Data science project - E9 report

Project name: Stanford Ribonanza RNA Folding
Team members: Pärl Pind, Marlene Ibrus

Task 1 - Github repository

Github repository: https://github.com/marleneibrus/Andmeteaduse-projekt

Task 2 - Business understanding

Identifying our business goals:

Background

Our project is from the Kaggle competition Stanford Ribonanza RNA Folding. The aim of our project (and the competition) is to create a model that predicts the structures of any RNA molecule and the resulting chemical mapping profile, which can be compared to data collected for each position in the RNA. In other words, we need to predict the chemical reactivity at each position of an RNA molecule.

Business goals

The goal is to better understand how RNA works and how to manipulate it and predict RNA structure. Try to identify unique RNA-based drug targets in the many bacterial, viral, neurological, and cancer genes that remain undruggable at the protein level. Predictively design RNA-based medicines such as mRNA vaccines and CRISPR gene therapeutics that promise to treat nearly all human disease. Full understanding of life requires a full, predictive understanding of RNA.

Business success criteria

The success of the resulting model will be measured by comparing the predictive values of the submission against experimentally obtained reactivity values. There is no threshold for a successful model in the Kaggle competition. The most accurate model (model with the lowest mean absolute score) is considered the most successful.

Assessing our situation:

Inventory of resources

Kaggle notebooks including tutorials from other people, such as instructions on where to begin, explanations on understanding the competition. Kaggle Discord server, where other contestants are discussing different methods and strategies, ways to approach this problem.

Datasets provided by the Kaggle competition, full of hundreds of thousands of rows for train data. Similar open-source models focused on predicting molecule structure (for example AlphaFold-v2 for protein folding).

Requirements, assumptions, and constraints

Kaggle submission deadline on the 7th, project deadline for Data Science on the 11th.
The Kaggle submission file must include a predicted reactivity for each RNA sequence position id of the test data set, two values per row, one corresponding to the DMS_MaP reactivity, one corresponding to the 2A3_Map reactivity. In other words, the submission file should include twice as many predictions as is the sum of the lengths of all of the sequences in the test file.

Risks and contingencies

One of the main risks is that the training time for our model will be too long. This can be solved by opting for a simpler machine learning model that requires less computational resources or by training our model based on batches of data and not the entire dataset. Another risk involved is overfitting the model. This can be solved by using a simpler model and by using as much of the data as possible.

Terminology

- RNA nucleotide - ribonucleic acid, that is made up of a sugar molecule (ribose), a phosphate group and a nitrogenous base: adenine (A), guanine (G), cytosine (C) uracil (U).
- RNA molecules - molecules composed of usually a single strand of nucleotides, essential for various biological roles.
- RNA folding - the process by which a single-stranded RNA molecule forms specific shapes or structures by pairing with complementary regions within itself. The way the RNA folds determines its function.
- RNA structure - folding pattern of the RNA molecule.
- RNA sequence - the specific order of nucleotides within an RNA molecule. This determines the genetic information encoded in the RNA, leading to the structure and function of the RNA.
- Chemical mapping - using various chemicals to analyze the structure of RNA molecules. The chemicals' reaction reveals information about RNA folding and structure.
- DMS_MaP - Dimethyl Sulfate-Mutation Profiling, technique that utilizes DMS to analyze RNA structure.
- 2A3_MaP - a technique that utilizes 2-aminopyridine-3-carboxylic acid imidazolide (2A3) as a probing agent to analyze RNA structure.
- Knot theory - a mathematical branch focused on closed curves in three-dimensional space. This relates to RNA folding, because RNA molecules form closed and open knots.

Costs and benefits

We think that this section is not relevant to our project, as all of the resources that we need have been provided to us free of charge.

Defining data-mining goals:

Data-mining goals

The main data mining goal of this project is to produce a regression model that predicts RNA molecule structure based on its sequence. Another goal is to produce a poster for the Introduction to Data Science (Sissejuhatus andmeteadusesse) - LTAT.02.002 course.

Data-mining success criteria

In the Kaggle competition, a mean absolute error (MAE) is calculated for the submission file, based on which the submissions are ranked, in ascending order. The experimental reactivity values for the test dataset are what the predicted reactivity values will be compared to. The formula for calculating MAE provided in the competition is the following:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \widehat{y_i}\right|$$

where where N is the number of scored ground truth values, and y and $\widehat{y}$ are the actual and predicted values, respectively. The y values will be clipped between 0 and 1 before calculating MAE, that is:

$$y_i = max(min(y_i^{RAW},\ 1.0)0.0)$$

where y^{RAW} are the raw data values. The model with the lowest MAE after the private test fold has been revealed is considered to be the most successful.

Task 3 - Data understanding

Gathering data:

Outline data requirements

For our data mining task, we need sufficient data for training our model based on different RNA sequences and their reactivities to the chemical mappings DMS_Map and 2A3_MaP (so for one sequence, one line for each mapping, two rows in total). This means that we need structured numerical experimental data and nominal data about the type of chemical mapping performed. There is no required time range for this data. The necessary data needs to be in CSV format and transformed to Parquet format for smaller size and smaller reading and writing times.

Verify data availability

Luckily for this project, the necessary data is already provided by the Kaggle competition. Since we are expected to use the available materials, we have no need to look for more data/any substitutes/create our own experiments with RNA chemical mappings. All of the necessary data is accessible. For our purposes, we will even need to limit the amount of data and use smaller batches to save on time and RAM.

Define selection criteria

In our gathering data stage, we discovered that some additional folders were too big in size for us to be able to use. The total amount of data provided in the Kaggle files was close to 92 GB, meaning the memory of our computers was not prepared to handle this. The main files needed for this project that we decided to go with were too big for GitHub, therefore, we had to use the GitHub Large Files extension. Even more, merely downloading the full sized data and unzipping it took hours, pushing it to GitHub would have taken even more. That is why we decided to limit the amount that we would use in our own work. Moreover, we converted the train and test datasets into the Parquet format from CSV, to reduce the size. We chose to use the provided cleaned train dataset in Kaggle, from which all of the columns except the "dataset_name" column will be necessary for our model.

Describing data:

Our data for this project comes from the supplied Kaggle files and folders. We will describe three of the available files a little further.
train_data (1643690 rows, 419 columns) - the training data, with one experimental profile per row. Includes columns sequence_id (string, identifier for each sequence), sequence (string, describes the RNA sequence), experiment_type (string, either DMS_MaP or 2A3_MaP to describe the type of chemical mapping experiment that was used to generate each profile), dataset_name (string, name of sequencing dataset from which the reactivity profile was extracted), reads (integer, number of reads that were assigned to the RNA sequence, and whose mutations were tabulated to compile the reactivity profile), signal_to_noise(float,

signal/noise value for the profile), SN_filter (boolean, depending on whether the profile has a high signal to noise and reads), reactivity_0001, reactivity_0002… (float, defines the reactivity profile for the RNA, the type of data that needs to be predicted), reactivity_error_0001, reactivity_error_0002,…(float, errors from reactivity)

test_data (1343823 rows, 5 columns) - the test set sequences, without any columns associated with the ground truth. Includes columns id_min, id_max (integer, minimum and maximum id values for the test sequence in a correctly formatted submission file), sequence_id (string, identifier for each test sequence), sequence (string, describes the RNA sequence), future (boolean, sequences whose data will be collected for the Kaggle competition)

train_data_QUICK_START (335616 rows, 416 columns) - filtered version of train_data.csv based on SN_filter. Includes only sequences that pass both DMS/2A3 SN_filter, some unnecessary columns are removed, duplicates are dropped.

Exploring data:

In the train_data, there are 806573 unique sequences and 2 experiment types for each of them (2A3_MaP and DMS_MaP). For the different datasets, the most frequent ones are DasLabBigLib_OneMil_OpenKnot_Round_2_train_DMS and DasLabBigLib_OneMil_OpenKnot_Round_2_train_2A3
with 125123 usages each. Most of the datasets are from DasLabBigLib, however, there are also datasets from OpenKnot, PK50 and PK90. The smallest ones are from PK90 with 2173 rows. For the reads column, the median is 148, while the mean is 2230. This means that the distribution is right-skewed. The maximum value for reads is 107394, but there are 37867 rows with the value 0, the filtered data train_test_QUICK_START takes into account only the rows with reads higher than 100, meaning that this huge range and skewage has already been taken care of. Since the filtered data also only looks at the data that has passed the SN_filter, we have already eliminated the rows with higher noise.

Verifying data quality:

The data we need for this project exists and we have access to a lot of it. Our main obstacle is concerning the size of all the data, which exceeds the memory and power our computers have available. The train_data_QUICK_START.csv file is already filtered and we do not have any major issues with it that would hinder our work.

Task 4 - Project planning (100 - 300)

Necessary steps for completing this project:

- Converting data into parquet mode for smaller datasets and faster loading times. Storing the necessary data in our project repository.
  - 2 h Marlene, 2 h Pärl
- Data preprocessing: the data has already been split into a train and test dataset, but we need to split the training data to include a validation dataset. We have been provided with a training dataset that has already been cleaned.
  - 2 h Marlene
- Feature selection:
  - 5 h Marlene, 5 h Pärl
- Model selection:
  - 2 h Marlene, 2 h Pärl
- Training the model
  - 3 h Marlene, 3 h Pärl
- Model tuning and improvements:
  - 6 h Marlene, 6 h Pärl
- Validation on test set:
  - 1 h Marlene, 3 h Pärl
- Visualization and interpretations of results:
  - 3 h Marlene, 4 h Pärl
- Organizing and annotating our code repository, creating the poster: making necessary comments to our code, including a read-me file that explains our process.
  - 5 h Marlene, 5 h Pärl
- Submit submission file to Kaggle
  - 1 h Marlene
- Submit project into our course
  - 1 h Pärl

The actual time needed to complete these tasks could be higher, since the amount of data could cause the training time to be very high.