Nearest Neighbour Regression Outperforms
Model-based Prediction of Specific Star Formation Rate

Kristoffer Stensbo-Smidt\textsuperscript{1} (k.stensbo@iku.dk), Christian Igel\textsuperscript{1}, Andrew Zirm\textsuperscript{2}, Kim Steenstrup Pedersen\textsuperscript{3}
\textsuperscript{1}Image Group, Department of Computer Science, UCPH; \textsuperscript{2}Dark Cosmology Centre, Niels Bohr Institute, UCPH

Objective

The goal is to test whether large-scale machine learning techniques can accurately estimate the specific star formation rate (sSFR) from colours of low-redshift galaxies.

Motivation

Astronomy faces typical big-data problems:

• Astronomical surveys have already produced images of $\sim 10^8$ objects, and will soon produce more than 30 TB of data per night.
• While spectra contain most information, they are very expensive to acquire and thus only available for $\sim 10^8$ objects. Relevant properties must therefore be inferred from less informative images.
• Near real-time data processing is necessary to allow for time-critical follow-up observations.

Thus, efficient data mining techniques are required.

Specific Star Formation Rate Prediction

• Specific star formation rate (sSFR) is the number of stars being formed per year in a galaxy, normalised by the stellar mass of the galaxy.
• The sSFR is an important quantity because it is linked to the galaxy evolution.
• The state-of-the-art astronomical model (referred to as the template-based model) for sSFR prediction uses simulated spectra to estimate the sSFR.

Experiments

Data come from the Sloan Digital Sky Survey (SDSS) DR7 and consist of:

• Colours, derived photometrically (see figure 1)
• sSFRs, derived spectroscopically

Smaller SDSS Subset

Samples: 11 450 galaxies, for which template-based model predictions are available.

Algorithm: k-nearest neighbours (k-NN) regression.

Evaluation: Cross validation (CV) consisting of:

• an outer 10 fold CV, for evaluation, and
• an inner 10 fold CV, used to determine the best $k \in \{1, 3, 5, \ldots, 25\}$.

Larger SDSS Subset

Samples: All 694 894 galaxies in SDSS with ground truth sSFRs; 30 000 are randomly selected as training set.

Algorithm: k-NN regression, $k = 21$.

Results

• k-NN regression performs significantly better than the template-based model.
• The template-based model undershoots the low sSFR peak.
• Extending to the full SDSS subset does not lead to a significant reduction of the k-NN accuracy.

<table>
<thead>
<tr>
<th>Model</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Template-based model\textsuperscript{a}</td>
<td>0.36 $\pm$ 0.01</td>
</tr>
<tr>
<td>k-NN regression, smaller subset\textsuperscript{a}</td>
<td>0.30 $\pm$ 0.01</td>
</tr>
<tr>
<td>k-NN regression, larger subset\textsuperscript{b}</td>
<td>0.30</td>
</tr>
</tbody>
</table>

\textsuperscript{a} Results based on the smaller SDSS subset of 11 450 galaxies.

\textsuperscript{b} Results based on the larger SDSS subset of 694 894 galaxies.

Conclusions

• Machine learning can predict the sSFR of galaxies from photometric data with a high accuracy.
• k-NN regression is a viable model for many astronomical applications, because of small feature space dimensionalities.
• The superior k-NN results indicate that there is room for improvement of the standard physical models for sSFR prediction.

Acknowledgements

We gratefully acknowledge support from The Danish Council for Independent Research | Natural Sciences through the project “Surveying the sky using machine learning” (SkyML).

References