilities Not easy to prove the system characteristics as it lacks mathematics
		Simple reasoning, application and can deal with uncertainties and nonlinearity	Lacks robustness
		It is able to detect hyperplane of optimal separation Deals with higher degrees of dimensionality Its kernels can learn precise concepts as they have infinite Vapnik–Chervonenkis dimension Works well usually	Positive and negative examples need to be used to train the model Kernel function choice needs care Consumes memory and computation time Suffers from numerical stability issues while solving the constraint QP

Machine Learning Techniques

ML and DL Algorithms used in O&G: Supervised & Unsupervised

Algorithms	Application	Advantages	Disadvantages
SVM Support vector machine	Regression/ classification/ clustering	Provide high accuracy classifiers.Overfitting occurrence is little, excellent in dealing with noise Preferred for text classification applications that are normally high dimensional problems Intensive memory consumption	It is a binary classification technique, so it needs pairwise classification to perform multi-class classification that means one class against all others, for all classes Runs slowly and require high computational power
		Get useful information from little datasets Has generalization capabilities	Low performance with big data or multi-classification tasks Kernel function parameters affect the performance
		Easy to understand and can be interpreted Data preparation is fast Can deal with numerical and categorized data It has white box interpretable model Statistical tests can be used to validate the model accuracy Robust Efficiently handle huge data in a little time	Even for most simple concepts, the learning of an optimal DT is known as NP-complete Complex DT models cannot generalize the data well Fails to learn some concepts as it is not easy for DT to express them
DT Decision tree	Regression/ classification/ clustering	Nonlinearity among parameters do not affect the performance of DT Interpretable and explainable	Complex Duplication might happen for same sub-tree of other paths
		In case of few predictor variables, it is easy to understand Can be used in building models that contain special data types, such as text	Have huge storage requirements The similarity function selection used to correlate instances is sensitive No clear principles for selecting k, excluding over cross-validation or alike Computational rate is high
		Classes do not need to be linearly divisible Modest and powerful	Tends to disregard the attributes importance Sluggish and expensive

Machine Learning Techniques

ML and DL Algorithms used in O&G: Supervised & Unsupervised

Algorithms	Application	Advantages	Disadvantages
KNN *K* nearest neighbors	Classification/ clustering	Understandable and easy to implement technique Can be trained quickly Robust in case of associated noise It is mainly well suited for multimodal classification	Sensitive to local Data structure Memory restriction Supervised type of learning Sluggish algorithm
		High performance Arise variable measures	Computationally expensive Overfit issues
		Quick in implementation Less complex	Does not depend on variables Disregards original geometry of data
RF Random forest	Classification/ clustering	Robust with noisy data Can learn in increments	Low performance with attribute-related training data
K-means	Classification/ clustering	Data point is allowed to exist in different clusters Normal representation of the behavior of genes	Need to define number of clusters c Membership cutoff value has to be set Initial assignment of centroids affects the clusters
		Changeable model that can adapt different dataset distribution If training data increase, the parameters number does not change	In some cases, the convergence in slow
Fuzzy C-means	Classification/ clustering	Modest and easy to-understand the working algorithm As a topological clustering unsupervised technique, it can deal with dataset nonlinearity Unique in directionality reduction being able to convert high dimensions problem to 1–2 dimensions	Time consuming technique

Machine Learning Techniques

ML and DL Algorithms used in O&G: Supervised & Unsupervised

Algorithms	Application	Advantages	Disadvantages
RNN Recurrent neural network	Regression/ classification/	Can record the information as activations with time Manipulate consecutive information that are random in length	It is affected by the gradient vanishing type Not able to be stacked within extra deep modeling
CNN Convolutional neural network	Regression/ classification/ image processing	Able to detect relevant features only from given dataset Same parameters can be utilized in different problems	Tuning of parameters is difficult Requires large amount of data
		Quick training	Quality might be low
GAN Generative adversarial network	Regression/ classification/	No approximation techniques needed Does not require several entries in the samples	Unstable training Generating discrete data is difficult
DBN Deep belief network	Regression/ classification/	Layer by layer strategy of learning makes it capable of learning the features Deals with non-labelled data and can be safe from the overfitting and underfitting issues	Some pre training algorithms decrease the performance as the input data is clamped Run time is long
		Not affected by the fragmentation of training data thus it reduces over-smooth problem	Lower output quality

Machine Learning Techniques

Limitations of AI Models

Limitation	Reason	Solution
Overfitting	Lack of an appropriate amount of data to be used for training	Using the ratio of input data points to the total number of network weights used by the connections (ρ)
Coincidence	Getting a good match by coincidence for a specific dataset	Using discriminant analysis
Overtraining	When the error keeps decreasing by updating the model structure and the model can be more complex to fit a specific dataset	A training methodology that is named "early stopping" can be used Reinforcement learning with in-stream supervision, for example, the generative adversarial networks
Data availability	Sometimes the gathered data is limited	Single-shot learning in which the AI model is pre-trained on a similar dataset and then is enhanced with experience
Interpretability	The single connections in the models do not affect alone but the whole model connections combined affect results	Local interpretable model and its agnostic explanations The generalized additive models method
Generalization	Model failure in the circumstances different from the set of circumstances, which were used in building the original model	Additional resources are to be utilized for training new datasets
Bias	The nature of black-box models makes it to be prone to biases	Using model-independent perturbations

Machine Learning Techniques

Model Selection Criteria

Vocabulary

When selecting a model, we distinguish 3 different parts of the data that we have as follows:

Training set	Validation set	Testing set
• Model is trained • Usually 80% of the dataset	• Model is assessed • Usually 20% of the dataset • Also called hold-out or development set	• Model gives predictions • Unseen data

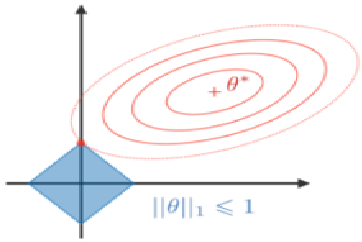
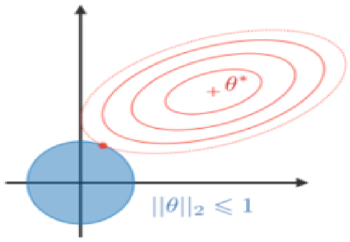
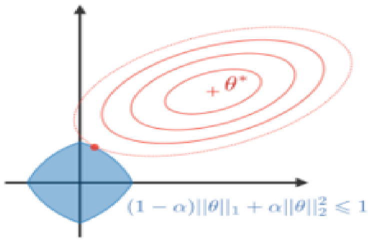
Once the model has been chosen, it is trained on the entire dataset and tested on the unseen test set. These are represented in the figure below:

Machine Learning Techniques

Model Selection Criteria

Regularization

The regularization procedure aims at avoiding the model to overfit the data and thus deals with high variance issues. The following table sums up the different types of commonly used regularization techniques:

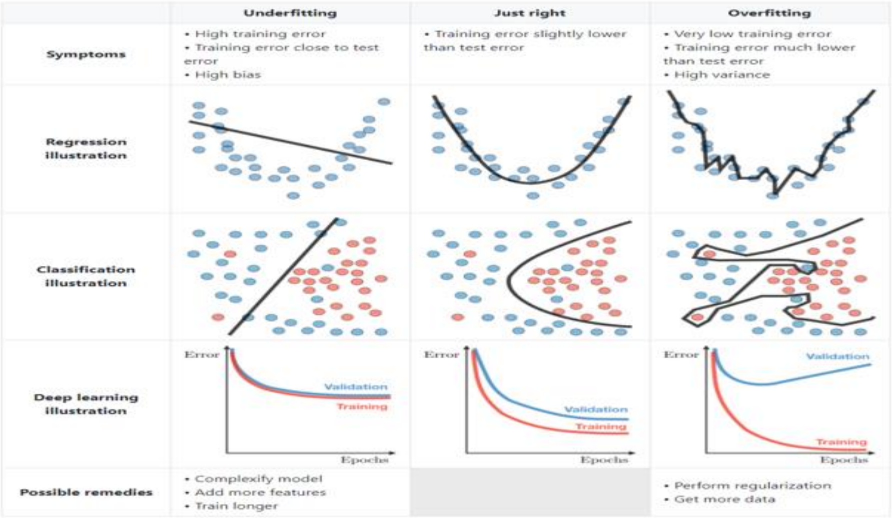
LASSO	Ridge	Elastic Net
• Shrinks coefficients to 0 • Good for variable selection	Makes coefficients smaller	Tradeoff between variable selection and small coefficients
		
$... + \lambda\|\|\theta\|\|_1$ $\lambda \in \mathbb{R}$	$... + \lambda\|\|\theta\|\|_2^2$ $\lambda \in \mathbb{R}$	$... + \lambda\Big[(1-\alpha)\|\|\theta\|\|_1 + \alpha\|\|\theta\|\|_2^2\Big]$ $\lambda \in \mathbb{R}, \alpha \in [0,1]$

Machine Learning Techniques

Model Selection Criteria

Bias/Variance Tradeoff

The simpler the model, the higher the bias, and the more complex the model, the higher the variance.



	Underfitting	Just right	Overfitting
Symptoms	• High training error • Training error close to test error • High bias	• Training error slightly lower than test error	• Very low training error • Training error much lower than test error • High variance
Regression illustration			
Classification illustration			
Deep learning illustration	Error, Validation, Training, Epochs	Error, Validation, Training, Epochs	Error, Validation, Training, Epochs
Possible remedies	• Complexify model • Add more features • Train longer		• Perform regularization • Get more data

Module 05

Deep Learning Techniques: Upstream E&P Deep Learning

MODULE 05

Deep learning techniques require substantial amounts of labeled data and significant computational resources for training. However, they have demonstrated remarkable capabilities in handling complex data and achieving state-of-the-art performance in various tasks. It's essential to carefully design deep learning architectures, preprocess the data, and fine-tune the models to extract the most meaningful insights from exploration and production data.

Deep learning techniques have gained significant attention in recent years for their ability to handle complex and high-dimensional data in various domains, including exploration and production in the oil and gas industry. Deep learning models, particularly neural networks with multiple layers, can automatically learn hierarchical representations from the data, enabling them to capture intricate patterns and relationships.

Here's an overview of deep learning techniques commonly applied to exploration and production data:

- Convolutional Neural Networks (CNNs)
- Recurrent Neural Networks (RNNs)
- Autoencoders
- Generative Adversarial Networks (GANs)

BAUERBERG KLEIN

TRAINING & CONSULTING

BAUERBERG KLEIN
TRAINING & CONSULTING