Automated Stellar Classification

TESSA THORSEN
Automated Classification

Why use automated algorithms?

- Faster classifications
- Less bias (more consistent classifications)
- Potential for higher accuracy

PROBLEM: we have too much data in each spectrum!

12,000 Spectra → Automated Algorithm → 12,000 Classifications
Principal Component Analysis (PCA)

Determine orthogonal basis vectors that preserve as much variance as possible

1\textsuperscript{st} basis vector preserves most variance, n\textsuperscript{th} preserves least variance
Locally Linear Embedding (LLE)

Useful when data has an underlying nonlinear structure of lower dimension that is linear locally

Algorithm unpacks underlying structure of the data

Source: Roweis, 2000
Isometric Mapping (Isomap)

Creates a partially connected graph where point \( i \) is connected to point \( j \) if it is one of its \( k \) nearest neighbors.

Approximate distance between far points as distance along graph from one point to the other.

Determine principal component vectors that preserve distances along graph.
Classification Algorithm

Train algorithm on labeled data (half of data set)
  ◦ Construct regions that correspond to classes

Apply algorithm to unlabeled data (other half of data set)
  ◦ Determine class labels based on region into which points fall

Classification of 2D data. Colors correspond to subclasses: points are used to determine colored regions
Data

From SDSS we extracted spectra for stars with:

- Signal-to-noise ratio of at least 10
- Wide range of wavelength coverage
- Subclass of K1, K3, K5, or K7

Processing data:

1. Shift spectra to rest-frame
   \[ \lambda_{emit} = \frac{\lambda_{obs}}{z+1} \]

2. Select region of spectra from 3800 – 9000 Å
   - Reject spectra without this range of coverage

3. Normalize spectra so that total flux = 1
2D Projections

7920 stars

9 stars

3 stars

with Principal Component Analysis
Outliers
PCA Components

First Component

Second Component
PCA Components Cont’d

Third Component

Fourth Component
PCA Approximation

PCA coefficients:

4.33675346e+01, 1.38431234e+01, 3.47935748e+00, 1.04308414e+00, 4.20062542e-02, 4.41135025e+00, 5.80523252e-01, -7.17547536e-01, -2.03134447e-01, -4.43795174e-01
Classification Accuracy

My Results

Results of Bu et al.

Note: $k$ is number of nearest neighbors used
Dependence of Accuracy on $k$

**Accuracy of Isomap Projections**
- Iso: $k = 6$
- Iso: $k = 8$
- Iso: $k = 10$
- Iso: $k = 15$
- Iso: $k = 20$
- Iso: $k = 30$
- Iso: $k = 40$
- Iso: $k = 50$
- Iso: $k = 60$
- Iso: $k = 70$
- Iso: $k = 80$
- Iso: $k = 90$
- Iso: $k = 100$

**Accuracy of LLE Projections**
- LLE: $k = 6$
- LLE: $k = 8$
- LLE: $k = 10$
- LLE: $k = 15$
- LLE: $k = 20$
- LLE: $k = 30$
- LLE: $k = 40$
- LLE: $k = 50$
- LLE: $k = 60$
- LLE: $k = 70$
- LLE: $k = 80$
- LLE: $k = 90$
- LLE: $k = 100$
Comparison of Algorithms

PCA is more accurate than both Isomap and LLE
LLE has high accuracy with high $k$, lower accuracy with lower $k$

Accuracy peaks at
- 5 output dimensions for PCA
- 3 output dimensions for Isomap
- 4 output dimensions for LLE
Conclusion

LLE is effective in finding outliers, not as effective at classifying (unless we use high $k$)

PCA is the most accurate for high numbers of output dimensions

Isomap is most accurate for low numbers of output dimensions
Future Work

Much more work needed!

◦ Why were my results different than expected?
◦ Look at where stars of different magnitudes lie on the manifold
◦ What are the outlying spectra?

References: