Data Mining
Graded Exam Assignment 2015

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 pages 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 at 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 `DataMiningAssignment2015.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 nth row of the table ending with `_Y` contains the label or response given the attributes in the nth row of the corresponding table ending with `_X`.

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

1 Photometric Redshift Estimation

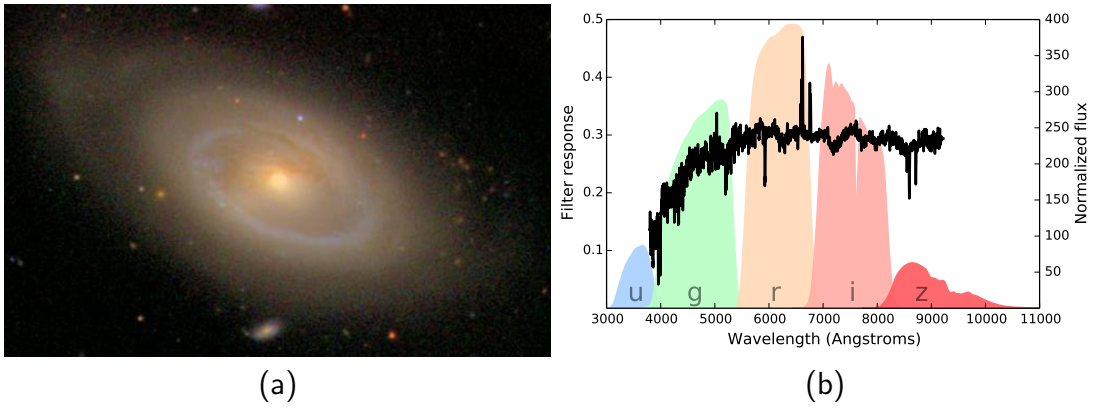
1.1 Background

Astronomy is rich with data. The advent of wide-area digital cameras on large aperture telescopes has led to ever more ambitious surveys of the sky. The data volume of an entire survey of a decade ago can now be acquired in a single night. Automatic data analysis methods are required to fully exploit this wealth of data.

Here we consider photometric redshift estimation of galaxies. The redshift phenomenon is caused by the Doppler effect, which shifts the spectrum of an object towards longer wavelengths if it is moving away from the observer. Because the universe is expanding uniformly, we can infer a galaxy's velocity by its redshift and, thus, its distance to Earth. Hence, redshift estimation is a useful tool for determining the geometry of the universe. A photometric observation contains the intensities of an object (in our case, galaxies) in 5 different bands (u,g,r,i,z), ranging from ultraviolet to infrared, see Figure 1. Spectroscopy, in contrast, measures the photon count at certain wavelengths. The resulting spectrum allows for identifying the chemical components of the observed object and thus, enables

determining many interesting properties, including the redshift. Spectroscopy, however, is much more time-consuming than photometric observation and therefore, costs could be greatly reduced if we could predict suitable candidates for follow-up spectroscopy from low-quality low-cost photometry.

Your task is to train and evaluate a photometric redshift estimator using astronomical data [Sheldon et al., 2012]. For each of the 5 bands a point spread function (*model*) and a composite model (*cmodel*) are fit to the photometric observation. We take the 4 magnitude differences between adjacent bands and the magnitude in the red band for *model* and *cmodel*. Thus, we arrive at $2 \times (4+1) = 10$ variables for each galaxy. The target variable *redshift* is the ground-truth as determined by spectroscopic observation.



(a) (b)

Figure 1: An example from the *Sloan Digital Sky Survey* (SDSS) [Aihara et al., 2011]. (a) An image of the spiral galaxy NGC 5750. (b) Its associated spectrum overlapping the five photometric intensity band filters *u,g,r,i,z*.

The dataset contains a sample of 5000 galaxies whose redshift has been determined by spectroscopy. It is split in training and test set, each consisting of 2500 galaxies, which are described by the features listed in Table 1.

1.2 Mean and sample variance

Let the output data in `Redshift_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*

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

Feature #	Feature name
0	model magnitude difference between bands u-g
1	model magnitude difference between bands g-r
2	model magnitude difference between bands r-i
3	model magnitude difference between bands i-z
4	model magnitude in band r
5	composite model magnitude difference between bands u-g
6	composite model magnitude difference between bands g-r
7	composite model magnitude difference between bands r-i
8	composite model magnitude difference between bands i-z
9	composite model magnitude in band r
target	redshift (determined by spectroscopy)

Table 1: Relevant features for photometric redshift estimation.

Deliverables: mean and biased sample variance of the redshift in the training data set

1.3 Linear regression

The goal of our modeling is to find a mapping $f : \mathbb{R}^{10} \to \mathbb{R}$ for predicting the redshift.

1.3.1 Build model

Build an *affine* linear model of the data using linear regression and the training data in `Redshift_Train_X` and `Redshift_Train_Y` only. Report the 11 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 `Redshift_Test_X` and `Redshift_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 Cyberfraud

According to the PWC "2011 Global Economics Crime Survey", cybercrime ranks among the top four economic crimes, not only leading to financial loss, but also bearing the risk of reputational damage for financial institutions and other online service providers.

In this part of the assignment, you are supposed to develop a system for detecting fraudulent logins. The idea is to identify a criminal who has stolen a password by the way s/he types the password. That is, the goal is to identify the user that produced a given keyboard input based on the keystroke timings. We consider data from the dataset from Killourhy and Maxion [2009], which consists of 400 keystroke samples from various persons all typing the same strong password "`.tie5Roanl`". The keystrokes are collected into a vector of flight and dwell times that contains 21 features [Moskovitch et al., 2009]. In the assignment, we consider only two users.

Consider the tables `Keystrokes_Train_X` and `Keystrokes_Train_Y` containing training data and the tables `Keystrokes_Test_X` and `Keystrokes_Test_Y` containing test data.

2.1 Classification

Train a nearest neighbor classifier (1-NN) using the training data stored in the tables `Keystrokes_Train_X` and `Keystrokes_Train_Y` (use the standard Euclidean distance as the underlying metric). Measure its performance on the test data stored in `Keystrokes_Test_X` and `Keystrokes_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 typing patterns in the training data set. Perform a principal component analysis (PCA) of the input attributes of the training data patterns in `Keystrokes_Train_X`.

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. 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 training data projected on first two principal components of the training data; brief discussion of results

2.3 Clustering

Perform 2-means clustering of the training input patterns and report the 21-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

H. Aihara, C. A. Prieto, D. An, S. F. Anderson, É. Aubourg, E. Balbinot, T. C. Beers, A. A. Berlind, S. J. Bickerton, D. Bizyaev, et al. The eighth data release of the Sloan digital sky survey: first data from SDSS-III. *The Astrophysical Journal Supplement Series*, 193(2):29, 2011.

K. S. Killourhy and R. A. Maxion. Comparing anomaly-detection algorithms for keystroke dynamics. In *Dependable Systems & Networks, 2009. DSN'09. IEEE/IFIP International Conference on*, pages 125–134. IEEE, 2009.

R. Moskovitch, C. Feher, A. Messerman, N. Kirschnick, T. Mustafic, A. Camtepe, B. Lohlein, U. Heister, S. Moller, L. Rokach, et al. Identity theft, computers and behavioral biometrics. In *IEEE International Conference on Intelligence and Security Informatics (ISI'09)*, pages 155–160. IEEE, 2009.

E. S. Sheldon, C. E. Cunha, R. Mandelbaum, J. Brinkmann, and B. A. Weaver. Photometric redshift probability distributions for galaxies in the SDSS DR8. *The Astrophysical Journal Supplement Series*, 201(2):32, 2012.