Stellar Characterization using Deep Learning and Machine Learning

Sankalp Gilda
University of Florida
Hon et al. (2017)
CNN
Data Preparation

- 1,440 PHOENIX spectra picked in ranges:
  - Teff: [3500,7000] in steps of 100K
  - log(g): [2,5.5] in steps of 0.5 dex
  - [Fe/H]: [-1,1] in steps of 0.5 dex
  - lambda: [5000,5700] in steps of 0.01A
- **Convolved** with MARVELS PSF in 13 bins of width 50A each. Median resolution in each bin, across 30 calibration stars.
- **Rotational broadening** of dwarfs by 3km/s.
- **Resampled** onto the same grid as MARVELS, with spacing of 0.1A
- **Continuum-normalized** using an iterative Legendre polynomial fitting function, following same prescription as Ghezzi.
- Finally, **interpolated** (using RBF, cubic spline, and trilinear) and step sizes reduced to 25K for Teff, .25dex for log(g), and .25 dex for [Fe/H].
This leaves us with 19,035 spectra (x3), which are then trimmed down to a wavelength range of 5059-5536 Å.

3,082 MARVELS spectra (odd beams, stacked) are continuum normalized as well. To deal with curvature on the blue end, the final wavelength range is selected as 5200-5532 Å, in steps of .1Å (3200 features)
CNN architecture

- 2 architectures were tested: a 9-layer net and a 15-layer net. The latter was considerably slower (40s vs 200s per epoch) and took longer to converge but gave better results.
- Training time for 100 epochs on 9-layer net was ~1hr, and on the 15-layer net was ~5hrs.
Pre-processing

- Before feeding in the $19,035 \times 3,200$ array to the net, both the input data (normalized spectra) and the output data (1d arrays of atmospheric parameters) were normalized along each column.
- Standard operating procedure in DL.
MC2 / KEPLER3.TRES-2, T_eff = 5598, log(g) = 4.44, [Fe/H] = .40
Data Hungry NN – Augment!
Application

- **Regression**: The same neural net was used 3 times, once each for each of the three parameters.
- Following Ciardi et al. 2011, a star is a dwarf, if:
Figure 1. Top: KIC-based Surface Gravity–effective temperature H-R diagram of the stars in the analysis sample. The dashed black line marks the delineation to separate dwarfs (blue) and giants (red). Center: histograms of the surface gravity for the dwarfs (blue) and giants (red). The vertical dashed lines mark the median surface gravity values. Bottom: histograms of the effective temperatures for the dwarfs (blue) and giants (red). The vertical dashed lines mark the median temperature values.
RMS = 104 K
mean absolute error (MAE) = 75 K
Distribution of log(g), for 30 MARVELS calibration dwarfs

RMS = 0.12 dex
MAE = .11 dex
RMS = 0.09 dex
MAE = 0.08 dex
Distribution of $T_{\text{eff}}$ from CNN, for 2,233 MARVELS RPMJ dwarfs

![Scatter plot showing the distribution of $T_{\text{eff}}$ from CNN against $T_{\text{eff}}$ from spectral indices. The data points are scattered across the plot with a linear trend line.]
Distribution of \([\text{Fe/H}]\) from CNN, for 2,233 MARVELS RPMJ dwarfs
Results

- MAE (Deep net, 100 epochs)
  - Synthetic: $\log(g) = 0.015$ dex
    $T_{\text{eff}} = 10$ K
    $[\text{Fe/H}] = 0.03$ dex
  - MARVELS: $\log(g) = 0.11$ dex
    $T_{\text{eff}} = 85$ K
    $[\text{Fe/H}] = 0.08$ dex
Issues with DL or: Enter ML

- Data Hungry
- Black-box
- Parameter-Sensitive
- Potentially long training times
• Depending on the algorithm, ML can work with much less data
• Intelligent feature selection
• Ensembles (in progress)
• Potentially quicker execution
Feature Selection

- A KNN based RF algorithm is designed to iteratively decrement the number of features
- 5 fold CV for validation
MAE for Teff for 30 MARVELS stars

MAE

num_of_features

5000 4000 3000 2000 1000 0
flux for 5820K, 4.03 logg, -0.24 Teff

wavelength in angstroms
The road ahead...

- Correlate selected features with known absorption features
- Explore the relationship between number of calibration spectra, resolution, SNR, and parameter space
- Work on end-to-end pipeline