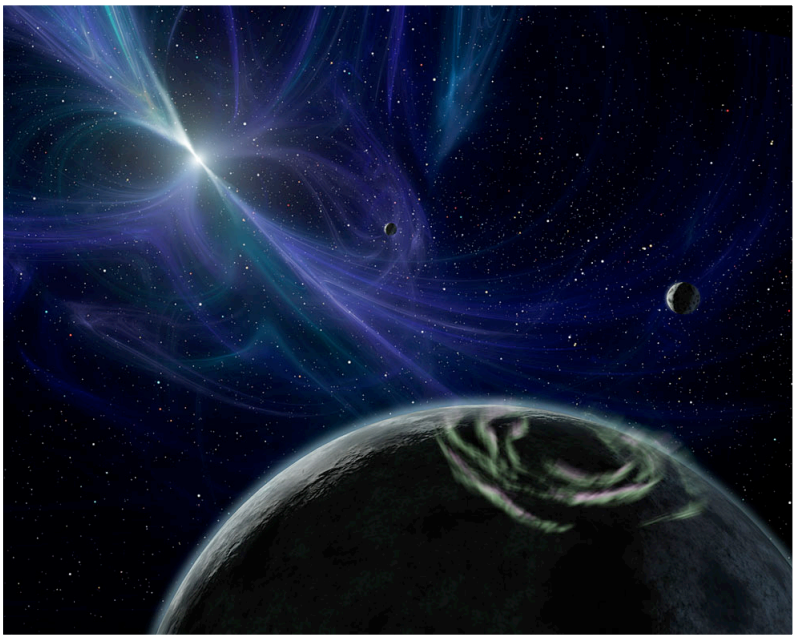
ENABLING THE SEARCH FOR EXOPLANETS
FINAL REPORT



MICHAEL D'ANDREA

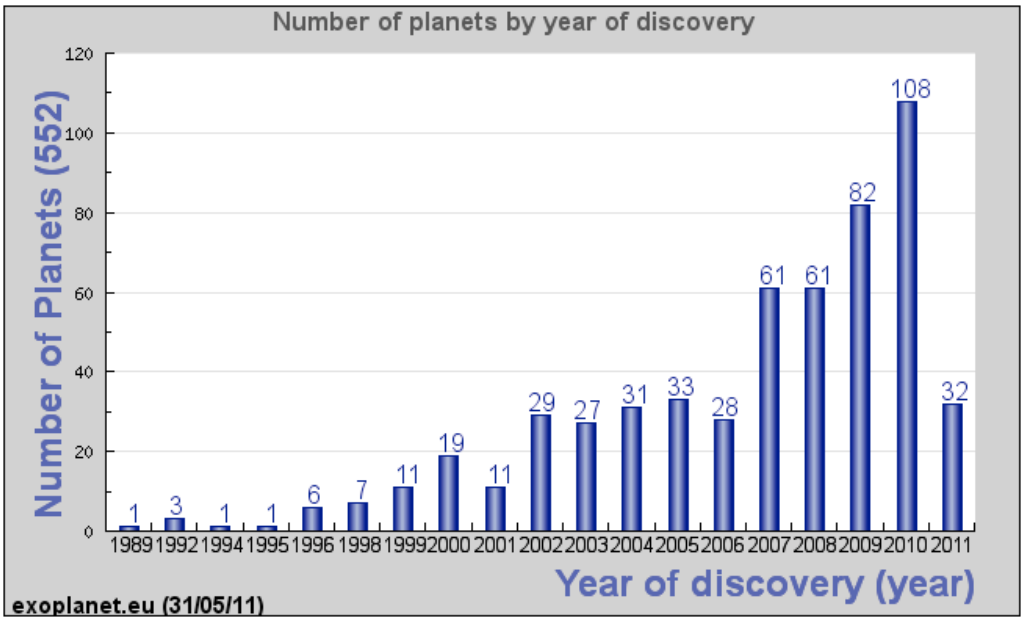
COMPUTER SCIENCE 34:
MACHINE LEARNING AND STATISTICAL ANALYSIS

DARTMOUTH COLLEGE ~ SPRING 2011

INTRODUCTION

In current astronomy, one of the most significant advances in data collection is the discovery of exoplanets, or planets outside of our solar system. The first ever *confirmed* exoplanet sighting, of PSR 1257b, was in 1992, and since then, especially in recent years, a boom in discovery has taken place, thanks to advances in detection methods and technology. In particular, there are (as of 30/5/2011) 552 exoplanet candidates, discovered by any means of detection (http://exoplanet.eu/catalog.php).

FIGURE 1: PLANETARY CANDIDATES DISCOVERED BY YEAR



The rapidly increasing database of known exoplanets, while fortunate and informative, is still very much a random sampling of "rechecking" familiar stars that were found at time before exoplanet discovery was refined. This is because the physics of exoplanet formation and retention in a star system is still poorly understood. Thus, I determined to investigate how certain stellar properties affected the star's likelihood of hosting an exoplanet, to help better inform where astronomers should 'look' to find new instances of exoplanets.

METHODOLOGY

The two major components of determining how to conduct this project were to choose an appropriate database of stellar information and to select an algorithm.

DATA

The data I used for this project came from the Henry Draper (HD) catalogue; I collected the data from the VizieR astronomical catalogue, provided by the Centre de données astromoniques de Strasbourg. There are 125 exoplanet candidates within this data set. In order to select the sample, of host and non-host stars in sum, I first needed to determine which stellar properties I was interested in investigating.

With some research in the field and some guidance from Professor Brian Chaboyer of the Dartmouth Physics and Astronomy Department, I decided upon the following seven stellar properties, in hopes that these would maximize the relevance to exoplanet hosting conditions, and to minimize the number of parameters as to avoid the curse of dimensionality. The seven continuously valued parameters are:

LOG TE ([K]) - the base-10 log of the effective temperature of the star

[FE/H] ([Sun]) - the metallicity of the star

DIST (pc) - distance to the star from Earth

AGE (Gyr) - age of star

VMAG - the Johnson V magnitude (scaled light frequency)

MASS (solMass) - mass of star

VSINI (km/s) - rotational velocity

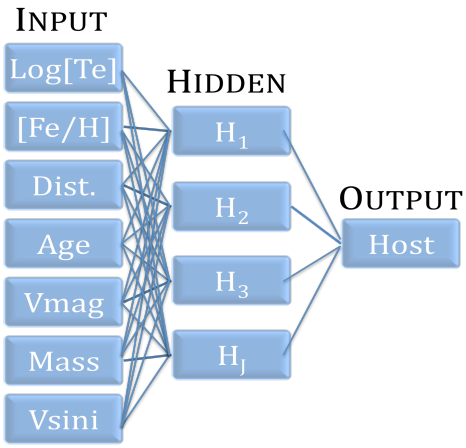
By narrowing down the HD catalogue to stars with complete information on these seven parameters, and to those discovered by means of the radial velocity technique or astrometry, I was able to collect data on 10564 samples. To supplement this data, I used information from the Exoplanet Data Explorer to classify the 125 exoplanet-hosting stars within the dataset.

An important note on data methodology: given the imprecision of the field, as well as the newness of the data, many of the identified exoplanets used for this project are just *candidates* based on the relevant astronomical observations and measurements, and are not yet *confirmed*. Furthermore, there will be inevitably stars classified as non-hosts that do in fact host exoplanets, but have not been yet discovered as such. The hope is that these flaws in methodology are not great enough to disrupt the overall trends that may be found in correlating stellar properties with being a host.

ALGORITHM

To perform the classification task proposed, I decided to code my own Back-Propagation Neural Network (BPN). Despite the existence of simpler and perhaps more reliable algorithms for the task, I chose this in particular for the following reasons: 1) this course does not otherwise provide experience with neural networks in general; 2) this type of algorithm is one of the most commonly utilized for data mining in astronomy; and 3) neural networks are also significantly related to another academic interest of mine, cognitive science, as this tool based on real neural systems in the brain are seen as a means to better understand how such cognitive systems operate in general.

FIGURE 2: BASIC NETWORK MODEL



The network written for this project is three-layers (i.e. with one hidden layer), with seven input nodes for the seven stellar parameters and one output node, indicating a host star when active and a non-host when inactive. For convenience, the input nodes were scaled from 0.0 to 0.1, although later investigation showed that the network was robust to this scaling. After some experimenting, I decided upon only four nodes in the hidden layer. Furthermore, the nodes on the hidden and output layers are sigmoidal and desaturated, so their values range from 0.1 to 0.9.

FIGURE 3: BASIC ALGORITHM STRUCTURE

	INPUT/HIDDEN	HIDDEN/OUTPUT
FORWARD-FEED:	$out_j = \frac{1}{1+\exp\left(-\sum_i x_i^n w_{ji}\right)}$	$y^n = \frac{1}{1+\exp\left(-\sum_j out_j^n w_{kj}\right)}$
ERROR CALCULATION:	$\delta_k = \left(y^n - t^n\right) y^n \left(1 - y^n\right)$	$\delta_j = out_j^n \left(1 - out_j^n\right) \delta_k w_{kj}$
WEIGHT UPDATES:	$w_{ji}(t+1) = w_{ji}(t) + \eta \delta_j x_i^n$	$w_{kj}(t+1) = w_{kj}(t) + \eta \delta_k out_j$

The learning rate (η) of the network was made to be adaptive, after preliminary testing with a fixed learning rate proved to be sub-optimal. Thus, for every iteration the rate was increased by a factor of 1.3 if the MSE decreased, and decreased by a factor of 1.2 if the MSE increased. The starting rate was 5e-4, determined after some testing. To compensate for the disparity in size between host and non-host data points, a weighting system for the error was also implemented; error for host points was multiplied by a factor: (number of hosts)/(number of non-hosts), and that for non-hosts was multiplied by this same factor inverted. Both the adaptive learning rate and error weighting system improved the efficiency of the algorithm, and in some cases reduced the overall MSE. Finally, from testing it was determined that the algorithm could perform well over 30 iterations of updating weights in most cases. Thus, for simplicity this condition was used in lieu of a convergence requirement.

DETERMINING PARAMETERS

FIGURE 4: NUMBER OF HIDDEN NODES

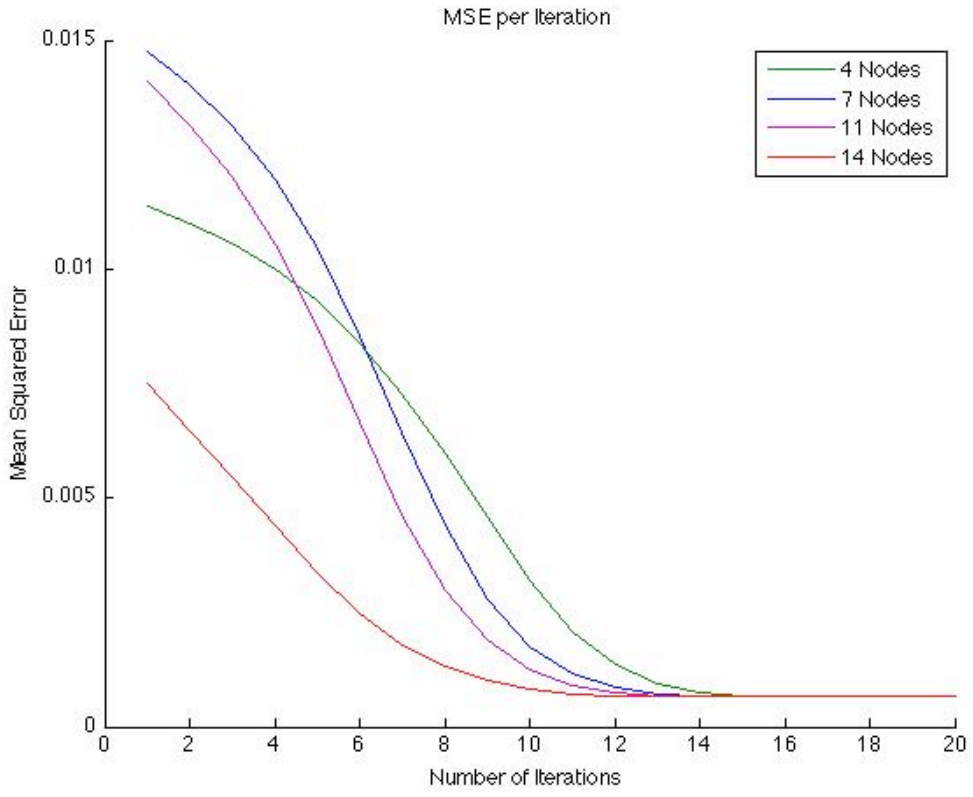
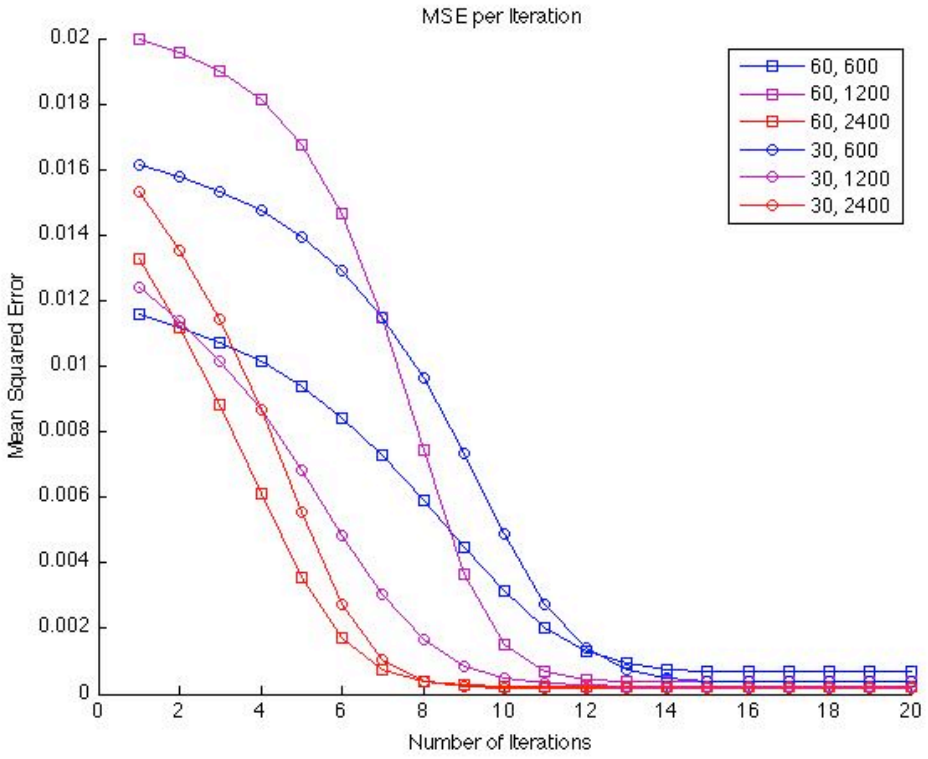


FIGURE 5: SIZE OF TRAINING GROUP

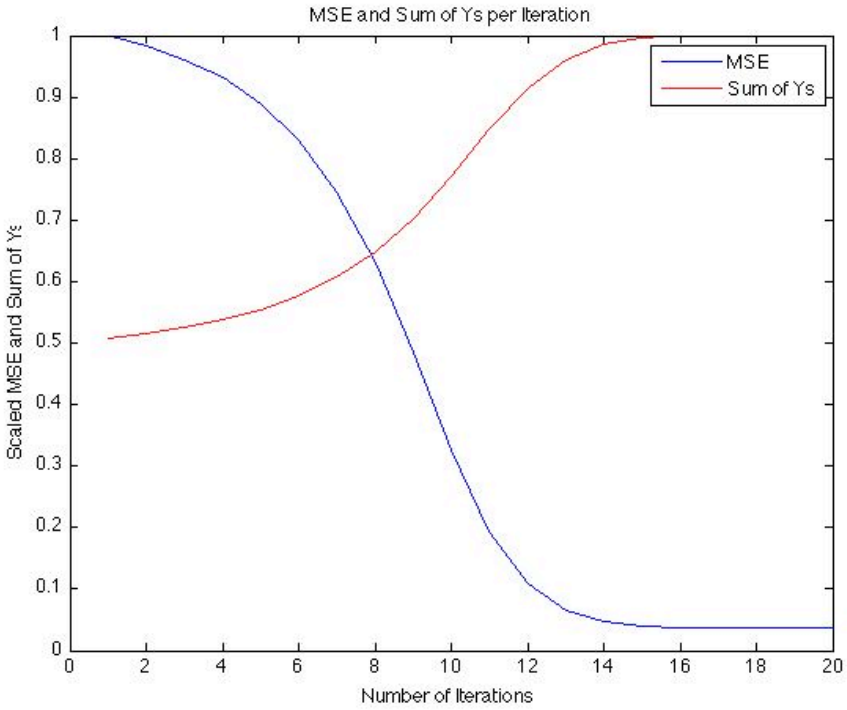


As can be gathered by the figures above, the overall shape of the error curve is not significantly influenced by the input parameters, when the other variable factors mentioned before are set the way described. Given that the data and initial weights for the algorithm are random with each implementation however, the curves given by the figures are not always true, but good for qualitative purposes. In particular, over the course of developing and running the program a "full" error curve (with a shallow then steep decline) was desirable for troubleshooting purposes, hence the early preference for the chosen parameters. Continued testing seemed to indicate that these parameters, for unclear reasons, were not terribly influential on the performance of the algorithm.

RESULTS

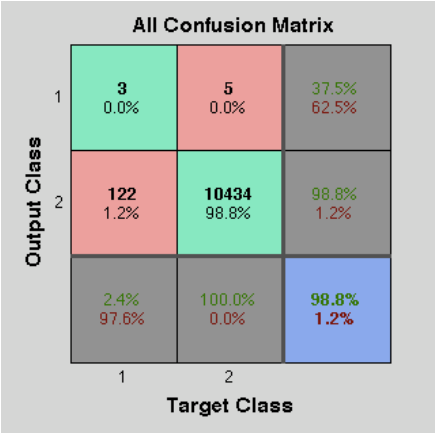
The robustness of the system as seen, for example, by the discussion immediately prior, means that the results derived from the algorithm are more indicative of the program in general and the qualities of the data used rather than specific variables such as the parameters described.

FIGURE 6: ERROR VERSUS CLASSIFICATION



Despite my best efforts, the algorithm that developed from my extended trial and error could not classify accurately; in fact, the algorithm almost always minimized the MSE by classifying all the stars as the same, such as in the figure above. This suggests two possibilities: the code is at fault, or the data is flawed. As discussed before, a few liberties were taken in the selection of data, which went along with the selection of the problem proposal in general.

FIGURE 7: NPRTOOL CONFUSION (4 HIDDEN NODES)



In order to address this, I ran my data through the Neural Pattern Recognition Tool (nprtool), a built-in tool in MATLAB, to see how well it performed on the data. As can be seen in *Figure 7*, on average nprtool only performed very marginally better than my algorithm; just like mine, the MATLAB algorithm performed maximally when all the data was approximately the same class. Similar results were produced by varying numbers of hidden nodes and training sets. This suggests that the source of mis-categorization lies not in my program, but in the data itself.

CONCLUSION

For better or worse, the determination that the data is not strongly correlated or statistically significant is not unreasonable. In particular, as mentioned before, the project procedure here assumes that those stars not labeled as hosting exoplanets do not, even though the reality is fairly uncertain. The assumption was that exoplanets are rare astronomical occurrences, and if this is the case the project proposal might not have been entirely unfeasible. In fact, these results may suggest that there are far more undiscovered exoplanets than one would believe (hence many misclassified samples); the rapidly increasing rate of discovery seems to suggest this, and experts in the field are open to this possibility. Even still, the parameters chosen may not be strongly correlated with the star's hosting an exoplanet, or no such strict set of parameters may exist at all. Unfortunately it seems the project was overly ambitious in its expectations of the data, underscoring the omnipresent naivety of scientists whose findings repeatedly supersede their expectations. Regardless, this project may yet have purpose in the future, as more exoplanets are discovered, a better understanding of the distribution and physics of exoplanets is attained, and detection methods and technologies improve as the quest to find other worlds continues to thrive.

DATA SOURCES

- http://en.wikipedia.org/wiki/List_of_exoplanetary_host_stars
- http://exoplanet.eu/catalog-RV.php
- http://exoplanets.org/table/
- http://nsted.ipac.caltech.edu/
- http://vizier.cfa.harvard.edu/viz-bin/VizieR-3

REFERENCES

1. Ball, Nicholas, and Robert J. Brunner. "Data Mining and Machine Learning in Astronomy." *International Journal of Modern Physics.* 11 June 2009.
2. Bazell, David, and Yuan Peng. "A Comparison of Neural Network Algorithms and Preprocessing Methods for Star-Galaxy Discrimination." *The Astrophysical Journal Supplement Series.* 116: 47-55. May 1998.
3. Dr. Brian Chaboyer, Dartmouth College Department of Physics and Astronomy.
4. Goos, Gerhard, Juris Hartmanis and Jan van Leeuwen, Eds. "Artificial Neural Networks: An Introduction to ANN Theory and Practice." Springer: Berlin, 1991.
5. Skapura, David M. "Building Neural Networks." ACM Press: New York, NY. 1996.
6. Soto, A., A. Cansado and F. Zavala. "Detection of Rare Objects in Massive Astronomical Datasets Using Innovative Knowledge Discovery Technology." *Astronomical Data Analysis Software and Systems XIV.* Ed.s Shopbell, P.L., M.C. Britton and R. Ebert. *ASP Conference Series.* 30: 1-5. 2005.