Questions & Answers Part 2

Please type your questions in the Question Box. We will try our best to get to all your questions. If we don't, feel free to email Jordan A. Caraballo-Vega (jordan.a.caraballo-vega@nasa.gov).

Question 0: Where should we submit assignment 1? I couldn't find information about this.
Answer 0: Homework from the assignments will be submitted after the training series via a Google form. You can find all training materials, recordings, and the homework submission form (once available) on the training webpage. https://appliedsciences.nasa.gov/join-mission/training/english/arset-fundamentals-machine-learning-earth-science

Question 1: Are the MODIS 250m data corrected for bow effects (i.e., enlargement and stretching of the pixels because Earth curvature at the edges of the swath)?
Answer 1: In this work we use the L2G product which is gridded data from the swath data. The gridded data are equal sized pixels but there is no "correction" for the bow-tie effect. You can find more information on this here: https://ladsweb.modaps.eosdis.nasa.gov/missions-and-measurements/products/MOD09GQ
In our production algorithm we exclude pixels whose underlying "observation" have viewing geometries with large angles as these pixels would be considered less reliable due to the spread you mentioned in your question.

Question 2: In the homework for Part 1, we did the K-Means with CPU and GPU and I'm getting very different results (the number of grid-points assigned to each cluster is very different). Is it supposed to be like that?
Answer 2: Oftentimes, there are differences between CPU and GPU algorithms across scikit-learn and cuml. These differences should not be significant, but can be present (simply due to calculational error inherent in different processors). There are many noted differences in the literature between sci-kit learn and cuml results. Another factor that could make the K-Means output different is the initial seed that controls

randomization. The initial cluster used by the K-Means could provide different results based on the starting point from where centroids are computed.

Question 3: What are some of the key methods used to remove outliers from the data once detected?
Answer 3: There are generally three methods for handling outliers in a tabular dataset. An example of removing outliers using pandas can be found here: https://stackoverflow.com/questions/23199796/detect-and-exclude-outliers-in-a-pandas-dataframe

More info here for pandas: https://stephenallwright.com/remove-outliers-pandas/

Question 4: Can bias correction help to improve the quality of datasets? Also why don't we use -999 to represent missing values rather than using mean values?
Answer 4: Bias correction can certainly help. Methods such as oversampling and under-sampling can be used to address class imbalance in the dataset. However, it's important to note that bias correction techniques can also have limitations and potential drawbacks. For example, oversampling or undersampling can lead to overfitting if not applied carefully.

Re: using -999 to represent missing values. That might not be the best idea due to that causing outliers in the data if the data value is -999. For remote sensing datasets, it would be best to simply remove rows that contain no-data or null data. Imputing a value such as mean could contaminate the dataset as it is synthesized. In other applications it may be applicable to impute a statistical value such as mean, median, or mode. We use these statistical values to keep the row in the dataset and also aim to not disturb the data distribution. We are assuming that the missing values are similar to the non-missing values and that the mean is a reasonable estimate of the missing values. Imputation can introduce bias if the imputed values are not representative of the true missing values.

I also want to add the following general principles when dealing with missing values. Any choice is problem dependent:

(1) We can remove the rows and columns that have missing values.

(2) We can replace the missing values by another “ acceptable value" (neighboring value, mean, etc.).
(3) We can use our model to determine what could replace the missing value.

Question 5: Is there any guideline or accepted practice for how high of a correlation coefficient should prompt you to discard one of the variables (there is often high correlation between spectral bands but they still contain distinct information)?

Answer 5: After computing the Pearson or Spearman correlation matrix (or both), the general practice is to drop bands with values higher than .95 correlation. In datasets of large number of bands like hyperspectral datasets, additional analysis using principal components might be needed to better understand variance between the correlated bands to drop some of these features. Algorithms that find non-linear relationships might not require dropping highly correlated variables, since they will be able to find relationships across the set of features. Some references here: Midi, H., Sarkar, S. K., & Rana, S. (2010). Collinearity diagnostics of binary logistic regression model. *Journal of interdisciplinary mathematics*, *13*(3), 253-267, Talukdar, S., Singha, P., Mahato, S., Pal, S., Liou, Y. A., & Rahman, A. (2020). Land-use land-cover classification by machine learning classifiers for satellite observations—A review. *Remote Sensing*, *12*(7), 1135.

Question 6: What is the difference between Random Tree Classifier (RTC) and Random Forest Classifier (RFC)? Which machine learning classifier is also good for coarse, moderate, and high resolution satellite images?

Answer 6: Generally speaking the best classifier for fine spatial resolution data is going to be different than the best classifier for moderate or coarse resolution. Example when we work with Commercial High resolution data (< 3m resolution) we often find that CNN is a much better classifier because a feature on the ground (example: a tree crown) will be made up of many pixels at fine spatial resolution. The CNN uses spatial texture to help determine the output class, this outperforms most pixel based algorithms.

Question 7: Can we differentiate minerals using deep learning models on hyperspectral data like Hyperion data?

Answer 7: This is an open area of research but there has definitely been some success in looking for certain types of minerals with hyperspectral data. The coarse spatial resolution of spaceborne instruments like Hyperion can limit the types of minerals you can find in this way.

Question 8: Can you discuss the bin width sensitivity in the histogram? What is the importance of the box plot over the histogram?
Answer 8: Choice of bin size can certainly affect the appearance and interpretation of a histogram. If a bind width is too large the histogram may be too coarse and fail to capture some important details of the distribution. Too small, and the resulting histogram could be too detailed and emphasize some noise in the data over the underlying distribution we are trying to capture.

A histogram can provide information on central tendency, spread, and skewness of the distribution. A box plot shows the median and the interquartile range and any outliers or extreme values.

Question 9: When you take a sample of the data, to conserve computing costs, is this process randomly selecting rows?
Answer 9: In this case we are just randomly selecting a subset for the simplicity of the demo. In a science scenario, you could use the entire dataset if you have enough computational bandwidth, or you could stratify a subset of your data that is representative of the training dataset to do some exploratory data analysis. Note that this will only be useful to inspect the data, and not for data cleaning since you might be missing some of the features that could still need to be cleaned.

Question 10: If we want correlation with the bands and multiple classes (water, forest, settlement, etc.), does it work with this heat map correlation chart?
Answer 10: That can certainly work. You can add a column for each class and use the same method and that will give us the correlation between those class columns and the other features.

Question 11: Can you tell me a bit about rank of data and plotting of spearson rank correlation in machine learning?
Answer 11: Correlation can be easily drawn using scatterplots, but you can use Spearson rank correlation to get a statistical value out of it. The ranking of the two variables is computed by providing a ranking of 1 to the biggest number in the column, 2 to the second value in the column, etc. Tied rankings are then just the average of these. Then you simply take the square differences to remove the negative values and

just sum them. We use the confusion matrix as an easy visualization to show the relationship between those two features.

Question 12: What is the maximum size of an image that is practical to work with pandas?
Answer 12: It all depends on computational and memory limits. Remember pandas only takes tabular format, we need to turn each pixel into a row in the dataset. A image of 4800x4800 that is turned into a per-pixel row dataset can fit comfortably in a pandas dataframe, computations and visuals might be slower to work with. One option for very large datasets (and large images) is to use a Dask dataframe (https://docs.dask.org/en/stable/dataframe.html). This utilizes parallel computing and other techniques to perform computations on really large dataframes.

Question 13: Will we be able to plot the results of PCA?
Answer 13: We are not plotting the PCA output in the exercise, but you can simply take the output array from the PCA output, reshape it, and plot it using any visualization tool like matplotlib.

Question 14: As far as I understand, Tukey IQR works well for normal distribution, but most of the band values in the dataset are skewed with a long tail. Are there any methods more fitted for data like this?
Answer 14: There are many other techniques, we use Tukey IQR as the base for data science and because of its applicability across many other use cases. Regarding multispectral remote sensing data, the idea is to use methods more focused in local and global neighborhoods, not necessarily given by simple distance calculations (e.g. isolation forests, calculating local outlier factors, etc.). Here are a couple of techniques you could look at further: Alvera-Azcárate, A., Sirjacobs, D., Barth, A., & Beckers, J. M. (2012). Outlier detection in satellite data using spatial coherence. *Remote Sensing of Environment*, *119*, 84-91., Liu, H., Jezek, K. C., & O'Kelly, M. E. (2001). Detecting outliers in irregularly distributed spatial data sets by locally adaptive and robust statistical analysis and GIS. *International Journal of Geographical Information Science*, *15*(8), 721-741., John, J. (2021). Outlier Detection and Spatial Analysis Algorithms. *arXiv preprint arXiv:2106.10669*. Sckit-learn also provides other outlier detection algorithms you could try.

Question 15: While cleaning the data, if there are many null values, what kind of techniques can be applied for filling the data gaps?
Answer 15: It would be better to exclude those data from the training than to fill them if you have enough training data. Another option when in the presence of small datasets,

you could exclude some column features if you know some of them are problematic, or you could perform interpolation techniques.

Question 16: How do you decide on the values for the hyperparameters? Do you do hyperparameter optimizations?
Answer 16: You can perform hyperparameter tuning or optimization to determine the best parameters that work well with the model and data. We will cover this in session 3!

Question 17: Do we use a training data set to train the algorithm? Even though we have 80% of data as test data, why do we classify it to 4 training data sets and only 1 test data set?
Answer 17: We do train the algorithm with a full training dataset. We did show K-fold where we do random splits of the training data into k equally-size “folds”, training the model on k-1 of the folds and then testing on the last fold. This allows us to train on the entire dataset.

The following link explains the k-fold cross validation method:

https://www.kdnuggets.com/2022/07/kfold-cross-validation.html

Question 18: How do we train the model without the masking data? As in unsupervised clustering first then using them for masking based on those clusterings.
Answer 18: The model is trained to understand the spectral values and make predictions based on this. It doesn’t matter that the training data are presented to the model in tabular form vs. an image. From a code perspective both are just arrays so it doesn’t make a difference.

Question 19: After getting a set of results, how can we determine whether we need to adjust our hyper parameters vs. selecting a different ML model?
Answer 19: This is where you would perform validation on the output map (aka the model prediction) to determine the quality of the output. If the map validates to an acceptable level (determined by the project PI) then you are all set. If not, you need to tweak the training or try a different method to improve the outputs.

Question 20: My Colab session is crashing every time I try to do the Colab assignments, even though I have selected GPU for my runtime. What should I do?
Answer 20: Try to restart your entire runtime and use a new session. Follow the prompts from the session to know when to restart the session after some of the packages are installed. There are specific cells that will restart the session to reload the GPU components.

.

Question 21: Could you talk more about why 50% is used as the prediction accuracy threshold?
Answer 21: For this specific Random Forest binary classifier example, the closer the probability is in the middle of the range 0-1, the least certain the model is. The closer the probability is to 0 the more certain the model is for the classification to be land. When dealing with multiclass problems, the higher the probability per class, the more certain the model is to predict that class.

Question 22: Are all types of indices datasets available on Cloud?
Answer 22: There are many different types of data available on the Cloud. You would simply have to search to determine if the specific data that you are interested in is available there.

Question 23: What options do you have when the train dataset is giving you results (probably prediction) you do not expect, especially when testing it?
Answer 23: The first order of business would be to evaluate your model training and testing statistics. These can give you a sense of how the model may perform on the data you want to apply to. Next, you can add more training data to help the model learn the features where it was underperforming. Lastly, you could try a different method entirely probably without even updating the training data.

Question 24: What techniques can be adopted to improve machine learning computation time, especially on large, high-resolution data sets?
Answer 24: Easiest method is to add more compute and run multiple scenes simultaneously on different systems. Other possibilities include breaking up an image into smaller pieces and parallelizing the runs that way.

Question 25: If we have a lot of features (100 for example), is there any technique to select the most important ones in python for random forest?

Answer 25: Not sure if you mean “predictors” (possibly spectral bands) or if you mean classes in the output classification.

After you trained a model, you can determine a feature importance that indicates how much each feature contributes to the model prediction. It determines the degree of usefulness of a specific variable for a current model and prediction. You can compute the feature importance with the random forest model used here.

Question 26: How do you provide a GEOTIFF image as an input to CNN in deep learning?
Answer 26: Generally the methods are to transform the GeoTiff image into what the CNN expects, for example with a deep learning framework such as Pytorch it expects a dataformat such as a numpy array that will be turned into a Tensor (Pytorch’s array type). Reading in a geotiff with python using something like GDAL or Rasterio will automatically convert the array to a numpy array.

Question 27: Around Lake Powell, there are some pixels that are water, but not being identified in the model. What are some suggestions for improvements to the model to get more water pixels identified? How good is the ML result in classifying water compared to traditional NDVI/NDWI thresholding?
Answer 27: We trained this on a very very small amount of data. We are using a small dataset just for the demonstration. In order to get a more accurate model, the first step would be to add more training data.

Question 28: Is permutation more informative than feature importance results different for Random Forest Classifier?
Answer 28: A random forest’s feature importance is calculated based on the decrease in impurity that each feature causes when it’s used to split the data in a node of the tree in the forest. Permutation importance is model-agnostic and calculates feature importance by randomly shuffling a feature’s values and measuring the effect on the model’s importance vs when the feature is not randomly shuffled.

The advantage of permutation is it’s more accurate and unbiased since it measures importance based on model performance rather than just looking at the number of times a feature might be used for splitting the data in a node in a tree.

Question 29: Is it possible to plot an ROC curve only from precision, recall, and f-score? What are the other parameters to explain the result other than ROC?
Answer 29: ROC curves are specific for the True Positive and False Positive rates. However, you could plot Precision-Recall curves to understand the performance of your model. F-score is a combination of both precision and recall, thus it might not be as beneficial as general ROC curves. Here is a simple example with some information on the topic:
https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/

Question 30: What do negative values in surface reflectance maps indicate? How should they be handled?
Answer 30: The valid range for the MODIS Surface Reflectance product according to the user guide is -100 - 10000. In conversations with the developer he has suggested that a value of “-100” is statistically close to “0” and therefore the values are still “valid”. The actual negative values are retained to give some indication of “real” variability between adjacent pixels (as opposed to simply binning them all to “0” for example would lose information).

Question 31: Are there any potential reasons that the lake in the lower left hand corner of the tile of interest was not captured by the model?
Answer 31: Simplest answer is insufficient training of the model to capture that particular feature, we only have 1000 observations which is definitely not representative enough for the model to classify all water pixels.

Question 32: How do you extract data from raster files in different buffers of NDVI for a given lat long?
Answer 32: You could apply your NDVI buffers and select pixels based on your domain knowledge. Given the threshold you specify for the features you are looking for, you could assign a label value to that particular lat long location and then extract that pixel into a dataframe.

Question 33: In EDA, the outliers were due to land or due to skewed distribution. How would we know which one was the reason exactly?
Answer 33: Generally it will depend on the dataset in hand. In this case it was because of land. You will be able to know exactly by taking some of the outliers detected and

looking at them in a map using your domain knowledge (e.g. instrument errors, anomalous landscape, etc.)

Question 34: Can you tell me more about the dataset module? How do we use it for downloading other data? I mean, how do we figure out the 'DATASET_URL' ?
Answer 34: You can identify a different DATASET_URL by searching for it on HugggingFace. HuggingFace has several datasets dedicated to machine learning model training https://huggingface.co/. You can also upload your own dataset to HuggingFace using data downloaded from other sources like EarthData and doing your own preprocessing.

Question 35: How can we use ungridded data like MODIS Level 2 data for the machine learning data input?
Answer 35: Ungridded data is highly challenging to work with because it is ungridded. There are tools that can put these data onto a “mesh” so that you can interrogate them more easily. I suggest looking at Healpix https://healpix.sourceforge.io/ or STARE https://www.earthdata.nasa.gov/esds/competitive-programs/access/stare

Question 36: How do I decide what technique can be used for prediction (like XGBoost, Random Forest, Neural Network)?
Answer 36: Take a look at Session 1 slides where we provide some questions based on your science problem that will allow you to take a more informed decision. Some people might recommend using a k-fold cross validation method (on a subset of the training dataset) on all the possible models you have available. The model that gives you the best score is the one you can consider to train the entire training set. Unfortunately, it really becomes a “trial and error” problem once you select the model.

Question 37: I would like to know how one can extract a library of spectra per feature, and then apply it without linking it to specific train/test points. Reason being, one doesn't know if the feature may have shifted - in estuaries vegetation could be highly dynamic.
Answer 37: There are many tools available to do this in image processing software such as ENVI, QGIS, etc. And you could certainly program this in Python using GeoPandas or similar to accomplish your goals.

Question 38: When using Random Forest in Regression mode, which is the rmse that is considered a 'valid' accuracy with respect to the input data?
Answer 38: Acceptable RMSE is subjective and determined by the PI.

Question 39: Why did you use 400 n_estimators in the RandomforestClassifier model? I understand n_estimators must be equal to the number of features (i.e., X values in the model).
Answer 39: 400 was relatively arbitrary, the default is 100 for RandomForestClassifier but it is up to the user to determine the right number. The n_estimators do not have to be equal to the number of features. That hyperparameter represents how many trees will be built in the forest. Generally the more trees you have, the more generalized a model you might get since we are not relying on the prediction of only a few trees.

Question 40: I did the question #2 and the differences I found are substantial. I plotted a histogram of the K Mean classification. That is, a histogram with only 4 'bars' (KMeans was run for 4 clusters). When I do it with the GPU 2 of the 4 clusters barely have any assigned pixels. When using CPU they get hundreds of thousands of pixels in each of the 4 clusters.
Answer 40: This is an interesting finding. My suggestion is to set a static seed between the two to compare the performance of both devices.

Question 41: What would be some other good indications of overfitting beyond a relatively high prediction accuracy score?
Answer 41: It will likely be obvious in the output map that the results are not as you would have expected in spite of a high correlation in training and testing of the model itself. You could see how the model is mimicking the training data when performing predictions. You can also take a look at the ROC curves to understand the rate of the model, and use XAI techniques to look at some of the samples, overconfidence in predictions is oftentimes an indication of overfitting.

Question 42: Do the expectations/process of EDA differ when doing GEOBIA as opposed to pixel-based classification?
Answer 42: The process should be very similar across both use cases. EDA will help you find insight, trends, etc. that will allow you to understand the performance of your GEOBIA workflow. EDA will help you with outliers and to better represent the input data stack for GEOBIA.

Question 43: Are there any possibilities to extract the value of multiple indices using the training sample? If I have five land cover types with 100 points each, can I extract multiple point values to each of them and use it as classification?
Answer 43: Yes, short answer is that you can simply use the column features from your dataset and calculate the indices you need. From there, you can set a threshold for the features of interest and create a discrete scale with them. Ideally, you will want to validate them after the threshold to make sure the classification corresponds to the actual value from the map.

Question 44: What are the guidelines to use different cross validation methods for different models? Like tenfold, fivefold?
Answer 44: The web page:

https://vitalflux.com/k-fold-cross-validation-python-example/

explains well what is the k-fold cross validation. It describes how the method is performed. The question is how do we select the value of k? For use, it is problem dependent. It is recommended to start with the value of k=10. However, if the data set is very large, k=5 can be considered.

If we are not sure which ML model to choose, it is a good approach to apply the k-fold cross validation on a set of ML models using a “small”” (but representative) data set. The model that provides the best score will be considered with the entire data set.

Question 45: How can we extract the area of each land cover class after classifying it with machine learning algorithms without exporting to any GIS/QGIS for vectorization?
Answer 45: You could easily use a combination of numpy and gdal to count the occurrence of pixels and then calculate the area based on the spatial resolution of the data in hand. Another option is to vectorize your data and calculate the area that way. Here is an example related to change detection:
https://hatarilabs.com/ih-en/land-cover-change-analysis-with-python-and-gdal-tutorial

Question 46: Why do we use Distribution of the probability of predicted values?

Answer 46: We always output prediction probabilities to get an understanding of how confident the model is when outputting predictions.

Question 47: Is the workflow similar for high-resolution datasets like Sentinel/Landsat?
Answer 47: The workflow itself would be exactly the same, only the inputs would change. The steps of the workflow we presented (EDA, splitting the data set, training the model, testing the model, etc.) here can be applied to any input data. However, we need to adapt each step to the specific input data we have.

Question 48: Can we use the precision, recall, and f-score interchangeably with kappa, overall accuracy?
Answer 48: Definitely no, they represent somewhat different things so you will need to make sure you use the appropriate measure for your project.

Question 49: In the session 1 Colab file, after we concat all the raster_arrays, while plotting it, why do we only plot the first 3 rasters from it? We can plot any 3 right?
Answer 49: Yes, you can choose any 3 bands from the raster. We choose the first three for simplicity, but you can read any three of your choice to represent your RGB image.

Question 50: Can I use autocorrelation if I am dealing with a single variable? How can I interpret the results from autocorrelation?
Answer 50: The only place where you can use autocorrelation with a single variable is when you have a time series of that variable and want to compare it across time. The idea of autocorrelation is to compare two variables. Depending on the type of correlation (different algorithms have different ranges), the lower the value the closer you will be to negative autocorrelation, the higher the value the closer you will be to positive autocorrelation.

Question 51: How can we learn such variation in soil/water constitution if data are skewed as they were in the histograms,to classify water/ground,is there any optimization for wavelength analysis to solve this problem,if it is the origin?
Answer 51: You could perform exploratory data analysis using signal processing techniques for visualizing the actual wavelength. In our case, the data can be skewed

per band, but the model is able to find those patterns for soil and water pixels based on the combination of the bands and not only one of them individually.

Question 52: Is there a measure of data quality based on the amount of missing data? For example, if a data has more than 20% missing data, is this data set suitable for modeling or not and we should go back for another data set.
And also there are categories for proper methods of recovering lost data? By lost data recovering I mean imputation of missing data.
Answer 52: Each band in MODIS surface reflectance has a QA layer that describes the missing data so yes one could eliminate particular bands from training (and prediction). All bands used in training must be present in the prediction phase otherwise the prediction will fail.

Interpolation is always possible. Whether it is advisable is subjective and determined by the project PI.

Question 53: When we manually do a binary classification (here, water and no water) from a surface reflectance scene, how do you manually pick the training pixels for a non-water class (such as urban, forests, barren, etc.)? Does the diversity in training pixels affect the final classification output?
Answer 53: In the case of this exercise the training pixels were, in fact, manually selected by drawing polygons in QGIS and extracting values under the polygons. There are other more automated ways after you have locations and/or labels to work with but since this was a concise and finite example it was easy enough to just do it manually.

Yes the diversity in training pixels absolutely determines the quality of the output. The goal of your training pixels is to capture the spectral diversity of the features you want to predict, this will get you the optimum model.