Identifying Tree Species via Photographs of Leaves

Geoffrey E. Schneider

9 April 2017

1 Introduction

My primary objective in this project is to practice, learn and experiment with machine learning techniques, as well as to form a kind of portfolio of my personal knowledge and abilities in this area. The project objective is to identify tree species from photos of their leaves. Because my objective is not the same as the project objective, I will not always take the most direct approach to the problem, experimenting with techniques that might not be absolutely necessary. I will however, try to mimic good practices in developing machine learning algorithms. As such, I will start with the easiest approach to the easiest version of the problem as outlined below in Section 2, as a baseline from which to compare other algorithms.

This is a work in progress, as apparent by the empty sections below.

2 Plans

Quick and dirty approach:

(a) Gather a small amount of data (1 or 2 species, 100 examples)

(b) Put data in a standard, usable form (minimal cleaning)

(c) Implement simple logistic regression

(d) Evaluate accuracy and ways to improve via learning curves and accuracy metric.

(e) Ceiling analysis (if appropriate).

Additional models to try:

- Perceptron
- CNN
- SVM

Data cleaning:

- PCA

- Autoencoder
- Background removal
- Feature choosing?

Additional data gathering:

- Manual data gathering
- Artificial data
- Other data repositories

3 First Approach

I begin by using the easiest possible approach to get results for comparison with later attempts. The first approach attempts to classify pictures of leaves from two species: *Abies Concolor* and *Abies Nordmanniana* using logistic regression.

3.1 Data

There were a total of 86 sample pictures of these leaves taken from the Leafsnap database [1]. There were 51 for *Abies Concolor* and 35 for *Abies Nordmanniana* (See Figure 3.2 for examples). They were color jpg photos.

3.2 Preprocessing

The Octave script 'loadDataScript.m' loads and preprocesses the images. Images were all approximately 600×800 pixels (± 3 pixels per dimension) when rotated such that the first dimension was shorter than the second. Because this would result in a very large feature vector, the size was reduced to 30×40 pixels for all images via the Octave command 'imresize' from the 'image' package. The 86 resulting $30 \times 40 \times 3$ matrices were unrolled into a 86×3600 whose rows were randomly permutedm the data was normalized, and then divided into matrices of training data (X), cross-validation data (X_cv) and test data (X_test). These are labelled with corresponding vectors y, y_cv and y_test. The label 0 corresponds to *Abies Concolor* and 1 to *Abies Nordmanniana*.

3.3 Model

The model trained was a logistic regression model with regularization. The hypothesis for a particular example is given by:

$$h_\theta(x) = \frac{1}{1 + e^{x \cdot \theta}}$$

Where x is a feature vector and θ is a vector of the model parameters.

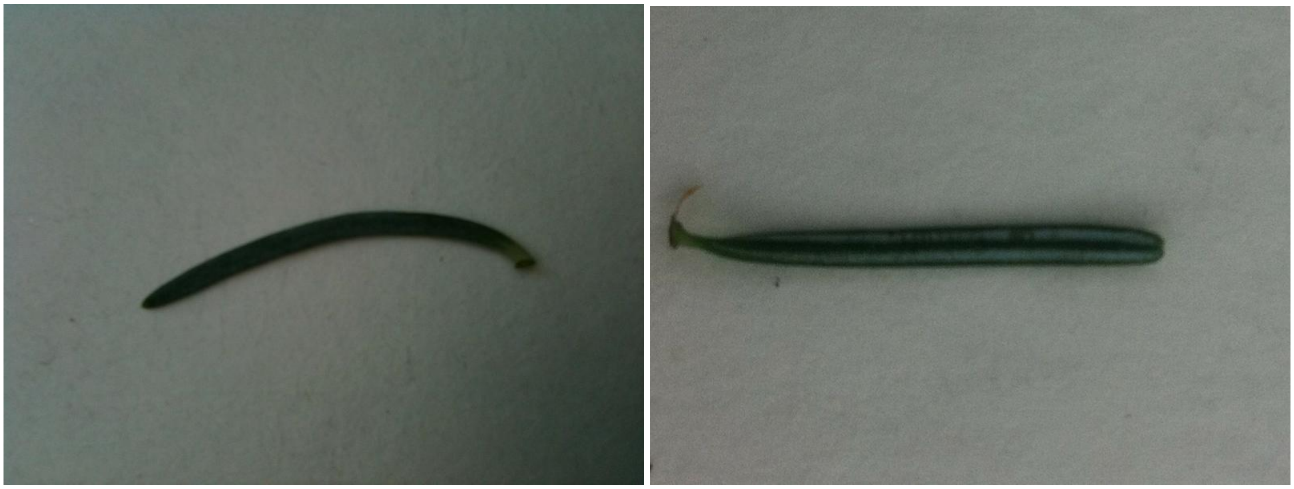

Figure 1: Example photos of Abies Concolor (left) and Abies Nordmanniana (right).

The cost function is given by

$$J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}(y_i \operatorname{Log}(h_\theta(x_i)) + (1-y_i)\operatorname{Log}(1-h_\theta(x_i)) + \frac{\lambda}{2m}\sum_{k=1}^{n}\theta_k^2$$

where y_i is the label and x_i is the feature vector of the ith sample.

3.4 Training

The model parameters were trained by minimizing the cost function using the Octave 'fminunc' function, with theta initialized as the 0 vector.

3.5 Evaluation of results

The table below shows cost results for the model trained for various choices of the regularization parameter λ (after training, cost was tested with $\lambda = 0$).

λ :	0	0.1	0.3	1
training :	1.3115e-004	0.0016223	0.0028836	0.0067081
cross-validation:	1.1418e-004	0.0023556	0.0015548	0.0019042

The pre-training cost was 0.69315 on both the training set and the cross-validation set. If we predict *Abies Concolor* when our hypothesis function gives a result < 0.5, this gives a 100% accuracy on the cross-validation set for all choices of λ.

4 Data Gathering

5 Preprocessing

6 Models

7 Training Methods

8 Evaluation of Results

References

[1] Neeraj Kumar, Peter N. Belhumeur, Arijit Biswas, David W. Jacobs, W. John Kress, Ida Lopez, João V. B. Soares, ”Leafsnap: A Computer Vision System for Automatic Plant Species Identification,” Proceedings of the 12th European Conference on Computer Vision (ECCV), October 2012.