

Geocomputation & Machine Learning course

Quick recap

This meeting was a transition lecture from geocomputation to machine learning, focusing on preparing a table for tree height estimation using the JEDI dataset. Giuseppe demonstrated how to extract point data from satellite LIDAR scans and prepare it for random forest modeling, explaining the dataset structure, quality control measures, and predictor variables including climate, soil, and topographic data. The team discussed technical issues with Ubuntu virtual machines and RAM allocation, with recommendations to increase resources or consider full Linux installations. Raffaele provided guidance on proper data splitting techniques to avoid bias in training and testing sets, explaining the importance of maintaining similar distributions between both datasets. The discussion covered random forest parameter tuning using grid search and cross-validation methods, with Raffaele introducing more advanced hyperparameter optimization techniques like Optuna and Hyperopt. The session concluded with practical demonstrations of prediction back to rasters and visual validation of model results, setting up the foundation for future machine learning applications in the course.

Next steps

Abdul

- Shut down RAM-heavy Windows applications, including antivirus, when running the Ubuntu virtual machine to free up resources.

Giuseppe

- Send an email to students regarding the time change for the next week's lecture due to US daylight savings.

Raffaele

- Take over the machine learning part of the course starting from the next lecture.

Shiva

- Increase the RAM in the virtual machine settings when the Ubuntu VM is shut down to prevent freezing/crashing.

Collaboration

- Students: Consider performing a full Ubuntu installation on a second computer or replacing the current OS if virtual machine performance remains an issue.
- Students: Run the provided Python script to extract the GEDI dataset from HDF files and create the table for machine learning.
- Students: Use the provided study case on tree height estimation to practice building a random forest model, including data cleaning, splitting, and evaluation.
- Students: Experiment with hyperparameter tuning using grid search or other libraries like Hyperopt/Optuna to improve model performance.
- Students: Apply quality control filters to the dataset (e.g., removing outliers, using quality flags) to improve model accuracy.
- Students: Practice predicting the model back to a raster format and perform visual validation of the results.
- Students: Send traveling and accommodation details to Giuseppe if attending the Matera event.

Summary

Ubuntu VM Performance Issues Discussion

The team discussed technical issues with Ubuntu virtual machines, with Shiva reporting frequent crashes and Abdul experiencing performance problems when running both Hekra simulation and Ubuntu simultaneously. Giuseppe advised increasing RAM and processor allocation in the virtual machine settings and suggested considering a full Linux installation for better performance. The meeting focused on transitioning from geocomputation to machine

learning, with plans to use the JEDI dataset for tree height estimation as a case study. Giuseppe explained the HDF format and demonstrated how to process the dataset to create a table for machine learning analysis, specifically using random forest algorithms.

LiDAR Data Preprocessing for ML

Giuseppe explained the process of cleaning and preparing LiDAR data for machine learning, focusing on filtering points based on forest area and quality metrics. He described how the response variable would be tree height derived from six algorithms, with quality flags available for model weighting. Giuseppe outlined the predictors used, including satellite data, climate indices, and other environmental variables, emphasizing that random forest models should include all available predictors rather than worrying about collinearity.

Tree Height Prediction Dataset Overview

Giuseppe presented a detailed overview of the dataset being used for tree height prediction, explaining various environmental variables including climate data, soil indices, geomorphometry, and tree cover information. He demonstrated how to analyze the data using histograms and statistical descriptions to identify outliers and assess data quality. The discussion covered methods for handling autocorrelation between variables and included guidance on data extraction techniques using GDAL and Rasterio tools, with Python being identified as the preferred approach for building the final dataset table.

Randall Forest Project Break Meeting

The meeting participants agreed to take a break and reconvene at 30 minutes later to work on the Randall Forest project. Giuseppe asked if anyone had experience using Random Forest, but no specific responses were captured in the transcript.

Random Forest Algorithm Discussion

Giuseppe led a discussion on random forest algorithms, explaining their principles, implementation, and the importance of proper data splitting for training and testing. He emphasized that random forest models can handle non-parametric data, including variables with different scales and distributions, but caution was needed to avoid overfitting by controlling the tree's complexity. Raffaele highlighted the importance of stratifying data splits

based on the target variable to ensure both training and testing sets reflect the true data distribution. Abdul asked about handling time series data, to which Raffaele responded that the approach depends on the analysis goals, suggesting time-based or geographic-based splits when appropriate.

Model Evaluation and Performance Metrics

Raffaele clarified that training performance metrics cannot be relied upon for model evaluation, as the model optimizes itself based on the target variable. Giuseppe explained that testing and training performance should be similar, with a difference of no more than one digit indicating acceptable performance, and warned about overfitting issues in Python compared to R. The discussion covered proper model parameterization, including setting random seeds for reproducibility and adjusting parameters like minimum sample split and maximum depth, with emphasis on monitoring out-of-bag scores to ensure model generalization.

Model Bias and Data Distribution

Raffaele explained that the model is experiencing a bias, predicting values around 30 due to the distribution of training data having two main peaks near 0 and 30. Giuseppe discussed the importance of feature selection and highlighted that tree cover and vegetation index are the most influential predictors. They agreed that improving the data distribution and selection, rather than just hyperparameter tuning, would be necessary to address the model's bias.

Data Quality and Model Performance

Raffaele explained strategies for addressing skewed data distribution, including bootstrapping and transforming target variables, while emphasizing the importance of maintaining reversible transformations. He also detailed the differences between holdout validation and k-fold cross-validation, noting that the choice depends on community standards and computational constraints. Giuseppe demonstrated the impact of data quality control on model performance, showing improved results after applying quality controls and selecting better data, with the final model achieving an R-squared value of 0.70-0.61 compared to the initial 0.50-0.60 without controls.

Dataset Downsampling and Hyperparameter Optimization

Raffaele explained his approach to downsampling the dataset to achieve a smoother distribution for the target variable, aiming to prevent model bias towards certain height predictions. He also discussed the grid search strategy for optimizing hyperparameters and introduced alternative methods like Optuna and Hyperopt for more efficient searches. The group reviewed the process of making predictions back to the raster, emphasizing the importance of visual validation and expert field knowledge. Giuseppe announced upcoming schedule changes and encouraged participants to plan their travel to Madeira, while Raffaele clarified questions about decision trees and feature importance in random forests.