Machine Learning for Ocean Colour

Retrieval of CDOM - regression

Ana B. Ruescas

ana.ruescas@brockmann-consult.de
bruescas@uv.es

July 28, 2022

Brockmann Consult GmbH
IPL Image Processing Laboratory

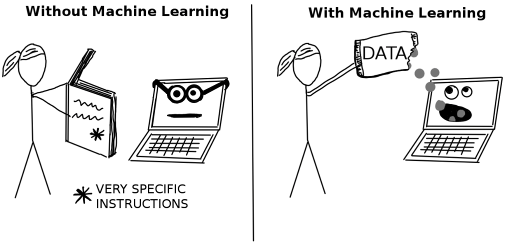
Table of Contents

About myself

- Since 2018: Associate professor in the Geography Department at University of Valencia (Spain): Hydrology, Biogeography, Fundamentals of Cartography and Remote Sensing
- Since 2016: Member of the Image and Signal Processing research group @UV: Machine learning RS applications
- Since 2011: Senior Remote Sensing Expert (External) @Brockmann Consult GmbH: Ocean Colour and SNAP
- Since 2017: Part of the Copernicus Marine Training Service Team @EUMETSAT: preparation and delivery of marine training courses

What is Machine Learning?

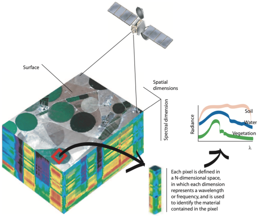
- Machine learning is a set of methods that computers use to make and improve predictions or behaviors based on data.
- The presentation will focus on supervised machine learning: we have a dataset for which we already know the outcome of interest and want to learn to predict the outcome for new data.
- If the output is categorial, the task is called classification, and if it is numerical, it is called regression.
- Clustering tasks (= unsupervised learning) or reinforcement learning are not the object of this lecture.

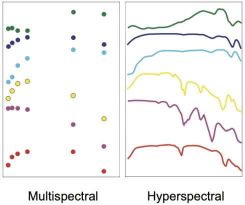


[C. Molnar, Interpretable Machine Learning, 2021]

Satellite image processing

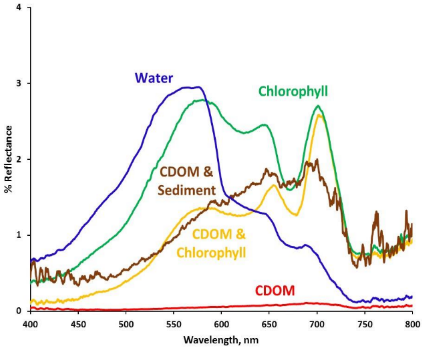
- Materials in a scene reflect, absorb, and emit electromagnetic radiation in a different way depending of their molecular composition and shape
- Remote sensing exploits this physical fact
- Absorption, depth, re-emissions and particular spectral features
- Accurate identification of bio-physical components and processes
- Image spectroscopy allows to identify materials in the scene with unprecedented accuracy



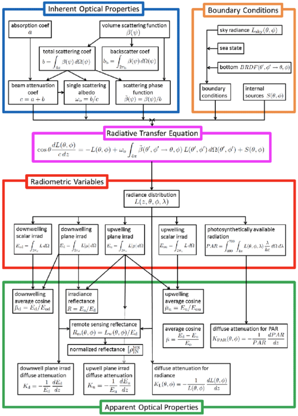


Ocean Colour processing

- Different materials produce different electromagnetic radiation spectra
- The spectrum shows absorptions and emissions at different wavelengths
- The spectral information contained in an image pixel can therefore indicate the various components in the water



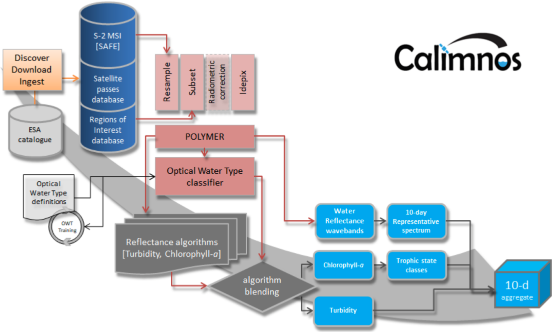
[Olmanson et al., 2016]



[Ocean Optics Web Book]

Standard processing chain

- Many steps and by-products from signal/image acquisition to the final product
- Transmission → Preprocessing → Processing
- A wide diversity of problems and dedicated tools



[Stelzer et al., 2020]

Challenges

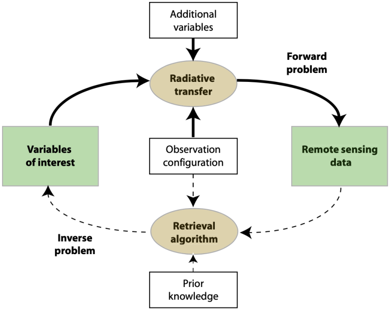
- How to select the best features that describe the problem
- Extract the best combination of spectral bands
- Automatically find groups of pixels in the image for screening, detection...
- Estimate geo-physical variables from the spectra (e.g chlorophyll-a)
- Assign semantic classes to objects/regions in the scene (e.g. water types)
- High dimensional data: multi-temporal, multi-angular and multi-source fusion
- Non-linear and non-Gaussian feature relations
- Few supervised (labeled) information is available (high cost)
- Tons of data to process in (near) real-time

Regression

Retrieval of biophysical parameters

The objective here is to transform measurements into biophysical parameter estimates with EO data.

- Forward modelling: simulate a database of pairs of reflectance spectra + parameters with RTMs
- Inverse modeling: numerical/statistical invert models from RS data to estimate the parameters by designing algorithms that, starting from radiance, can give estimates of the variables of interest



Methods for model inversions

Three main families of methods:

1. *Statistical* inversion models: parametric and non-parametric
    - Parametric models rely on physical knowledge of the problem and build explicit parameterized expressions that relate a few spectral channels with the bio-geo-physical parameter(s) of interest.
    - Non-parametric models are adjusted to predict a variable of interest using a training dataset of input-output data pairs.
2. *Physical* inversion models: try to reverse RTMs
    - After generating input-output (parameter-radiance) datasets, the problem reduces to, given new spectra, searching for similar spectra in the dataset and assigning the most plausible ('closest') parameter.
3. *Hybrid* models try to combine the previous approaches.

Statistical approaches

Two main approaches:

1. **Parametric regression** inversion models: assume an explicit model for retrieval, e.g. discrete band approaches like indices, band ratios...
2. **Non-parametric regression**: do not assume explicit feature relations

Linear non-parametric models
Stepwise multiple linear regression (SMLR)
Partial least squares regression (PLSR)
Ridge regression (RR)
Least Absolute Shrinkage and Selection Operator (LASSO)

Non-linear non-parametric models
Decision trees, bagging and random forests
Neural networks
Kernel methods: SVR, RVM, KRR, GPR
Bayesian networks

Parametric approaches

Weaknesses

- Makes only poorly use of the available information within the spectral observation; at most a spectral subset is used. Therefore, they tend to be more noise-sensitive as compared to full-spectrum methods
- Parametric regression puts boundary conditions at level of chosen bands, formulations and regression function.
- Statistical function accounts for one variable at the time.
- A limited portability to different measurement conditions or sensor characteristics.
- No uncertainty estimates are provided. Hence the quality of the output maps remain unknown.

Strengths

- Simple and comprehensive regression models; little knowledge of user required.
- Computationally inexpensive.
- Fast in processing.

Non-parametric approaches

Weaknesses

- Training can be computational expensive.
- Can create over-complex models that do not generalize well (overfitting).
- Expert knowledge required, e.g. for tuning. However, toolboxes exist that automate some steps.
- Some regressors behave rather unstable when applied to data that deviate from statistically different from those used for training.
- Most of them act as a black box.

Strengths

- Can make use of all bands, full spectrum.
- Build advanced, adaptive (nonlinear) models.
- Some methods cope well with redundancy and noisy data.
- Once trained, fast processing images.
- Some of them (e.g. NN, decision trees) can be trained with high numbers of samples.
- Some methods provide insight in model development (e.g. GPR: relevant bands; decision trees: model structure).
- Some methods provide uncertainty intervals (e.g. GPR, KRR).

How to measure the goodness of a model?

Given two variables y_i and $\hat{y}_i$, i = 1, ..., N

- Error (residuals): $e_i = y_i - \hat{y}_i$
- Bias: mean error (ME): $ME = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)$
- Accuracy: $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$
- Goodness-of-fit: Pearson‘s correlation coefficient

In Python:

```
from sklearn.metrics import r2_score, mean_absolute_error, mean_squared_error
print("r2=",r2_score(ytest, ypred))
print("MAE =", mean_absolute_error(ytest, ypred))
print("RMSE =", mean_squared_error(ytest, ypred))
```

Advanced statistical retrieval

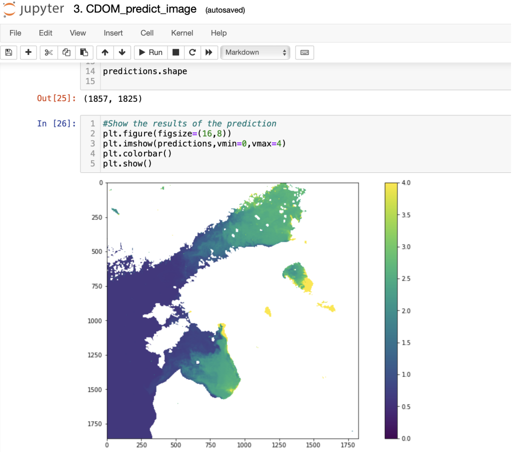
1. What is Gaussian Process (GP):
   - Gaussian Processes are a generalization of the Gaussian probability distribution.
   - It is a probability density over functions, non-linear and non-parametric.
   - GPR is still a form of supervised learning.
   - A Gaussian process is a Gaussian random function, and is fully specified by a mean function m(x) and covariance function k(x, x). This covariance function is called the latent function or the "nuisance" function.
   - They are a type of kernel model and they are capable of predicting highly calibrated class membership probabilities
   - The hyperparameters of the GPR algorithm on a given dataset can be tuned.
2. Deep Learning (it is all about scale)
   - Deep Learning is a subfield of ML concerned with algorithms inspired by the structure and function of the brain called artificial neural networks.
   - Multi-output regression involves predicting two or more numerical variables.
   - Deep learning neural networks are an example of an algorithm that natively supports multi-output regression problems.
   - Neural network models for multi-output regression tasks can be easily defined and evaluated using the Keras deep learning library.
   - What is deep learning?

Summary on regression

- Biophysical parameter estimation is perhaps the most important (and challenging) problem in remote sensing
- Traditional methods were focused on simplistic approaches using only few spectral bands
- New regression-based approaches alleviate the problems by exploiting the wealth of spectral and auxiliary information
- The common approaches consider:
  - Empirical models (e.g. indices) are easy, fast but too general
  - Physical radiative transfer models are flexible but slow and require specific information (e.g. aerosols, geometry) which is not always available
  - Non-parametric regression may offer a robust alternative that can be easily implemented in operative processing chains

Python exercise

Machine Learning Regression for Ocean Parameters Retrieval



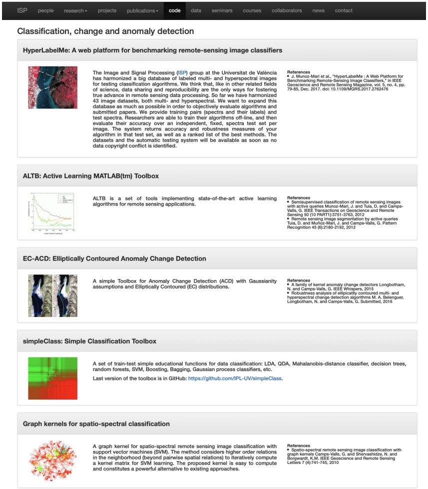
https://gitlab.com/benloveday/mlregocean_cdom

Thoughts

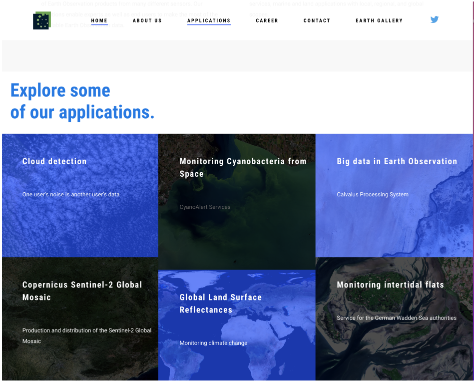
- A major disadvantage of using machine learning is that insights about the data and the task the machine solves is hidden in increasingly complex models.
- The higher the interpretability of a machine learning model, the easier it is for someone to comprehend why certain decisions or predictions have been made. Miller (2017), "Interpretability is the degree to which a human can understand the cause of a decision". Another one is: "Interpretability is the degree to which a human can consistently predict the model's result".

Links and references

Image and Signal Processing Group, Universitat de València



Brockmann Consult GmbH



[Special thanks to G. Camps-Valls and D. Tuia]