2023 Ariel Data Challenge Solution

Team Name: The Gators

June 22 2023

Below we answer the specific questions from the organizers.

1 Did you use the training data from ADC2022?

No, we did not. We only used the training data from the 2023 challenge.

2 How did you use the training data?

We used a suitable analytical ansatz for the posterior distributions and used the provided ground truth distributions to derive the values for the parameters of the ansatz. Those parameters were later used as labels for our regression models.

We used the labelled training data to build a main regression model which predicts the parameters of our posterior ansatz, given the spectral information and the auxiliary parameters alone.

For the temperature predictions only, we also used the remaining unlabelled data in a semi-supervised approach, where we first build a second regression model which also includes the FM parameters among its inputs; then we used that supplementary model to derive pseudolabels for the unlabelled data, and finally we retrained our main regression model with both the labelled and pseudolabelled data, in order to improve our predictions for the temperature.

3 Did you perform any data preprocessing step?

1. For planets with spectral values higher than 0.1, we replaced these anomalously high values with a value constructed from the other (lower than 0.1) spectral values.

2. In order to focus on the effect of the atmosphere, we try to subtract the contribution of the planet itself, which we approximate using the spectral bin with the lowest value:

$$M'_\lambda = M_\lambda - \min_\lambda (M_\lambda) \tag{1}$$

3. We used the analytical solution of the thermal equilibrium between the star and planet to estimate the equilibrium temperature T_p of the planet from the auxiliary parameters alone. We constructed the additional features $\hat{R}_P = R_S\sqrt{\min_\lambda (M_\lambda)}$, $\frac{\hat{R}_P}{R_S}$, $\frac{D}{H}$, $\frac{R_S}{H}$, and $\max_\lambda (M'_\lambda)$, and concatenated them with the auxiliary parameters. The atmospheric scale height H of the planet is computed with the equilibrium temperature T_p.

4. We standardized the auxiliary features and normalized M'_λ and the noise data by dividing by $\max_\lambda (M'_\lambda)$.

4 What kind of model did you go for?

Our basic strategy was described in the second question above. More concretely, we used several fully connected neural networks some of which use concatenations or products of the outputs of previous modules as inputs. We train separately for the planetary radius, the temperature and the chemistry. The radius and chemistry models are trained with labelled data only. The temperature model also uses the unlabelled data as explained above.

5 What is the input/output of the model?

The model takes in the 52 wavelength bins of the flux modulation data, the 52 corresponding uncertainties (noise), and the eight provided auxiliary parameters plus the additional features we constructed. The model takes these three vectors as input, and has a total of 117 input neurons.

The model outputs 24 values which are the parameters of our posterior ansatz.

6 Did you do any post-processing to the output?

We constrain the output to be within the interval $[-12, -1]$ for each of the absorber concentrations. We also restrict the radius prediction to be between 0 and 3 R_J, and the temperature prediction to be between 0 and 7000. We also require that the predicted radius is smaller than $\hat{R}_P$.

7 Did you perform any sampling step? If so please describe.

For the radius and temperature we are sampling from a bivariate Gaussian distribution with the predicted means and standard deviations, and covariance 0.7. For each chemical absorber, we sample the corresponding ansatz.

8 Did you use any external library and/or forward model?

We did not use any forward model. We used standard libraries such as pytorch, scipy, pandas, and numpy.