# Data Mining
# Gallactic Graded Exam Assignment 2014

**Christian Igel**
Department of Computer Science
University of Copenhagen

This is the graded exam assignment for the Data Mining part of the course *Databases and Data Mining* at the University of Copenhagen.

This assignment must be made and submitted individually. However, feel free to discuss the solution of the assignment with your fellow students. Submissions in English are preferred, but submissions in Danish are also accepted.

The assignment will be graded using the 7-point scale. This will be combined with the grades of the previous exams on databases to give the final grade for the course. To obtain the best grade of 12 in this assignment, you must fulfill all the learning objectives at an excellent level. In terms of the questions in this assignment, this means that you have to answer all questions with none or only a few mistakes or parts missing. To obtain the passing grade of 02, you need to fulfill the learning objectives at a minimum level.

### Solution format

The deliverables for each question are listed at the end of each question. The deliverable "description of software used" means that you should hand in the source code you have written to solve the problem. If you have used a tool to solve the problem, this tool should be described and reasons for the particular choice of this tool should be given.

Thus, a solution should contain:

- A report with detailed answers to the questions. Describe the way you solved the problems. Your report should include graphs and tables with comments if needed (**max. 8 page of text** including figures and tables). Use meaningful labels, captions, and legends.

  The way you solved the problems and your results must be comprehensible without looking a the attached source code.

- Your solution code (preferable Python scripts or C++ or Java code) with comments about the major steps involved in each question. The code must be submitted in original format (i.e., not as `.pdf` files). Use meaningful names for files, constants, variables, functions and procedures etc.

  Your code should also include a README text file describing how to compile (if necessary) and run your program. It should also contain a list of all required libraries. If you use the Shark machine learning library, you need not include the library in your submission. If we cannot make your code run we have to consider your submission to be incomplete.

**Database**

All data considered in this assignment are stored in a single SQLite database named `DataMiningAssignment2014.db`.

There are two tables per data set. The tables ending with `_X` contain the input data and the tables ending with `_Y` the corresponding target (output) data. That is, the $n$th row of the table ending with `_Y` contains the label or response given the attributes in the $n$th row of the corresponding table ending with `_Y`.

The file `readData.py` gives an example how to access the data using Python.

The database as well as the source code are available from the course page in the Absalon system.

# 1 Predicting the Specific Star Formation Rate

## 1.1 Background

It is believed that the observable universe contains more than $10^{11}$ galaxies—many like our own Milky Way. Galaxies contain between $10^7$ and $10^{14}$ stars and

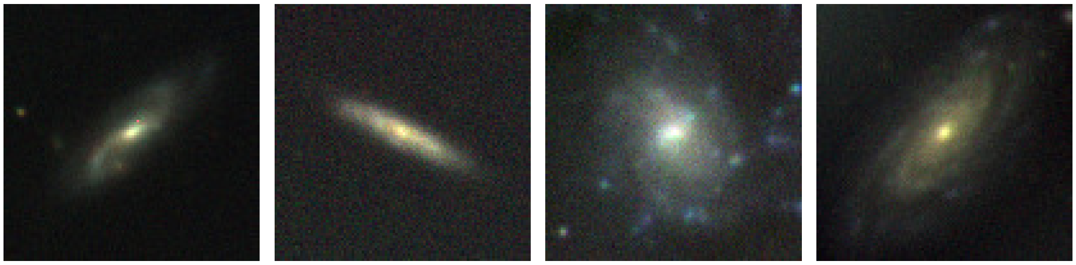

Figure 1: Examples of well-resolved galaxies in the SDSS database.

come in many variations: some have a flat, pancake-like structure, perhaps even containing a spiral pattern, some have an ellipsoidal shape with no clear internal structure or boundary and yet some show only a chaotic structure. The size and appearance of a galaxy is tightly linked to its environment. By determining the properties of galaxies we can therefore learn not only about the galaxies themselves, but also about the universe as a whole. Unfortunately, determining properties of galaxies is difficult, and the best way to do it is by obtaining spectra of their light. Spectra are, however, extremely expensive and time-consuming to acquire, whereas images of galaxies are easily obtained. Indeed, the Sloan Digital Sky Survey (SDSS), one of the most extensive astronomical surveys to this day, has images of more than 200 million galaxies, but only spectroscopic data for about 1.5 million of those. Extracting "spectroscopic information" from images is therefore a major issue in astronomy.

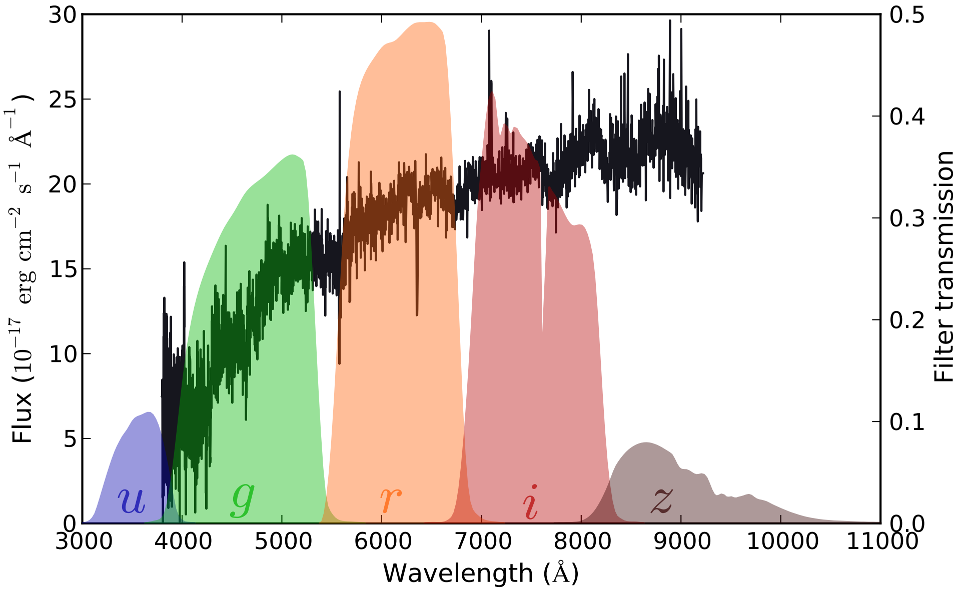

Figure 2: An example spectrum of a galaxy from the SDSS database (black curve) overlaid by the five bandpass filters of SDSS (taken from Stensbo-Smidt et al. [2013]). The five bands are termed $u$, $g$, $r$, $i$ and $z$. They give rise to four colors by subtracting the intensities from neighboring bands, which are the basis of our sSFR prediction.

In this exercise, we will try to do just that. We will use data obtained from images to predict a spectroscopic quantity, namely the specific star formation rate (sSFR). It is basically the number of stars being formed per year in the galaxy, normalised by the galaxy's mass, but in that lies a wealth of information about the formation and evolution of the galaxy.

The input features for our prediction are the so-called colours of the galaxy, which are derived from the intensity of the galaxy's light in different band-pass filters, see Figure 2. The label is the galaxy's sSFR (actually, it is the logarithm of the sSFR).

The training and test data are the tables `SSFR_Train_X` and `SSFR_Train_Y` as well as `SSFR_Test_X` and `SSFR_Test_Y`, respectively.

## 1.2 Mean and sample variance

Let the output data in `SSFR_Train_Y` be given by $y_1, \ldots, y_\ell$. Compute the sample mean

$$\hat{\mu} = \frac{1}{\ell} \sum_{i=1}^{\ell} y_i$$

and the *biased sample variance*

$$s^2 = \frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - \hat{\mu})^2 \ .$$

*Deliverables:* mean and biased sample variance of the number of sunspots in the training data set

## 1.3 Linear regression

The goal of our modeling is to find a mapping $f : \mathbb{R}^4 \to \mathbb{R}$ for predicting the number of sSFR based on the colors.

### 1.3.1 Build model

Build an *affine* linear model of the data using linear regression and the training data in `SSFR_Train_X` and `SSFR_Train_Y` only. Report the five parameters of the model.

### 1.3.2 Training error

Determine the training error by computing the mean-squared-error of the model over the complete *training* data set.

Compare this mean-squared-error with the biased sample variance calculated above. Have a look at the definitions of both quantities and briefly describe what it means if the mean-squared-error is below or above the biased sample variance.

### 1.3.3 Test error

Compute the mean-squared-error on the test data set (i.e., use `SSFR_Test_X` and `SSFR_Test_Y`). Comment very briefly on the result.

*Deliverables:* description of software used; parameters of the regression model; mean-squared error on the training and test data set; brief discussion relating mean-squared-error to the biased sample variance; short discussion of results on the test set

# 2 Stars vs. Galaxies

A crucial question in astronomy is: What kind of object are we actually observing? The further we look into space with our telescopes, the lower the resolution and the more difficult it is to disambiguate between a point source, e.g., a star, and an extended object like a galaxy.

Figure 3: Mosaic of the sky as observed by the SDSS telescope.

The data we are using consists of a random 6000 sample-subset of the SDSS (http://www.sdss3.org/dr10/) for astronomical objects whose properties have been confirmed by spectral follow-up observations. The features describing each sample consist of 2 magnitudes, which are observed in 5 different bands ($u$ = ultraviolet, $g$ = green, $r$ = red, $i$ = near infrared, $z$ = infrared, see Figure 2). The composite model magnitude (*cModelMag*) belongs to a galaxy model that is fitted to the observed telescope image. The point spread function model magnitude (*psfModelMag*) likewise expresses the magnitude of a fitted point source model. In total, we consider 10 features, which are listed in Table 2. The label for one sample can either be 0 for a galaxy or 1 for a star. Both training and test data set contain 3000 samples.

Consider the tables `Objects_Train_X` and `Objects_Train_Y` containing training data and the tables `Objects_Test_X` and `Objects_Test_Y` containing test data.

| # | Feature name |
|---|---|
| 0 | cModelMag_u |
| 1 | cModelMag_g |
| 2 | cModelMag_r |
| 3 | cModelMag_i |
| 4 | cModelMag_z |
| 5 | psfModelMag_u |
| 6 | psfModelMag_g |
| 7 | psfModelMag_r |
| 8 | psfModelMag_i |
| 9 | psfModelMag_z |

Table 1: Features describing the astronomical object in the binary classification task.

## 2.1 Classification

Train a nearest neighbor classifier (1-NN) using the training data stored in the tables `Objects_Train_X` and `Objects_Train_Y`. Measure its performance on the test data stored in `Objects_Test_X` and `Objects_Test_Y`. How high is the classification accuracy on the test data?

*Deliverables:* description of software used; test accuracies of nearest neighbor classifier

## 2.2 Dimensionality reduction and visualization

In this exercise, we look more closely at the galaxies in the training data set. Perform a principal component analysis (PCA) of the input attributes of the 1849 training data patterns in `Objects_Train_X` *corresponding to galaxies* (i.e., have label 0).

Plot the eigenspectrum. How many components are necessary to "explain 90 % of the variance"? Visualize the data by a scatter plot of the first two principal components using a different color for each class. Briefly discuss the results.

*Deliverables:* description of software used; plot of the eigenspectrum; number of components necessary to explain 90 % of variance; scatter plot of the galaxy training data projected on first two principal components of the galaxy training data; brief discussion of results

## 2.3 Clustering

Perform 2-means clustering of the 1849 training input patterns corresponding to galaxies in `SGTrain2014.dt` and report the 10-dimensional cluster centers. *After that*, project the cluster centers to the first two principal components of the training data. Then visualize the clusters by adding the cluster centers to the plot from the previous exercise 2.2. Briefly discuss the results: Did you get meaningful clusters?

*Deliverables:* description of software used; cluster centers; one plot with cluster centers and data points; short discussion of results

## Acknowledgment

Funding for the SDSS and SDSS-II has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Japanese Monbukagakusho, the Max Planck Society, and the Higher Education Funding Council for England. The SDSS Web Site is `http://www.sdss.org/`.

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are the American Museum of Natural History, Astrophysical Institute Potsdam, University of Basel, University of Cambridge, Case Western Reserve University, University of Chicago, Drexel University, Fermilab, the Institute for Advanced Study, the Japan Participation Group, Johns Hopkins University, the Joint Institute for Nuclear Astrophysics, the Kavli Institute for Particle Astrophysics and Cosmology, the Korean Scientist Group, the Chinese Academy of Sciences (LAMOST), Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy (MPIA), the Max-Planck-Institute for Astrophysics (MPA), New Mexico State University, Ohio State University, University of Pittsburgh, University of Portsmouth, Princeton University, the United States Naval Observatory, and the University of Washington.

# References

M. Charytanowicz, J. Niewczas, P. Kulczycki, P. A. Kowalski, S. Łukasik, and S. Żak. Complete gradient clustering algorithm for features analysis of x-ray images. In E. Piętka and J. Kawa, editors, *Information Technologies in Biomedicine*, volume 69 of *Advances in Intelligent and Soft Computing*, pages 15–24. Springer, 2010.

K. Stensbo-Smidt, C. Igel, A. Zirm, and K. Steenstrup Pedersen. Nearest neighbour regression outperforms model-based prediction of specific star formation rate. In *IEEE International Conference on Big Data*, pages 141–144. IEEE Press, 2013.