Chapter 10 Case Study: Galaxy Classification

10.3 Data Preparation

After removing a large number of the columns from the raw SDSS dataset, introducing a number of derived features, and generating two target features, Jocelyn generated an ABT containing 327 descriptive features and two target features. We lists these features.

Once Jocelyn had populated the ABT, she generated a data quality report (the initial data quality report covered the data in the raw SDSS dataset only, so a second one was required that covered the actual ABT) and performed an in-depth analysis of the characteristics of each descriptive feature.

The magnitude of the maximum values for the FIBER2FLUXIVAR_U feature in comparison t the median and 3rd quartile value was unusual and suggested the presence of outliers. The difference between the mean and median values for the SKYIVAR_R feature also suggested the presence of outliers. Similarly, the difference between the mean and median values for the LNLSTAR_R feature suggested that the distribution of this feature was heavily skewed and also suggested the presence of outliers.

With Edwin's help, Jocelyn investigated the actual data in the ABT to determine whether the extreme values in the feature displaying significant skew or the presence of outliers were due to **valid outliers** or **invalid outliers**. In all cases the extreme values were determined to be valid outliers. Jocelyn decided to use the **clamp transformation** to change the values of these outliers to something closer to the central tendency of the features. Any values beyond the 1st quartile value plus 2.5 times the inter-quartile range were reduced to this value. The standard value of 1. times the inter-quartile range was changed to 2.5 to slightly reduce the impact of this operation.

Jocelyn also made the decision to normalize all the descriptive features into standard scores. The differences in the ranges of values of the set of descriptive features in the ABT was huge. For example, DEVAB_R had a range as small as [0.05, 1.00] while APERFLUX7IVAR_U had a range as large as [-265,862, 15,274]. Standardizing the descriptive feature in this way was likely to improve the accuracy of the final predictive models. The only draw-back to standardization is that the models become less interpretable. Interpretability, however, was not particularly important for the SDSS scenario (the model built would be added to the existing SDSS pipeline and process thousands of galaxy objects per day), so standardization was appropriate.

Jocelyn also performed a simple first-pass feature selection using the 3rd-level model to see which features might stand out as predictive of galaxy morphology. Jocelyn used the **information gain** measure to rank the predictiveness of the different features in the dataset (for

this analysis, missing values were simply omitted). The columns identified as being most predictive of galaxy morphology were expRad_g (0.3908), expRad_r (0.3649), deVRad_g (0.3607), expRad_i (0.3509), deVRad_r (0.3467), expRad_z (0.3457), and mRrCc_g (0.3365). Jocelyn generated histograms for all these features compared to the target feature - for example, we show the histograms for the EXPRAD_R feature. It was encouraging that in many cases distinct distributions for each galaxy type were apparent in the histograms. We show small multiple box plots divided by galaxy type for a selection of features from the ABT. The differences between the three box plots in each plot gives an indication of the likely predictiveness of each feature. The presence of large numbers of outliers can also be seen.

10.4 Modeling

The descriptive features in the SDSS dataset are primarily continuous. For this reason, Jocelyn considered trying a similarity-based model, the **k nearest neighbor**, and two error-based models, the **logistic regression** model and the **support vector machine**. Jocelyn began by constructing a simple baseline model using the 3-level target feature.

10.4.1 Baseline Models

Because of the size of the ABT, Jocelyn decided to split the dataset into a **training set** and a large **hold-out test set**. Subsets of the training set would be also used for **validation** during the model building process. The training set consisted of 30% of the data in the ABT (approximately 200,000 instances), and the test set consisted of the remaining 70% (approximately 450,000 instances). Using the training set, Jocelyn performed a 10-fold cross validation experiment on models trained to use the full set of descriptive features to predict the 3-level target. These would act as baseline performance scores that she would try to improve upon. The classification accuracies achieved during the cross validation experiment were 82.912%, 86.041%, and 85.942% by the k nearest neighbor, logistic regression, and support vector machine model respectively.

These initial baseline results were promising; however, one key issue did emerge. It was clear that the performance of the models trained using the SDSS data was severely affected by the **target level imbalance** in the data-there were many more example of the elliptical target level than either the spiral or, especially, the other target level.

10.4.2 Feature Selection

In the SDSS dataset, many of the features are presented multiple times for each of the five different photometric bands, and this made Jocelyn suspect that many of these features might be

redundant and so ripe for removal from the dataset. **Feature selection** approaches that search through subsets of features (known as **wrapper** approaches) are better at removing redundant features than rank and prune approaches because they consider groups of features together. For this reason, Jocelyn chose to use a **step-wise sequential search** for feature selection for each of the three model types. In all cases overall classification accuracy was used as th fitness function that drove the search. After feature selection, the classification accuracy of the model on the test set were 85.557%, 88.829%, and 87.188% for the k nearest neighbor, logistic regression, and support vector machine models respectively. In all cases performance of the models improved with feature selection. the best performing model is the logistic regression model. For this model, just 31 out of the total 327 features were selected. this was not surprising given the large amount of redundancy within the feature set.

Based on these results, Jocelyn determined that the logistic regression model trained using the reduced set of features was the best models to use for galaxy classification. This model gave the best prediction accuracy and offered the potential for very fast classification times, which was attractive for integration into the SDSS pipeline. Logistic regression models also produce confidences along with the predictions, which was attractive to Edwin as it meant that he could build tests into the pipeline that would redirect galaxies with low confidence classifications for manual confirmation of the predictions made by the automated system.