Mining Astronomical Massive Data Sets within the Virtual Observatory

Giuseppe Longo

Department of Physical Sciences
University Federico II in Napoli – ITALY
INAF – Italian Institute of Astrophysics
INFN – Italian Institute of Nuclear Physics

Astrostatistics School – Penn State, June 2008
An overview

• methodological introduction to why astronomy needs statistical data mining

  • what I mean for “data mining” and why I believe that statistics and DM will prove crucial for the future of astronomy

  • some clustering methods

  • some applications to observational cosmology
An historical perspective (following Eric’s introduction)

1910
Final settling of stellar statistics, by the work of Kapteyn, Oort, etc.)

S.I.L.

1960’s
Photographic wide field plates (PSS)

XXI century
Renaissance of statistical astronomy (synoptic surveys)

Rush for the larger and the bigger

20’s
Few objects, few λ’s, heterogeneous data

80’s
Virtual Obs.

Palomar

Hooker

J. Kapteyn

LSST (2013)
Most discoveries take place immediately after a technological breakthrough.
And now, the question is.....

Where to search ... for the next discoveries?
And what has statistics got to do with it?
The parameter space

Any observed (simulated) datum $p$ defines a point (region) in a subset of $\mathbb{R}^N$. Es:

- RA and dec
- time
- $\lambda$
- experimental setup (spatial and spectral resolution, limiting mag, limiting surface brightness, etc.) parameters
- fluxes
- polarization
- Etc.

$100 \gg \mathbb{R} \in \mathbb{N}^p$ $N \gg 100$

The parameter space concept is crucial to:

1. Guide the quest for new discoveries (observations can be guided to explore poorly known regions), ...
2. Find new physical laws (patterns)
3. Etc,
The universe is densely packed

1/160,000 of the sky, moderately deep (25.0 in r)

55,000 detected sources
(0.75 mag above m lim)
The scientific exploitation of a multi band, multiepoch (K epochs) universe implies to search for patterns, trends, etc. among N points in a DxK dimensional parameter space:

\[ p = \{ \text{isophotal, petrosian, aperture magnitudes concentration indexes, shape parameters, etc.} \} \]

\[ p^1 = \{ RA^1, \delta^1, t, \{ \lambda_1, \Delta \lambda_1, f_1^{1,1}, \Delta f_1^{1,1}, \ldots, f_1^{1,m}, \Delta f_1^{1,m} \}, \ldots, \{ \lambda_n, \Delta \lambda_n, f_n^{1,1}, \Delta f_n^{1,1}, \ldots, f_n^{1,m}, \Delta f_n^{1,m} \} \} \]

\[ p^2 = \{ RA^2, \delta^2, t, \{ \lambda_1, \Delta \lambda_1, f_1^{2,1}, \Delta f_1^{2,1}, \ldots, f_1^{2,m}, \Delta f_1^{2,m} \}, \ldots, \{ \lambda_n, \Delta \lambda_n, f_n^{2,1}, \Delta f_n^{2,1}, \ldots, f_n^{2,m}, \Delta f_n^{2,m} \} \} \]

............................

\[ p^n = \{ RA^n, \delta^n, t, \{ \lambda_1, \Delta \lambda_1, f_1^{N,1}, \Delta f_1^{N,1}, \ldots, f_1^{N,m}, \Delta f_1^{N,m} \}, \ldots \} \]

\[ D = 3 + m \times n \]

\[ N > 10^9, \; D >> 100, \; K > 10 \]
Every time you improve the coverage of the PS....

Every time a new technology enlarges the parameter space or allows a better sampling of it, new discoveries are bound to take place.
Improving coverage of the Parameter space - II

Projection of parameter space along (time resolution & wavelength)

Projection of parameter space along (angular resolution & wavelength)
Considerations on the next breakthroughs

• We have reached the physical limit of observations (single photon counting) at almost all wavelengths...
• Detectors are linear
• All electromagnetic bands have been opened

Hence

Our capability to gain new insights on the universe will depend mainly on:

• Capability to recognize patterns or trends in the parameter space (i.e. physical laws) which are not limited to the human 3-D visualization
• Capability to extract patterns from very large multiwavelength, multiplepoch, multi-technique parameter spaces

The answer to these needs is the International Virtual Observatory which (like it or not like it) is bound to be implemented and to change the way astronomers work!
Experimental astronomy has become a three players game

- **astronomy**: problems, data, understanding of the data structure and biases
- **statistics**: evaluation of the data, falsification/validation of theories/models, etc.
- **computer science**: implementation of infrastructures, databases, middleware, scalable tools, etc.
The various parts of the V.O. task

Data Gathering (e.g., from sensor networks, telescopes...)

Data Farming:
- Storage/Archiving
- Indexing, Searchability
- Data Fusion, Interoperability, ontologies, etc.

Data Mining (or Knowledge Discovery in Databases):
- Pattern or correlation search
- Clustering analysis, automated classification
- Outlier / anomaly searches
- Hyperdimensional visualization

Data understanding
- Computer aided understanding
- KDD
- Etc.

Database technologies

Key mathematical issues

Ongoing research

New Knowledge
The Curse of Hyperdimensionality

The computational cost of clustering analysis:

- K-means: $K \times N \times I \times D$
- Expectation Maximisation: $K \times N \times I \times D^2$
- Monte Carlo Cross-Validation: $M \times K_{\text{max}}^2 \times N \times I \times D^2$
- Correlations $\sim N \log N$ or $N^2$, $\sim D^k$ ($k \geq 1$)
- Likelihood, Bayesian $\sim N^m$ ($m \geq 3$), $\sim D^k$ ($k \geq 1$)
- SVM $> \sim (NxD)^3$

$N =$ no. of data vectors, $D =$ no. of data dimensions
$K =$ no. of clusters chosen, $K_{\text{max}} =$ max no. of clusters tried
$I =$ no. of iterations, $M =$ no. of Monte Carlo trials/partitions

$N >10^9, D>>100, K>10$

Some dimensionality reduction methods are needed (e.g., PCA, ICA, class prototypes, hierarchical methods, etc.), but more work is needed

Terascale (Petascale?) computing and/or better algorithms
3-D is always better than 2-D

Is it worth the effort? ... YES!
Models implemented in VO-Neural

- **MLP** (Multi layer perceptron): slow, supervised, non linear
- **SOM** (self organizing maps): faster, unsupervised, non linear, great visualization, non physical output
- **GTM** (generative topographic mapping): slow, unsupervised, great visualization, physical output
- **PCA & ICA** linear and non linear: terrible visualization, physical output, good performances on uncorrelated data
- **Fuzzy C Means**: slow on MDSs, effective in “fuzzy problems”
- **PPS**: great (the best ones for unsupervised clustering, classification and visualization)
- **Competitive Evolution on Data** (CED): bad visualization, great accuracy as unsupervised clustering tool, …
- **RBF** and many other methods.
PART II

why Data Mining is crucial to face this data tsunami....
Machine learning methods can be broadly grouped in:

**Supervised methods**

They learn how to partition the parameter space by means of a training phase based on examples.

Neural Networks such as the Multi Layer Perceptron (MLP), Support Vector Machines (SVM), etc.

**Pro’s & Con’s**

- They are good for interpolation of data, very bad for extrapolations
- They need extensive bases of knowledge (i.e. uniformly sampling the parameter space) which are difficult to obtain;
- Errors are easy to evaluate
- Relatively easy to use

- They reproduce all biases and preconceived ideas present in the BoK
Unsupervised (clustering) methods

They cluster the data relying on their statistical properties only. Understanding takes place through labeling (very limited BoK).

Generative Topographic Mapping (GTM), Self Organizing Maps (SOM), Probabilistic Principal Surfaces (PPS), Support Vector Machines (SVM), etc.

Pro’s & Con’s

• In theory they need little or none knowledge a-priori
• Do not reproduce biases present in the BoK

• Evaluation of errors more complex (through complex statistics)
• They are computationally intensive
• They are not user friendly (... more an art than a science; i.e. lot of experience required)
PART III

Three applications to observational cosmology
MINING THE SDSS ARCHIVE. I. PHOTOMETRIC REDSHIFTS IN THE NEARBY UNIVERSE

RAFFAELE D’ABRUSCO,1,2 ANTONINO STALANO,3 GIUSEPPE LONGO,1,4,5 MASSIMO BRESCEA,5,4 MAURIZIO PAOLILLO,1,4 ELISABETTA DE FILIPPIS,5,4 AND ROBERTO TAGLIAFERRI5,4

Received 2006 October 11; accepted 2007 March 2

ABSTRACT

We present a supervised neural network approach to the determination of photometric redshifts. The method was fine-tuned to match the characteristics of the Sloan Digital Sky Survey, and as base of “a priori” knowledge, it exploits the rich wealth of spectroscopic redshifts provided by this survey. In order to train, validate, and test the networks, we used two galaxy samples drawn from the SDSS spectroscopic data set, namely, the general galaxy sample (GG) and the luminous red galaxy subsample (LRG). The method consists of a two-step approach. In the first step, objects are classified as nearby \((z < 0.25)\) and distant \((0.25 < z < 0.50)\), with an accuracy estimated as 97.52%. In the second step, two different networks are separately trained on objects belonging to the two redshift ranges. Using a standard multilayer perceptron operated in a Bayesian framework, the optimal architectures were found to require one hidden layer of 24 (24) and 24 (25) neurons for the GG (LRG) sample. The final results on the GG data set give a robust \(\sigma_z \approx 0.0208\) over the redshift range \([0.01, 0.48]\) and \(\sigma_z \approx 0.0197\) and \(\approx 0.0238\) for the nearby and distant samples, respectively. For the LRG subsample we find instead a robust \(\sigma_z \approx 0.0164\) over the whole range, and \(\sigma_z \approx 0.0160\) and \(\approx 0.0183\) for the nearby and distant samples, respectively. After training, the networks have been applied to all objects in the SDSS table GALAXY matching the same selection criteria adopted to build the base of knowledge, and photometric redshifts for circa 30 million galaxies having \(z < 0.5\) were derived. A catalog containing redshifts for the LRG subsample was also produced.
Photometric redshifts

Photometric system - $S_{\lambda}(\lambda)$

Galaxy spectrum - $F(\lambda)$

$X = \begin{cases} 
  m_U = -2.5\log_{10} \frac{\int F(\lambda)S_U(\lambda)d\lambda}{\int S_U(\lambda)d\lambda} + c_u \\
  m_B = -2.5\log_{10} \frac{\int F(\lambda)S_B(\lambda)d\lambda}{\int S_B(\lambda)d\lambda} + c_B 
\end{cases}$

Etc...

Color indexes

$U - B \equiv m_U - m_B$

$B - R \equiv m_B - m_R$

etc.
The Sloan Digital Sky Survey (SDSS) data set & BoK

8000 sq degrees
>210 million galaxies
data are public

Extensive but biased
spectroscopic BoK:
700,000 galaxy spectra

Benchmark for almost
everything in observational
cosmology

Subsample of about $10^7$ Luminous Red Galaxies (LRG)
Some results

![BC models graph with σ=0.051]

![Bayesian graph with σ=0.0415, Δz=0.0144]

<table>
<thead>
<tr>
<th>type</th>
<th>method</th>
<th>data</th>
<th>$\Delta z_{rms}$</th>
<th>Notes</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEDF</td>
<td>Bruzual-Charlot</td>
<td>EDR</td>
<td>0.0666</td>
<td></td>
<td>(Csabai et al. 2003)</td>
</tr>
<tr>
<td></td>
<td>Interpolated</td>
<td>EDR</td>
<td>0.0552</td>
<td></td>
<td>(Csabai et al. 2003)</td>
</tr>
<tr>
<td></td>
<td>Polynomial</td>
<td>EDR</td>
<td>0.0451</td>
<td></td>
<td>(Csabai et al. 2003)</td>
</tr>
<tr>
<td></td>
<td>KD-tree</td>
<td>EDR</td>
<td>0.0318</td>
<td></td>
<td>(Csabai et al. 2003)</td>
</tr>
<tr>
<td></td>
<td>ANNz</td>
<td>EDR</td>
<td>0.0254</td>
<td></td>
<td>(Csabai et al. 2003)</td>
</tr>
<tr>
<td>ML</td>
<td>SVM</td>
<td>EDR</td>
<td>0.0229</td>
<td></td>
<td>(Collister &amp; Lahav 2004)</td>
</tr>
<tr>
<td>ML</td>
<td>MLP-feed forward</td>
<td>SDSS-DR1</td>
<td>xx.xxx</td>
<td>yes</td>
<td>(Vanzella et al. 2003)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SDSS-RLG</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
• the color space is partitioned (KD-tree - a binary search tree) into cells containing the same number of objects from the training set
• In each cell fit a second order polynomial.
Multi Layer Perceptron

- Input layer (n neurons)
- M hidden layer (1 or 2)
- Output layer (n' < n neurons)

Neurons are connected via activation functions.

Different NN's given by different topologies, different activation functions, etc.
VO-Neural approach

SDSS-DR4/5 - SS

training  validation  Test set

60%, 20%, 20%

MLP, 1(5), 1(18)

0.01<Z<0.25  0.25<Z<0.50

MLP, 1(5), 1(23)  MLP, 1(5), 1(24)

σ rob = 0.196  σ rob = 0.201

99.6 % accuracy
$\sigma = 0.183$

SDSS – DR4/5 - LRG
Uneven coverage of parameter space:

General galaxy sample

$\sigma = 0.0208$
$\Delta z = -0.0029$

LRG sample

$\sigma = 0.0178$
$\Delta z = -0.0011$

Non LRG only

$\sigma = 0.0363$
$\Delta z = -0.0030$
Errors can be easily evaluated

And are, on average, well behaved....
Second example
Searching for candidate quasars in the SDSS archive

astro-ph/0805.0156v1; to appear soon in MNRAS

Quasar candidates selection in the Virtual Observatory era

D’Abrusco¹,², R., Longo¹,³,⁴, G., Walton², N. A.

Unsupervised method (PPS + NEC clustering) with small BOK for labeling
Several algorithms for “general purpose” photometric identification of candidate QSOs select sources according to different techniques exist.

- **Optical surveys**: looking for counterparts of strong radio sources (but only $\sim 10\%$ of QSO are radio-loud).
- **Ultraviolet and optical surveys**: looking for star-like sources bluer than stars.
- **Multi-colour surveys**: looking for star-like objects in colour parameter space lying outside compact regions ("star locus") occupied by stars.

**Overall performances of a generic targeting algorithm are usually expressed by two parameters:**

**Completeness**

$$c = \frac{\text{candidate quasars identified by the algorithm}}{\text{a priori known quasars}}$$

**Efficiency**

$$e = \frac{\text{confirmed quasars identified by the algorithm}}{\text{candidate quasars selected by the algorithm}}$$
An exemplification: finding QSO’s

Traditional way to look for candidate QSO in 3 band survey

Cutoff line

In 4 bands degeneracy is partially removed

More are the bands the lower is the degeneracy

Candidate QSOs for spectroscopic follow-up’s

Ambiguity zone

How to find the interesting regions (clusters)?
• Data Mining is the answer

How to visualize them?
• Dimensionality reduction
SDSS QSO candidate selection algorithm (Richards et al, 2002) targets star-like objects as QSO candidate according to their position in the SDSS colours space \((u-g, g-r, r-i, i-z)\), if one of these requirements is satisfied:

- QSOs are supposed to be placed \(> 4\sigma\) far from a cylindrical region containing the “stellar locus” (S.L.), where \(\sigma\) depends on photometric errors.

OR

- QSOs are supposed to be placed inside the inclusion regions, even if not meeting the previous requirement.

\[ c = 95\%, \; e = 65\% \] locally less
**SDSS QSOs targeting algorithm (II)**

1. **inclusion regions** are regions where S.L. meets QSO’s area (due to absorption from Lyα forest entering the SDSS filters, which changes continuum power spectrum power law spectral index). All objects in these areas are selected so to sample the [2.2, 3.0] redshift range (where QSO density is also declining), but at the cost of a worse efficiency (Richards et al, 2001).

2. **exclusion regions** are those regions outside the main “stellar locus” clearly populated by stars only (usually WDs). All objects in these regions are discarded.

**Overall performance of the algorithm:** completeness $c = 95\%$, efficiency $e = 65\%$, but locally (in colours and redshift) much less.
Our candidate QSO selection algorithm is based on unsupervised clustering inside colours space and exploits mixed (spectroscopic+photometric) datasets. Once clusters have been detected by the chosen algorithm, knowledge-base (spectroscopic types) is used (i.e., “labels” associated to objects within each cluster) to understand the mixture of objects contained in each cluster and to perform statistical analysis of these populations.
**Clustering strategy**

Clustering is usually performed on single objects, but this approach may be too sensitive to single outliers to be extensively used in highly non linear parameter space as astronomical ones. We perform a **pre-clustering** on the real distribution of points inside the parameter space, and then use a **clustering algorithm** to aggregate the pre-clusters produced.

1. **Pre-clustering algorithm:** this phase can be accomplished performing a reduction of dimension of the feature space; this reduction via feature extraction/selection can be supervised or unsupervised (our choice in unsupervised).

2. **Agglomerative clustering:** both distance definition and a linkage model (simple, average, complete, Wards, etc.) need to be provided to perform clustering.
The method:

1. **PPS** determines a large number of distinct groups of objects: nearby clusters in the colours space are mapped onto the surface of a sphere.

2. **NEC** aggregates clusters from PPS to a (a-priori unknown) number of final clusters.

3. These clusters are examined and “interesting” ones are selected through the Base of knowledge.

Two free parameters to be set are the number of latent variables for **PPS** ("resolution" of the initial clustering) and the critical value(s) of dissimilarity threshold $D_{th}$ for **NEC**.

A high number of initial latent bases (i.e. clusters from PPS) is good for almost all applications (empty clusters, if any, can be discarded); critical values for $D_{th}$ are classically determined by two similar methods both embodying a **stability criterion**:

1. **Plateau analysis**: final number of clusters $N(D)$ is calculated over a large interval of $D$, and critical value(s) $D_{th}$ are those for which a plateau is visible.

2. **Dendrogram analysis**: the stability threshold(s) $D_{th}$ can be determined observing the number of branches at different levels of the graph.
Our goal is an objective classifier which can achieve spectroscopic-like classification using only photometric attributes of objects.

Id est, a statistical device aimed at discovering unknown correlation between points (sources) in a photometric only parameter-space using clustering techniques. Our choice was an unsupervised (no a-priori categories) neural network-based combination of algorithms:

PPS (Probabilistic Principal Surfaces) + NEC (Negentropy Clustering) & Kmeans

We need a “knowledge base”: spectroscopic measured features (in our case, spectral classification represented by specClass) are needed and will be used as labels, before applying clustering to the only photometric objects.
Brief sketch of PPS and NEC

**NEC: a matter of Gaussians**

Clustering method based on the “neg-entropy” NegE, a measure of non-gaussianity of a variable. If A is gaussian, then NegE(A) = 0. Given a threshold d:

\[
\text{If } \text{NegE}(A \cup B) < d, \text{ then clusters } A \text{ and } B \text{ are replaced by cluster } A \cup B
\]

**PPS: the Beauty of Spheres**

The original \(m\)-dimensional data space is mapped in a lower \(n\)-dimensional space, called “latent space”. **Visualization ease** as a spherical manifold is fitted to the data, then projected into the manifold in \(R^3\) and plotted as points on the sphere surface.

Each latent variable on the sphere is responsible for a number of projected points, which form a “cluster”.

**NegE=750**

**NegE=4**
Case Study

3D PCA of Yeast Gene Microarray Data
Case Study

Data Projections in latent space
Case Study

Results: clusters (30)
Case Study

Results: pdf + clusters superimposed
Case Study

Gene Prototypes corresponding to 30 computed clusters
Once partition of colors space is completed (as a function of $D_{th}$), clusters mainly populated by QSO (according the knowledge-base at our disposal) are selected and informations about these clusters are exploited for the candidate QSO selection.

To determine the critical dissimilarity $D_{th}$ threshold we rely not only on a stability requirement. Given the following definition:

\[
\text{cluster is “successful”} \quad \text{Def} \quad \text{its fraction of confirmed QSO is higher than a fixed value}
\]

we ask $D_{th}$ to maximise the **Normalised Success Ratio** (NSR):

\[
\text{NSR} = \frac{\text{Number of successful clusters}}{\text{Number of total clusters}}
\]

The process is recursive: feeding merged unsuccessful clusters in the clustering pipeline until no other successful clusters are found. The overall efficiency of the process $e_{\text{tot}}$ is the sum of weighed efficiencies $e_i$ for each generation:

\[
e_{\text{tot}} = \sum_{i=1}^{n} e_i
\]
An example of “tuning”

To assess the reliability of the algorithm, the same objects used for the “training” phase have been re-processed using photometric informations only. Results have been compared to the BoK.

- Efficiency and completeness:
  - QSOs: 759, not QSOs: 72
  - QSOs: 83, not QSOs: 1327
  - e = 83.4 %, c = 89.6 %
Selection of new candidates

Different methods to extract QSOs candidates

- **“Re-labelling”**: both spectroscopic and photometric objects put into the same clustering process: candidate QSOs are selected as those objects belonging to clusters where spectroscopic confirmed QSOs (“tracers”) are found.

- **“Photometric cuts”**: “goal-successful” clusters are described in terms of their colours distribution; associated cuts are applied to photometric sample for candidate selection.

- **“Mahalanobis’ distance”**: it is used to measure the distances of a given photometric object from each cluster; the object is assigned to the nearest “goal-successful cluster” or rejected.
Data and experiments

Data samples:

1. **Optical**: sample derived from SDSS database table “Target” queried for QSO candidates, containing $\sim 1.11 \cdot 10^5$ records and $\sim 5.8 \cdot 10^4$ confirmed QSO (‘specClass == 3 OR specClass == 4’).

2. **Optical + NIR**: sample derived from positional matching (‘best’) between SDSS-DR3 database view “Star” queried for all objects with spectroscopic follow-up available and detection in all 5 bands (u,g,r,i,z) with high reliability for redshift estimation and line-fitting classification (‘specClass’) and high S/N photometry, and UKIDSS-DR1 star-like (‘mergedClass == -1’) objects fully detected in each of the four lasSurvey bands (Y,J,H,K) and clean photometry. **This sample is formed by 2192 objects.**

Experiments:

<table>
<thead>
<tr>
<th>Optical (1)</th>
<th>Optical+NIR (2)</th>
<th>Optical (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>candidate QSO</td>
<td>star-like objects</td>
<td>star-like objects</td>
</tr>
<tr>
<td>4 colours</td>
<td>4 + 3 colours</td>
<td>4 colours</td>
</tr>
</tbody>
</table>
Experiment 2: SDSS \( \cap \) UKIDSS

Only a fraction (43\%) of these objects have been selected as candidate QSO’s by SDSS targeting algorithm in first instance: the remaining sources have been included in the spectroscopic program because they have been selected in other spectroscopic programmes (mainly stars).
Experiment 2: local values of $e$
Experiment 2: local values of $c$
In this experiment the clustering has been performed on the same sample of the previous experiment, using only optical colours.
## Results (I)*

<table>
<thead>
<tr>
<th>Sample</th>
<th>Parameters</th>
<th>Labels</th>
<th>$\epsilon_{\text{tot}}$</th>
<th>$\chi_{\text{tot}}$</th>
<th>$\eta_{\text{gen}}$</th>
<th>$\eta_{\text{succ.clus}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Optical QSO candidates (1)</strong></td>
<td>SDSS colours</td>
<td>‘specClass’</td>
<td>83.4 % (± 0.3 %)</td>
<td>89.6 % (± 0.6 %)</td>
<td>2</td>
<td>(3,0)</td>
</tr>
<tr>
<td><strong>Optical + NIR star-like objects (2)</strong></td>
<td>SDSS colours + UKIDSS colours</td>
<td>‘specClass’</td>
<td>91.3 % (± 0.5 %)</td>
<td>90.8 % (± 0.5 %)</td>
<td>3</td>
<td>(3,1,0)</td>
</tr>
<tr>
<td><strong>Optical + NIR star-like objects (3)</strong></td>
<td>SDSS colours</td>
<td>‘specClass’</td>
<td>92.6 % (± 0.4 %)</td>
<td>91.4 % (± 0.6 %)</td>
<td>3</td>
<td>(3,0,1)</td>
</tr>
</tbody>
</table>
Third example
Classifying AGN in SDSS with SVM

Narrow Line Region

Seyfert 1

Torus

Seyfert 2

Broad Line Region

Central Engine: Accretion Disk + Black Hole
Kauffman et al. (2003) $0.02<z<0.3$

**Kewley line**

$$\log \frac{[\text{OIII}]\lambda 5007}{H_\beta} = \frac{0.61}{\log \frac{[\text{NII}]\lambda 6583}{H_\alpha}} - 0.47 + 1.19$$

G. Kauffman et al. (2003) $0.02<z<0.3$

**Heckman line**

$$\frac{[\text{OIII}]\lambda 5007}{H_\beta} = 2.1445 \frac{[\text{NII}]\lambda 6583}{H_\alpha} + 0.465$$
Support Vector Machines in two slides

given a training set formed by pairs [features-label]: \((x_i, y_i)\), \(i = 1\ldots l\)
where \(x_i \in \mathbb{R}^n\) e \(y_i \in \{1,-1\}\).

Support Vector Machines (SVM) try to solve the following optimization problem:

\[
\min_{\omega, b, \xi} \frac{1}{2} \omega^T \omega + C \sum_{i=1}^{l} \xi_i
\]

With the condition:

\[
y_i (\omega^T \phi(x_i) + b) \geq 1 - \xi_i
\]

Vectors \(x_i\) are mapped into an higher dimensionality space where the SVM identify an hyper plane which maximizes the distances from the two classes

\(C > 0\) is a classification error correction term

\[
K(x_i, x_j) = \phi(x_i)^T (x_j)
\]

Is the so called Kernel function

\[
K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \quad \gamma > 0
\]

radial basis function (RBF)
What is a Good Decision Boundary?

• Consider a two-class, linearly separable classification problem
• Many decision boundaries!
  – The Perceptron algorithm can be used to find such a boundary
• Are all decision boundaries equally good?
Large-margin Decision Boundary

- The decision boundary should be as far away from the data of both classes as possible
  - We should maximize the margin, $m$

\[
\begin{align*}
\mathbf{w}^T \mathbf{x} + b &= 1 \\
\mathbf{w}^T \mathbf{x} + b &= -1 \\
\mathbf{w}^* &= \frac{2}{\|\mathbf{w}\|}
\end{align*}
\]
Transforming the Data

- Computation in the feature space can be costly because it is high dimensional
  - The feature space is typically infinite-dimensional!
- The kernel trick comes to rescue
The optimal values of the two parameters C and gamma cannot be estimated a priori and need to be evaluated on a trial and error procedure.

Usually they are varied as: $C = 2^{-5}, 2^{-3}, \ldots, 2^{15}$, $\Gamma = 2^{-15}, 2^{-13}, \ldots, 2^3$

This process is computationally heavy and it requires GRID (Cloud computing).

Cross-Validation: in order to avoid overfitting effects we use Cross-Validation to estimate the best configuration of the SVM:

The training set is divided into 5 folder: ABCDE, and 5 trainings are performed, with 5 different training set:

ABCD; ABCE; ABDE; ACDE; BCDE

The excluded folder is used for testing the results and the worse result is taken
Experiment 2 with SVM

Efficiency isocontours = $e_{\text{max}} = 79.69\%$

PON-SCOPE GRID Infrastructure (110 nodes)

Training set 30380 objects

- $e = 79.69\%$
- $e$ Seyfert: $e_{\text{sey}} = 74.76\%$
- $e$ LINER: $e_{\text{LIN}} = 81.09\%$
- $c$ Seyfert: $c_{\text{sey}} = 52.77\%$
- $c$ LINER: $c_{\text{LIN}} = 91.69\%$
Some references

• **Tomorrow’s lectures**


• Bishop, C.M.: Neural Networks for Pattern Recognition, Oxford University Press (1995)

• I.T. Nabney, Netlab: Algorithms for Pattern Recognition, Springer-Verlag, 2002


• Chang C.C., Lin C.J., LIBSVM a library for Support Vector Machines

• Staiano 2006 Ph.D. Thesis (VO-Neural website)

• Cauoiti 2008, Thesis (VONeural website)
The company who is making the journey...
(... almost all of them) The VO-Neural team