Statistical Methods for Machine Learning

In a Galaxy Far, Far Away

Exam Assignment

Christian Igel and Aasa Feragen
Department of Computer Science
University of Copenhagen

This is the final exam assignment on the course *Statistical Methods for Machine Learning*, block 3 2014, at the University of Copenhagen. It is based on the full course curriculum as stated in the lectures schedule in the Absalon system. The assignment is centered around real world pattern recognition tasks.

This assignment must be made and submitted *individually*. It is also acceptable (and we encourage) to use bits and pieces of your solutions and the hand-out code from the previous assignments.

Your solution to this assignment will be graded using the 7-point scale and will be the final grade for the course. To obtain the best grade of 12, you must fulfill all the course learning objectives (see below) at an excellent level. In terms of the questions in this assignment, this means that you have to answer all questions with non 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, which means you have to make a serious attempt at solving the central questions in the assignment (but not necessarily all) with some mistakes allowed.

The deadline for this assignment is **April 3, 2013**. You must submit your solution electronically via the Absalon home page. Go to the assignments list and choose this assignment and upload your solution prior to the deadline.

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 software library to solve the problem, this library should be described and reasons for the particular choice should be given.

Thus, a solution should contain:

- A PDF document showing your results and giving detailed answers to the questions. If relevant, this may include graphs and tables with comments (**max. 10 page of text including figures and tables**). Use meaningful labels, captions, and legends. Do **not** include your source code in this PDF file. You will be graded mainly on the basis of this report.
- Your solution code (Matlab / R / Python scripts or C / C++ ? Java code) with comments about the major steps involved in each question. The code must be submitted in its original format (e.g., in `.m` or `.R` file format – not as PDF files). Use meaningful names for files, constants, variables, functions and procedures etc. Add comments to the code to make it more readable. Your code should be structured such that there is one main file that we can run to reproduce all the results presented in your report. This main file can, if you like, call other files with functions, classes, etc.
- Your code should also include a README text file describing how to compile (if relevant) and run your program, as well as a list of all relevant libraries needed for compiling or using your code. If we cannot make your code run we will consider your submission incomplete.

Learning objectives

The grade will be based on a judgement of how well you fulfill the following learning objectives:

1. Recognize and describe possible applications of machine learning for pattern recognition and data mining.
2. Explain, contrast and apply basic Bayesian probability theory for modeling stochastic data, including both parametric and non-parametric representations.
3. Explain and contrast the concept of supervised and unsupervised learning.
4. Explain the concepts of classification and clustering.
5. Identify, explain and handle the common pitfalls of machine learning.
6. Describe and apply linear techniques for classification.
7. Implement selected machine learning techniques.
8. Use software libraries for solving machine learning problems.

9. Visualize and evaluate results obtained with a machine learning method.

10. Compare, appraise and select methods of machine learning for solving specific problems of pattern recognition and data mining.

1 Predicting the Specific Star Formation Rate

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 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.

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).

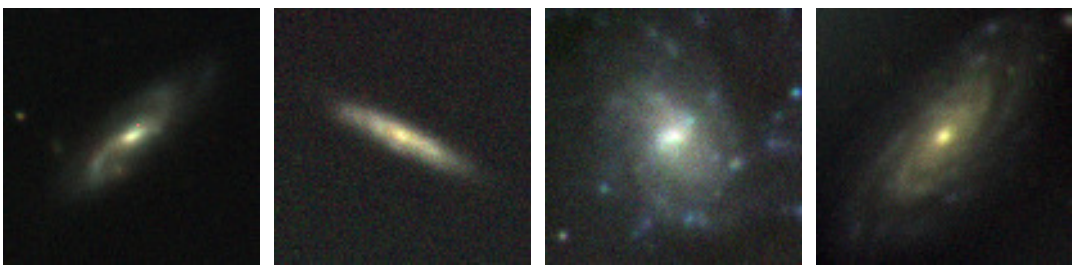

Figure 1: Examples of well-resolved galaxies in the SDSS database.

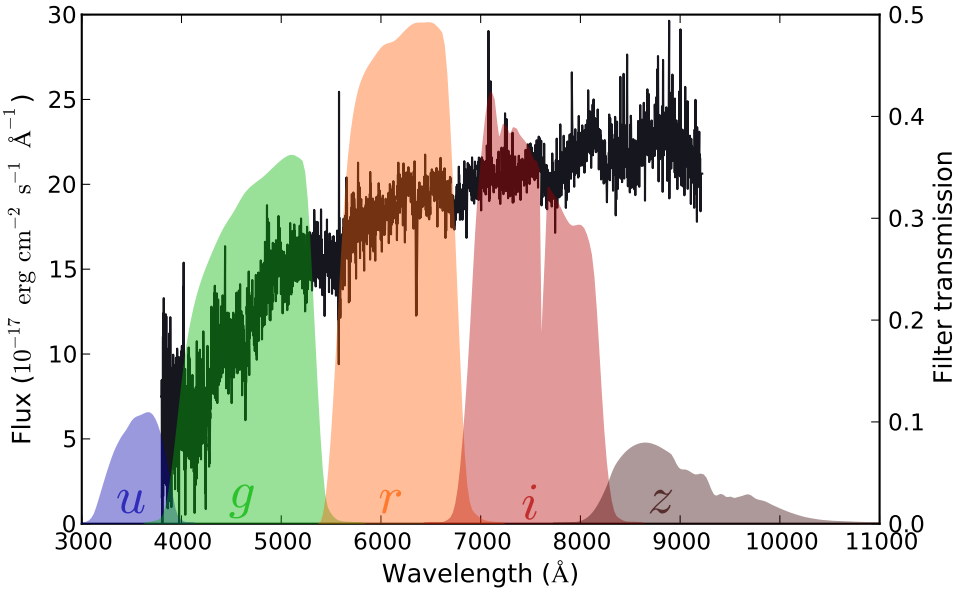

Figure 2: An example spectrum of a galaxy from the SDSS database (black curve) overlaid by the five bandpass filters of SDSS (taken from Stensbo-Smidt et al. [2013]). The five bands are termed u, g, r, i and z. They give rise to four colors by subtracting the intensities from neighboring bands, which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

The training and test data are the files `SSFRTrain2014.dt` and `SSFRTest2014.dt`, respectively. Each row correspond to the four input features (predictor variables) and, in the last column, the desired target (response variable).

Question 1 (linear regression)**.** The goal of our modeling is to find a mapping $f : \mathbb{R}^4 \to \mathbb{R}$ for predicting the logarithmic sSFR based on the colors.

Build an affine linear model of the data using linear regression and the training data in `SSFRTrain2014.dt` only. Report the parameters of the model (do not forget the offset/bias parameter).

Determine the training error by computing the mean-squared-error of the model over the complete *training* data set. Compute the mean-squared-error on the test data set `SSFRTest2014.dt`. 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; short discussion of results

Question 2 (non-linear regression)**.** Now fit the data to a non-linear regression function using one of the regression methods covered in the lectures. Describe the measures you have taken to ensure good generalization. Use the same split in training and test data, and use the mean-squared error to evaluate your model.

Deliverables: Description of software used; mean-squared error on the training and test data set; describe what you did to achieve good generalization performance; discussion of results

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.

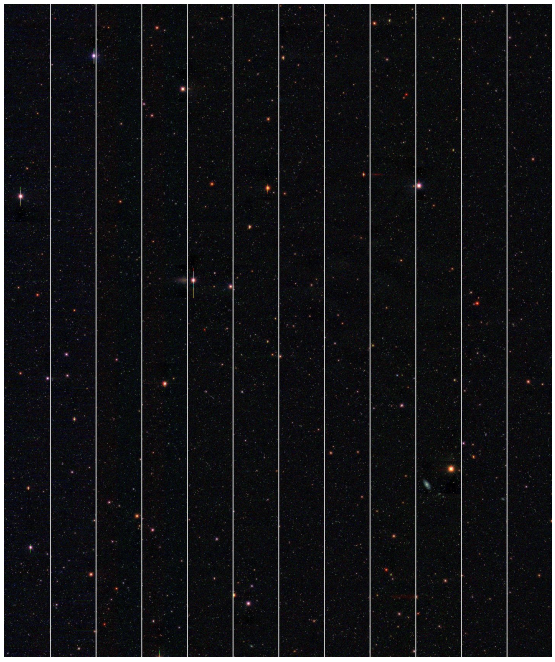

Figure 3: Mosaic of the sky as observed by the SDSS telescope.

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 training data is in the file `SGTrain2014.dt` and the test data in `SGTest2014.dt`.

The features that describe 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.

Question 3 (binary classification using support vector machines)**.** The task is to perform binary classification using support vector machines (SVMs). For this exercise, use standard C-SVMs as introduced in the lecture. Employ radial Gaussian kernels of the form

$$k(\boldsymbol{x}, \boldsymbol{z}) = \exp(-\gamma \|\boldsymbol{x} - \boldsymbol{z}\|^2) \ .$$

Here $\gamma > 0$ is a bandwidth parameter that has to be chosen in the model selection process. Note that instead of γ often the parameter $\sigma = \sqrt{1/(2\gamma)}$ is considered.

#	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.

Jaakkola's heuristic provides a reasonable initial guess for the bandwidth parameter σ or γ of a Gaussian kernel [Jaakkola et al., 1999]. To estimate a good value for σ, consider for every training example $\boldsymbol{x}_i$ the distance to the closest training example $\boldsymbol{x}_j$ having a different label (i.e., $y_i \neq y_j$). The median of these distances can be used as a measure of scale and therefore as a guess for σ. More formally, compute

$$G = \left\{ \min_{(\boldsymbol{x}_j, y_j) \in S \wedge y_i \neq y_j} \{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|\} \,|\, (\boldsymbol{x}_i, y_i) \in S \right\}$$

based on your training data S. Then set σ_{Jaakkola} equal to the median of the values in G:

$$\sigma_{\text{Jaakkola}} = \text{median}(G)$$

Compute the bandwidth parameter γ_{Jaakkola} from σ_{Jaakkola} using the identity given above.

Use grid-search to determine appropriate SVM hyperparameters γ and C. Look at all combinations of

$$C \in \{b^{-2}, b^{-1}, 1, b, b^2, b^3\}$$

and

$$\gamma \in \left\{ \gamma_{\text{Jaakkola}} \cdot b^i \,|\, i \in \{-3, -2, -1, 0, 1, 2, 3\} \right\} \ ,$$

where the base b can be chosen to be either 2, the base e of the natural logarithm (Euler's number), or 10. Feel free to vary this grid. For each pair, estimate the performance of the SVM using 5-fold cross validation (see section 1.3 in the textbook by Bishop [2006]). Pick the hyperparameter pair with the lowest average 0-1 loss (classification error) and use it for training an SVM with the complete training dataset. Only use the data from `SGTrain2014.dt` in the model selection and training process.

Report the values for C and γ you found in the models selection process. Compute the classification accuracy based on the 0-1 loss on the training data as well as on the test data.

Deliverables: Description of software used; a short description of how you proceeded; initial γ or σ value suggested by Jaakkola's heuristic; optimal C and γ found by grid search; classification accuracy on training and test data

Question 4 (principal component analysis)**.** In this exercise, we look more closely at the galaxies in `SGTrain2014.dt`. Perform a principal component analysis of the 1849 training input patterns corresponding to galaxies. Plot the eigenspectrum (see Figure 12.4 by Bishop [2006] for an example). Visualize the data by a scatter plot of the data projected on the first two principal components.

Deliverables: Description of software used; plot of the eigenspectrum; scatter plot of the data projected on the first two principal components

Question 5 (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 4. 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

Question 6 (kernel mean classifier)**.** Given a training set with observed input variables $\mathcal{X} = \bigcup_{c\in\mathcal{C}} \mathcal{X}_c$ split into subsets $\mathcal{X}_c$ of observations belonging to classes $\mathcal{C} = \{c_1, \ldots, c_q\}$, the *nearest mean classifier* is defined as

$$h(x) = \operatorname{argmin}_{c\in\mathcal{C}} \|x - \mu(\mathcal{X}_c)\| \ ,$$

where $\mu(\mathcal{X}_c)$ denotes the mean of all training data points belonging to the class c.

For a positive definite kernel $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, define a nearest mean classifier in the feature space which only depends on the value of the kernel evaluated on pairs of data points.

That is, both the mean vector for each class as well as the distance of a training sample from the mean vectors have to be computed in the kernel-induced feature space.

Deliverables: Formula and proof

Notation: Above, the input data $x_1, x_2, \ldots x_l$ come from some normed vector space $\mathcal{X}$ (for instance, the usual $\mathbb{R}^D$), and the training set consists of input data and class labels $S = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ where the labels y_n belong to the

classes $\mathcal{C} = \{c_1, c_2, \ldots, c_q\}$. The split of the input data into class memberships is defined as $\mathcal{X}_c = \{x | (x, y) \in S \wedge y = c\}$ for each $c \in \mathcal{C}$, giving $\mathcal{X} = \bigcup_{c \in \mathcal{C}} \mathcal{X}_c$.

3 Variable Stars

A variable star changes its intensity, as observed by a telescope, over time. This can be caused extrinsically, for example by other objects temporarily occluding it, but also intrinsically, when the star changes its physical properties over time. Figure 4 shows an example. The graph of the varying intensity as a function of time is called the light curve. Variable stars can be further divided into many classes depending on other physical properties. The task we are trying to solve is to predict the class of a variable star by its light curve. To achieve this, we train a classifier in a supervised setting using labeled data from the All Sky Automated Survey Catalog of Variable Stars (ACVS) [Pojmanski, 2000].

The data considered in the following is based on the study by Richards et al. [2012]. We have a training and a test set, in the file `VSTrain2014.dt` and `VSTest2014.dt`, respectively, with 771 labeled samples each. Each sample encodes the astronomical properties of a variable star in a 61-dimensional feature vector. The features are listed in Table 3, for a detailed description of their meaning we refer to Dubath et al. [2011] and Richards et al. [2011]. The labels indicate the class a variable star has been assigned to. In total there are 25 different classes, see Table 3.

Question 7 (multi-class classification)**.** Use a linear and a non-linear classification method (picking from the methods presented in the course) for classifying the classes, for example k-nearest neighbor and linear discriminant analysis (LDA). Only use the training data in `VSTrain2014.dt` in the model building process. After you trained a model, use the test data in `VSTest2014.dt` to evaluate it. Report the classification error on both training and test set.

Deliverables: Description of software used; arguments for your choice of classification methods; a short description of how you proceeded and what training and test results you achieved

Question 8 (overfitting)**.** John Langford, who is "Doctor of Learning at Microsoft Research", maintains a very interesting blog (web log). Read the very true blog entry: "Clever methods of overfitting," `http://hunch.net/?p=22`, 2005. Choose three of the different types of overfitting and discuss if and how they can occur when applying machine learning techniques to the variable star classification task. Ignore the last type of overfitting and issues related to reviewing of scientific papers (still, it is good to keep them in mind).

Deliverables: Short discussion addressing three "methods of overfitting" listed in the blog entry

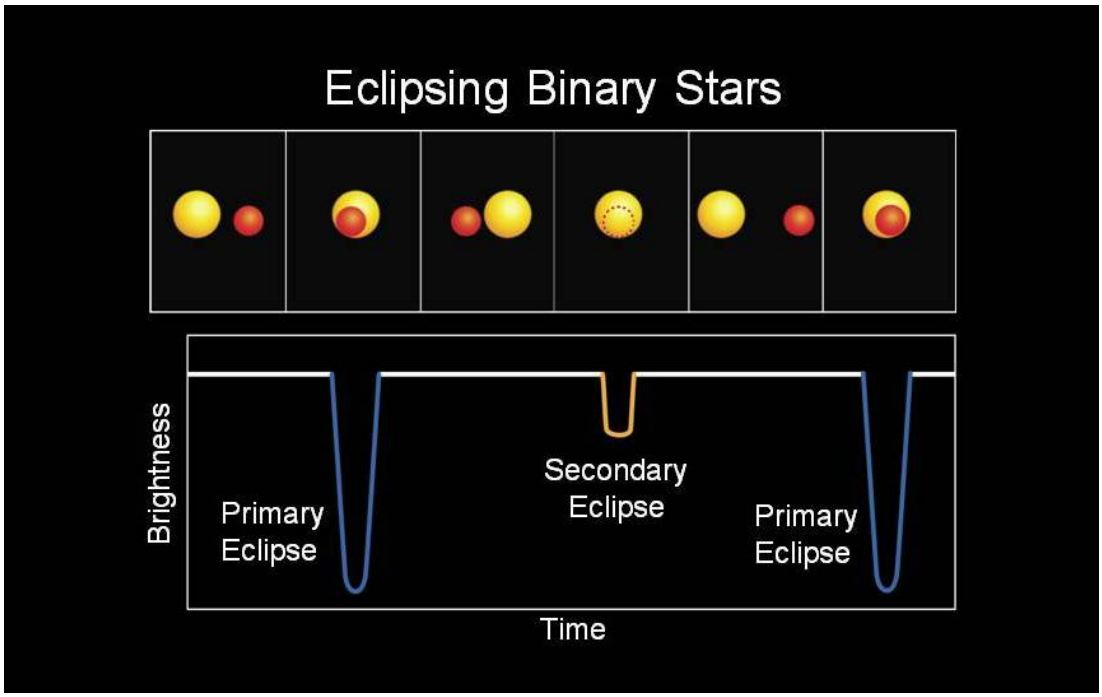

Figure 4: A variable star changes its intensity as observed from a telescope due to another smaller orbiting star. The image is taken from the NASA, `http://kepler.nasa.gov/news/nasakeplernews/index.cfm?FuseAction=ShowNews&NewsID=152`.

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 Uni-

#	Feature name	#	Feature name
0	amplitude	31	freq3_harmonics_rel_phase_2
1	beyond1std	32	freq3_harmonics_rel_phase_3
2	flux_percentile_ratio_mid20	33	freq_amplitude_ratio_21
3	flux_percentile_ratio_mid35	34	freq_amplitude_ratio_31
4	flux_percentile_ratio_mid50	35	freq_frequency_ratio_21
5	flux_percentile_ratio_mid65	36	freq_frequency_ratio_31
6	flux_percentile_ratio_mid80	37	freq_signif
7	fold2P_slope_10percentile	38	freq_signif_ratio_21
8	fold2P_slope_90percentile	39	freq_signif_ratio_31
9	freq1_harmonics_amplitude_0	40	freq_varrat
10	freq1_harmonics_amplitude_1	41	freq_y_offset
11	freq1_harmonics_amplitude_2	42	linear_trend
12	freq1_harmonics_amplitude_3	43	max_slope
13	freq1_harmonics_freq_0	44	median_absolute_deviation
14	freq1_harmonics_rel_phase_1	45	median_buffer_range_percentage
15	freq1_harmonics_rel_phase_2	46	medperc90_2p_p
16	freq1_harmonics_rel_phase_3	47	p2p_scatter_2praw
17	freq2_harmonics_amplitude_0	48	p2p_scatter_over_mad
18	freq2_harmonics_amplitude_1	49	p2p_scatter_pfold_over_mad
19	freq2_harmonics_amplitude_2	50	p2p_ssqr_diff_over_var
20	freq2_harmonics_amplitude_3	51	percent_amplitude
21	freq2_harmonics_freq_0	52	percent_difference_flux_percentile
22	freq2_harmonics_rel_phase_1	53	QSO
23	freq2_harmonics_rel_phase_2	54	non_QSO
24	freq2_harmonics_rel_phase_3	55	scatter_res_raw
25	freq3_harmonics_amplitude_0	56	skew
26	freq3_harmonics_amplitude_1	57	small_kurtosis
27	freq3_harmonics_amplitude_2	58	std
28	freq3_harmonics_amplitude_3	59	stetson_j
29	freq3_harmonics_freq_0	60	stetson_k
30	freq3_harmonics_rel_phase_1		

Table 2: Different features are used to describe the light curve of a variable star.

versity of Washington.

References

C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 1 edition, 2006.

P. Dubath, L. Rimoldini, M. Süveges, J. Blomme, M. López, L. Sarro, J. De Ridder,

Label	Class name	Label	Class name
0	Mira	13	Gamma Doradus
1	Semireg PV	14	Pulsating Be
2	RV Tauri	15	Per. Var. SG
3	Classical Cep	16	Chem. Peculia
4	Pop. II Cephe	17	Wolf-Rayet
5	Multi. Mode C	18	T Tauri
6	RR Lyrae, FM	19	Herbig AE/BE
7	RR Lyrae, FO	20	S Doradus
8	RR Lyrae, DM	21	Ellipsoidal
9	Delta Scuti	22	Beta Persei
10	Lambda Bootis	23	Beta Lyrae
11	Beta Cephei	24	W Ursae Maj
12	Slowly Puls.		

Table 3: The 25 different classes a variable star can be assigned to.

J. Cuypers, L. Guy, I. Lecoeur, et al. Random forest automated supervised classification of hipparcos periodic variable stars. *Monthly Notices of the Royal Astronomical Society*, 414(3):2602–2617, 2011.

T. Jaakkola, M. Diekhaus, and D. Haussler. Using the Fisher kernel method to detect remote protein homologies. In T. Lengauer, R. Schneider, P. Bork, D. Brutlad, J. Glasgow, H.-W. Mewes, and R. Zimmer, editors, *Proceedings of the Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 149–158. AAAI Press, 1999.

G. Pojmanski. The all sky automated survey. catalog of about 3800 variable stars. *Acta Astronomica*, 50:177–190, 2000.

J. W. Richards, D. L. Starr, N. R. Butler, J. S. Bloom, J. M. Brewer, A. Crellin-Quick, J. Higgins, R. Kennedy, and M. Rischard. On machine-learned classification of variable stars with sparse and noisy time-series data. *The Astrophysical Journal*, 733(1):10, 2011.

J. W. Richards, D. L. Starr, H. Brink, A. A. Miller, J. S. Bloom, N. R. Butler, J. B. James, J. P. Long, and J. Rice. Active learning to overcome sample selection bias: Application to photometric variable star classification. *The Astrophysical Journal*, 744(2):192, 2012.

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.