11/24/2019
Leopold Ringmayr
COMP 347

COMP 347: Project 2–SVM and Naive Bayes to Determine Bird Molting Status

Problem:

Understanding molting patterns of birds is of great interest to ornithologists, who seek to understand the behavioral and morphological underpinnings of when and how molting takes place. Molting describes the process of growing new wings, oftentimes during specific times of the year and oftentimes noticeable by a disruption of the wing through new feathers. In this project, I seek to leverage data about a given bird, including geographical and morphological data to determine whether a bird is in molt or not. I aim to use a soft-margin SVM classifier and a Naive Bayes classifier to determine whether birds in a given dataset derived from an image database EBirds are in molt or not in molt. Based on a set of features, can we determine whether the bird is in molt or not in molt? The dataset consists of 110 entries for images of birds; the data includes entries for location, date, season the image was captured, as well as wing linearity and species classification of the bird in the image.

The objective is two build and train two models–a soft-margin SVM classifier and a Naive Bayes (built-in MATLAB Naive Bayes)–and to establish which model best suits the classification problem. In the initial stages, feature selection is used to find the most useful features for the model. Then cross validation is used to find the best hyperparameters for the models; then, the “best” model is trained. I compare the two models using bootstrap resampling, and a paired t-test to determine which model is most suitable for this classification problem.

Feature Selection:

Two main steps were taken for the feature selection process. Based on domain knowledge, I chose a subset of features I deemed to be useful. Then I checked these chosen features on correlation/collinearity and whether features were irrelevant or not via lasso regression. I converted string type data such as species or location into numerical data by creating columns with ID numbers corresponding to specific locations/species. For the models, I eliminated the string-type columns and used the newly created numerical ones. It made sense to keep a feature for location, seeing as

the location of a bird (i.e. where an image of a bird was taken), could be a feasible indicator of whether the bird is in molt. Certain locations could be where birds tend to molt. Similarly, it was important to keep the wing linearity feature. Wing linearity can be used as a physical/morphological indicator of whether a bird is in molt or not. Consulting with Professor Terrill confirmed this: Birds with linear wings (no “break” in the wing outline) is a good indicator that a bird is not in molt. New feathers from molting tend to obstruct the wing linearity. Running a correlation test for my chosen features showed that none of the features were correlated with one another.

Running lasso regression revealed that none of the selected features were irrelevant. I saw that the coefficient for the species feature was shrunk to zero first, however, not until late in the iterations. Therefore, I opted in favor of leaving the species feature in the dataset. It is important to consider because molting patterns, and indications of molting could be very different across different species. The date feature was eliminated, seeing as temporal molting patterns could generally be captured using the season feature; molting generally occurs in the Fall.

The features I ultimately kept were season, species number, location number, and wing linearity.

Training the models:

The dataset was split into test and training sets, with a 20% proportion of the data being used in the test set. I used the remaining features for training the SVM and Naive Bayes model and for tuning the hyperparameters. In addition, a 10-fold cross validation framework was utilized for the hyperparameter selection. For this process, I tried various values for lambda on the 10 folds of the cross validation, training the model each time to see which parameter values would yield the highest accuracy. For my final SVM model, I had the following final weights [8.485713605382967,16.855713605382928,-3.564286394617064,-28.324286394617] with a bias value of -38. The lambda of 1.7. Once this best lambda was established, I trained the “best” SVM model using this parameter, which is also the model that was used for testing during bootstrap resampling.

For Naive Bayes, no hyperparameters were adjusted, hence no cross-validation was utilized for this model. The Naive Bayes model was trained on the training set (88 data entries).

Testing the models and Results:
For overall model testing and comparison, bootstrapping was employed.
After testing the models, I compared the statistics of both models in a paired t-test.
On the test set, SVM had an accuracy of 100%, and a corresponding error of 0%. The data was completely linearly separable using this model. Naive Bayes had an accuracy of 90% on the bootstrap sample.

The bootstrap resampling yielded a p-value of 0.0037we therefore reject the null hypothesis. Between the Naive Bayes model and the SVM model, the SVM classifier has the higher results.

Conclusion:

From the t-test conducted using the bootstrap resampling results from both models, I conclude that the SVM model is the better model for this classification problem. The data generally is linearly separable. It is likely that other morphological features not used in this project, such as more detailed wing features, could serve as strong indicators on whether a bird is in molt or not. In addition, it is important to consider the relatively small amount of data (88 entries to train on) could impact how accurate a molt classifier is.