

Geocomputation & Machine Learning course

Quick recap

The meeting focused on an introduction to machine learning and the use of support vector machines (SVM) for classification and regression tasks. Giuseppe opened with a brief recap of the previous session on random forests and addressed a question from Ali about the HyperAuth library, explaining that while R has better default settings for overfitting, Python is generally faster and more flexible for advanced machine learning and neural networks. Raffaele then led the main lecture, outlining the course structure and defining machine learning as the process of identifying patterns in data. He explained the differences between supervised learning, where models learn from labeled data to classify or predict, and unsupervised learning, where models organize unlabeled data, using PCA as an example. Raffaele detailed the mechanics of SVM for classification, including the concept of hyperplanes, margins, support vectors, and the use of a loss function and gradient descent to optimize weights. He then transitioned to support vector regression (SVR), explaining the epsilon-tube concept and the role of the C hyperparameter in penalizing points outside the margin. Through a hands-on tutorial using a tree height dataset, Raffaele demonstrated data preprocessing, model training, and hyperparameter optimization using grid search to improve R-squared performance. The session concluded with a discussion on the relationship between epsilon and C and a reminder from Giuseppe that these concepts form the foundation for future lessons on deep learning and neural networks.

Next steps

Annabell

- Find out using the grid search tool of scikit-learn how to optimize epsilon and C hyperparameters at the same time and determine the best combination for the model's testing performance.

Raffaele

- Share the link to the presentation and notebook in the chat for the students to access.

Collaboration

- All students: Perform the rsync of the SE data as usually done.
- All students: Review the workflow and video recording, and run the scripting procedure to experiment with changing parameters and observing how predictions change.
- All students: Be ready to ask clarification questions to Raffaele next week, as the upcoming deep learning and neural network content will build on the support vector machine concepts.
- All students: Set alarms/update calendars for the time shift next week; the lecture will start one hour later due to the time change in Italy.

Summary

Machine Learning Course Planning Discussion

Giuseppe led a discussion on the transition to machine learning using random forest, addressing questions about the HyperAuth library's compatibility with R versus Python, with R being better calibrated for overfitting prevention. Raffaele presented the upcoming machine learning course structure, outlining topics including supervised and unsupervised learning, support vector machines, perceptrons, and convolutional neural networks, with plans for interactive sessions and quizzes to engage participants. The course will conclude with a "bring your own paper" session where participants can present and discuss machine learning applications in geocomputation.

Machine Learning Concepts Overview

Raffaele explained the concepts of machine learning, supervised learning, and unsupervised learning using examples of image classification. He described supervised learning as identifying patterns in labeled data to make predictions, using strawberries and bananas as examples, and explained that unsupervised learning organizes unlabeled data into new representations without predefined classes. Raffaele also introduced mathematical notation to represent these concepts, with X as input data, Y as output variables, and F as the model function that learns to minimize prediction errors.

Machine Learning Fundamentals Overview

Raffaele explained the differences between supervised and unsupervised learning tasks, using examples like classification and clustering. He clarified that deep learning is a method to solve machine learning tasks rather than a type of machine learning itself. The session concluded with a quiz on machine learning fundamentals, where students demonstrated understanding of concepts like supervised and unsupervised learning, as well as the differences between classification and regression.

Deep Learning and SVM Concepts

Raffaele explained the concept of deep learning as a technique for approaching various types of machine learning tasks, including supervised, unsupervised, and reinforcement learning. The discussion then moved to a two-dimensional vector space classification problem, where participants were asked to identify the correct decision boundary. The correct answer was confirmed to be "all of the above" (options A, B, C, and D), as all hyperplanes could separate the classes. Raffaele then explained the concept of support vectors and the importance of maximizing the margin between different classes in a Support Vector Machine (SVM) model.

Support Vector Machines Overview

Raffaele explained the concept of Support Vector Machines (SVM) in detail, focusing on maximizing the margin between classes and minimizing the norm of the weights. He described how the loss function measures classification errors and penalizes incorrect predictions, with a correct classification resulting in zero loss. Raffaele also explained the process of computing gradients to optimize the weights and find the minimum of the loss function.

Gradient Descent Strategy Explanation

Raffaele explained the gradient descent strategy for minimizing loss functions, describing how it works by moving toward negative gradients to find local minima. He clarified that the learning rate α controls the step size, with too large a value potentially overshooting the minimum and too small a value causing slow convergence. When asked by Annabell about updating weights, Raffaele explained that this can be done using different approaches including stochastic gradient descent with full datasets, mini-batches, or individual data points, depending on computational requirements.

Support Vector Regressor Implementation Discussion

Raffaele and Abdul discussed adapting a support vector machine into a support vector regressor for classifying circles and squares. They explored the concept of finding an optimal regression line and explained how the SVR model works with a tolerance range (epsilon) for loss optimization. Raffaele clarified that if the prediction is within epsilon of the actual value, the loss is zero, otherwise, the loss is the absolute difference.

SVM Optimization and Classification

Raffaele explained the optimization of support vector machines (SVM) for classification, focusing on how to penalize misclassifications through the C hyperparameter and the concept of margins and slack variables. He demonstrated the impact of different C values on model performance using a house prices dataset, showing how higher C values can improve both the percentage of data points within the margin and the test mean absolute error. Raffaele clarified that SVM can be used for both classification and regression tasks, but there is no such thing as unsupervised classification.

SVM Tutorial for Tree Height

Raffaele led a tutorial on applying Support Vector Machines (SVM) to tree height estimation using the JD dataset, covering data preprocessing, model training, and hyperparameter optimization. The discussion included explaining how to handle the ID column (determined to be irrelevant for prediction) and the importance of proper data scaling. Annabell asked about the relationship between epsilon and C parameters in hyperparameter optimization, and Raffaele clarified that both should be optimized simultaneously rather than sequentially. The session concluded with a reminder from Giuseppe about the upcoming time change for next week's lesson and encouragement for participants to review the material before the next meeting.