THIS IS THE DRAFT AS OF ARTEFACT SUBMISSION
THIS IS NOT THE FINAL REPORT

Blob-match: Machine learning for cross-identification of radio surveys

James Gardner, supervisors: Cheng Soon Ong, Matthew Alger

Semester 2 2019

Abstract

Success in radio-radio survey cross-identification is determining the real, physical objects that we're looking at. The naivest measure of two sources (or blobs) being a match for an actual object is their separation on the sky. Using this separation, we train a logistic regression classifier on the TGSS (TIFR GMRT Sky Survey Alternative Data Release 1) and NVSS (NRAO VLA Sky Survey) radio surveys. Then use its predictions to partition a patch of the sky into objects, by transitively grouping any chain of predicted matches. Although the classifier successfully learns the importance of separation, we find that the naive partitioning fails to convincingly identify objects in the sky.

1 Introduction

2 Positional matching

2.1

2.2

2.3 Catalogue

3 Logistic regression

3.1 Features

3.2 Napkin-sized introduction to machine learning

all machine learning is a function: $f_w : X \to Y$

- f_w is the predictor/classifier
- X is the input (to the test/predictor), for blobmatch: the two surveys of gaussians on planes
- Y is the output, for blobmatch: probabilities of match for every pair

data is the known inputs and corresponding outputs, (i.e. "the right action")

$$\text{data} = (x_1, y_1), \ldots, (x_N, y_N)$$

loss is a metric of the difference of output of the function to the (known) data. NB: cost is not the same as loss, follow this up!

$$\text{loss} = l(y_1, f(x_1))$$

the average is not the expected value(!), which would be over the whole universe

$$\text{avg loss} = \frac{1}{N} \sum_{n=1}^{N} l(y_n, f(x_n))$$

training is to **minimise the objective**, which is the average loss "regularized" by some parameter, λ, with the norm of the weight vector, ω

$$\text{objective} = \text{avg loss} + \lambda ||\omega||^2$$

There are two meanings of "algorithm" in machine learning:

(1) the training, building the function, f_w, from training data

(2) testing/predicting, running the function on other data

3.3

3.4 Three types of predictors

3.5

4 Discussion

5 Appendix

5.1 Artefact

All the code, data, and plots used in this report can be found in the library: blobmatch, found here: `https://github.com/MatthewJA/blobmatch`

It contains (along with the various plots produced by each script):

- README.txt; detailing that the data is drawn from the two surveys: TGSS from `http://tgssadr.strw.leidenuniv.nl/doku.php`, and NVSS from `https://heasarc.gsfc.nasa.gov/W3Browse/all/nvss.html`
- positional_catalogue.py; (sky_catalogue.csv)
- feature_vectors.py; (patch_catalogue.csv, tgss.csv, nvss.csv)
- score_feature_vectors.py; (labels and sorts the above feature vectors)
- logistic_regression.py; (weights.csv, objects.csv)
- manual_labels.csv; by eye from cut-outs