Classification and parameter estimation in astronomy

Coryn Bailer-Jones
Max Planck Institute for Astronomy, Heidelberg

UCL, 4 February 2008
Overview

- learning from astronomical data
- Gaia Galactic survey mission
- object classification
- astrophysical parameter estimation
- optimizing the survey: heuristic filter design
- summary

http://www.mpia.de/homes/calj
Acknowledgements

- Gaia @ MPIA group
  - Christian Elting
  - Paola Re Fiorentin
  - Kester Smith
  - Carola Tiede
  - Vivi Tsalmantza

- More information on Gaia:
  http://sci.esa.int/Gaia
Astronomical spectra

Coryn Bailer-Jones, MPIA, Heidelberg

images from G. Smith, UCSD and J. Dalcanton, Washington
What can we learn from spectra?

- object type (classification)
  - overall spectral shape
  - characteristic features
- stars
  - atmospheric temperature, pressure, abundances
  - derive mass, age, radius (via evolutionary models)
  - interstellar extinction (along line-of-sight)
- galaxies
  - stellar and gas composition of galaxy
  - infer evolution
- quasars
  - redshift (cosmological distance)
  - use as background lights to study cosmology
Astronomical surveys
The Gaia Galactic survey mission
The Gaia Galactic survey mission

- five-year all sky survey of one billion objects
- detect objects in real time and measure
  - distances and space velocities
  - spectra (to determine physical properties)
- Science goals
  - structure, origin and evolution of our Galaxy
  - nature of dark matter
  - better understand stars and star clusters
  - find planets outside of the solar system
- strength of Gaia is both accuracy and large number statistics
  - distances to accuracy of 1% for over 10 million stars
Measuring distances via parallax

parallax measurements are relative to reference (background) objects
Quasar identification

- extragalactic objects at large distances
- used to define an inertial reference frame ("fixed background")
- characteristic broad emission lines
- position of lines in spectrum varies due to redshift
The classification problem

- "blind" survey
- object (spectrum) classification via pattern recognition
- four class problem
  - star, quasar, galaxy, physical binary star
- support vector machine classifier (libsvm)
  - RBF kernel
  - multiclass using the one-against-one method
- 8000 objects (~2000 of each class) in training set, same for test
Data simulation

instrument simulator including noise

synthetic spectra

training set

fluxes and estimated covariances

test set

Coryn Bailer-Jones, MPIA, Heidelberg
Example Gaia spectra

- Blue: star
- Red: quasar
Gaia spectral data

- blue: star
- red: quasar
- solid: median
- dashed: upper and lower quartiles
Classification results: contingency table

<table>
<thead>
<tr>
<th></th>
<th>galaxy</th>
<th>physbin</th>
<th>quasar</th>
<th>star</th>
</tr>
</thead>
<tbody>
<tr>
<td>GALAXY</td>
<td>99.65</td>
<td>0.00</td>
<td>0.35</td>
<td>0.00</td>
</tr>
<tr>
<td>PHYSBIN</td>
<td>0.00</td>
<td>94.38</td>
<td>0.15</td>
<td>5.46</td>
</tr>
<tr>
<td>QUASAR</td>
<td>2.93</td>
<td>0.05</td>
<td>96.81</td>
<td>0.21</td>
</tr>
<tr>
<td>STAR</td>
<td>0.10</td>
<td>13.99</td>
<td>3.16</td>
<td>82.75</td>
</tr>
</tbody>
</table>

- stars and physbin (physical binaries) are hard to distinguish, even in principle
- use output probabilities to vary completeness and contamination of a sample
Classifier confidence

how well does it find the correct objects?

\[ P(\text{class} \mid \text{CLASS}) \]
Classifier confidence

how well does it find the correct objects?

\[ P(\text{class} | \text{CLASS}) \]

how much many false positives in the quasar sample?

\[ P(\text{quasar} | \text{CLASS}) \]
Sample completeness & contamination

select threshold to build a sample

blue = completeness
red = contamination

threshold probability
Classifier confidence

P(\text{quasar} \mid \text{CLASS})

small but high confidence contamination of quasars by stars

classifier output probability
What is contaminating the quasar sample?

- **blue:** star
- **red:** quasar
- **solid:** median
- **dashed:** upper and lower quartiles

black: stars classified as quasars with high probability
Meet reality: population fractions

- in reality quasars are rare compared to stars (ca. 1000 times)
- don't modify training sample distribution/re-train
- instead
  - modify priors
  - treat SVM outputs as likelihoods
Meet reality: population fractions

- in reality quasars are rare compared to stars (ca. 1000 times)
- don't modify training sample distribution/re-train
- instead
  - modify priors
  - treat SVM outputs as likelihoods

\[ P^{\text{mod}}(i) \propto P^{\text{nom}}(i) \frac{f^{\text{mod}}_j}{f^{\text{nom}}_j} \]

- \( P(i) \) = posterior probability that object is in class \( i \)
- \( f_j \) = prior fraction of objects in true class \( j \)
- \( \text{mod} = \text{modified}, \text{nom} = \text{nominal} \)
Classifier confidence: nominal vs. modified

\[ P(\text{quasar} \mid \text{CLASS}) \]

**modified population**

**nominal population**

Coryn Bailer-Jones, MPIA, Heidelberg
blue = completeness  
red = contamination  
relative fractions:  
star 1  
physbinary 1  
quasar 0.001  
galaxy 1
Modified population fractions (priors)

blue = completeness
red = contamination
solid = nominal fractions (equal)
dashed = modified fractions (quasars 1000 times rarer)

threshold probability
Comparison of different ML models

<table>
<thead>
<tr>
<th></th>
<th>Spectra only</th>
<th>Spectra + Astrometry</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>K=4</td>
<td>K=3</td>
</tr>
<tr>
<td>SVM</td>
<td>10.4</td>
<td>2.5</td>
</tr>
<tr>
<td></td>
<td>8.3</td>
<td>0.6</td>
</tr>
<tr>
<td>Boosting</td>
<td>37.8</td>
<td>33.9</td>
</tr>
<tr>
<td></td>
<td>13.0</td>
<td>1.5</td>
</tr>
<tr>
<td>MLP</td>
<td>9.4</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td>7.5</td>
<td>0.2</td>
</tr>
<tr>
<td>Mclust</td>
<td>10.9</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td>9.4</td>
<td>0.1</td>
</tr>
<tr>
<td>RBF</td>
<td>38.5</td>
<td>24.0</td>
</tr>
<tr>
<td></td>
<td>27.1</td>
<td>15.3</td>
</tr>
</tbody>
</table>

Overall classification error (percent)

Class assignment via highest probability

K=3: 3 classes only (star and phsysbin merged)
Estimating stellar parameters

- several astrophysical parameters (APs) of interest
  - effective temperature, $\text{T}_{\text{eff}}$
  - surface gravity, $\log g$
  - abundance, $[\text{Fe}/\text{H}]$
  - interstellar extinction, $A_v$
- different nature and degree of impact on the spectra
Ranges:

\[ A_v = \{0, 10\} \]
\[ [\text{Fe/H}] = \{-2.5, 0.5\} \]
\[ \log g = \{0.5, 4.5\} \]
\[ T_{\text{eff}} = \{3350, 35000\} \]

Fixed values:

\[ A_v = 0, \quad [\text{Fe/H}] = 0.0 \]
\[ \log g = 4.5, \quad T_{\text{eff}} = 3500 \]
Strong vs. weak parameters

Minimum distance (nearest neighbour) method

**AP ranges:**
- $A_v = 0$
- $[\text{Fe/H}] = 0$
- $\log g = \{-0.5, 5.5\}$
- $T_{\text{eff}} = \{2000, 29000\}$
An inverse problem

log(Teff)

flux in band

Coryn Bailer-Jones, MPIA, Heidelberg
An inverse problem

\[ \log(T_{\text{eff}}) \]

flux in band

Coryn Bailer-Jones, MPIA, Heidelberg
Stellar parameters from SDSS spectra

- train regression model using simulated spectra (true physical parameters are known)
- predict (unknown) parameters of real objects
- model: MLP neural network (statnet)
- large input space (thousands of correlated pixels)
  - dimension reduction via PCA
  - also removes some noise
  - can act as a junk filter
PCA

original spectra 5+5 reconstruction 25+25 reconstruction

Teff=4752, logg=4.30, [Fe/H]=−1.94

Teff=5251, logg=2.13, [Fe/H]=−0.45

Teff=7720, log g=2.85, [Fe/H]=−0.57

Flux

Wavelength

Re Fiorentin, Bailer-Jones et al. 2007
Results: comparison of parameter determinations with independent estimates

Re Fiorentin, Bailer-Jones et al. 2007
Two stellar components in the halo of the Milky Way

Daniela Carollo\textsuperscript{1,2,3,5}, Timothy C. Beers\textsuperscript{2,3}, Young Sun Lee\textsuperscript{2,3}, Masashi Chiba\textsuperscript{1}, John E. Norris\textsuperscript{5}, Ronald Wilhelm\textsuperscript{6}, Thirupathi Sivarami\textsuperscript{2,3}, Brian Marsteller\textsuperscript{2,3}, Jeffrey A. Munn\textsuperscript{7}, Coryn A. L. Bailer-Jones\textsuperscript{8}, Paola Re Fiorentin\textsuperscript{8,9} & Donald G. York\textsuperscript{10,11}

The halo of the Milky Way provides unique elemental abundance and kinematic information on the first objects to form in the Universe, and this information can be used to tightly constrain models of galaxy formation and evolution. Although the halo was once considered a single component, evidence for its dichotomy has slowly emerged in recent years from inspection of small samples of halo objects. Here we show that the halo is indeed clearly divisible into two broadly overlapping structural components—an inner and an outer halo—that exhibit different spatial density profiles, stellar orbits and stellar metallicities (abundances of elements heavier than helium). The inner halo has a modest net prograde rotation, whereas the outer halo exhibits a net retrograde rotation and a peak metallicity one-third that of the inner halo. These properties indicate that the individual halo components probably formed in fundamentally different ways, through successive dissipational (inner) and dissipationless (outer) mergers and tidal disruption of proto-Galactic clumps.
Incorporating domain knowledge

- much redundant information in spectra
- can directly extract sensitivity of inputs (pixels) from simulated data (physical models)
  - soft dimension reduction
- use in various ways
  - data weighting (global)
  - local regression via local weighting (iterative) [in progress...]
- ideal case
  - design optimal filters to use in telescope in first place
Photometric filters

Transmission

Wavelength

SDSS
CFHTLS

u g r i z

image from R. Sharp, IoA
Heuristic filter design

- **algorithm**
  - parametrize filter system (three per filter)
  - establish a figure-of-merit (FoM) of filter system performance
  - maximize FoM with respect to filter system parameters

- **optimizer**
  - cannot define derivatives
  - local minima guaranteed!
  - evolutionary algorithm (genetic algorithm)
Figure of merit (fitness)

- space: flux in each filter of a filter system
- maximum fitness: local directions of variance of the different astrophysical parameters are orthogonal
- calculated from labelled data set
Fitness evolution

Bailer-Jones 2004
Filter system evolution

Evolution of all filter system parameters
(200*5 for each parameter type at each generation)

Bailer-Jones 2004
Discovered filter system

- unconventional: overlapping filters
- learned filter system competitive in parameter estimation work
- method could be extended to design “software filters” in algorithms (e.g. hierarchical classification system)
Summary

- astronomy driven by what we can learn from data (surveys)
- extract optimal information from spectra
  - classification (star, quasar etc.)
  - estimate physical parameters
- use of machine learning methods
  - probabilistic system for classification (SVM)
  - issues of inverse problem and “weak” parameters
- exploit domain knowledge
  - post hoc adjustment based on population fractions (priors)
  - deriving sensitivities from simulated data
  - EA-based heuristic filter design
- future work
  - nonlinear dimensionality reduction
  - degenerate solutions and error estimation