Paraguay benefit allocation

K. Olding, T. Deveney, E. Barry, J. Taylor, J. Faraway, J. Molinas

ITT8 SAMBa Presentation

June 15, 2018

Data

The data we have include:

- Approx 30,000 questionnaire responses each with 234 questions during 1998-2017
- A data set of 60 questions asked to 500,000 households from 2013-2017
- Images of 20,000 houses

Objectives

We had two key goals:

- Image classification of houses into 'poor' or 'not poor'
- Apply a more sophisticated regression model for income

K. Olding, T. Deveney, E. Barry, J. Taylor, J. Faraway, J. Molinas

Classification by images

Given 20,000 images of houses, can we apply image recognition to detect houses belonging to poor people?

For example:

Poor

Not Poor

Issue: The data was unlabelled...

To get around this we could:

- Use the ground truth information in the data to assign labels (difficult)
- Manually assign labels based on how they look (easy but labourious)

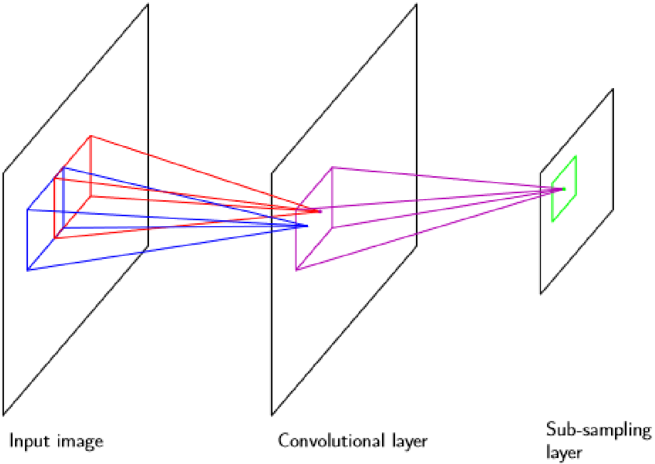
The first wasn't feasible so we went with the latter...

Labelled data:

- 672 labelled images (all rural from the same region)
- 80% (537) used for training
- 20% (135) for testing

Method

Train a convolutional neural network



Pattern recognition and machine learning - Bishop 2006

Results

After training our neural network the classification accuracy on the test set was:

	Labelled Poor	Labelled Not Poor
Predicted Poor	55	24
Predicted Not Poor	12	44

Total accuracy: 73%.

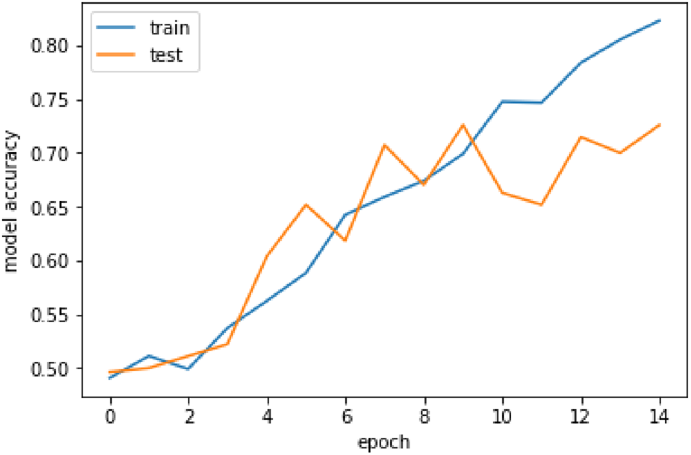
Example Classifications

	Labelled Poor	Labelled Not Poor
Predicted Poor		
Predicted Not Poor		

K. Olding, T. Deveney, E. Barry, J. Taylor, J. Faraway, J. Molinas

Accuracy

Plot of test set accuracy against Epochs



Urban?

We considered trying urban areas but the variation between residences would cause issues.

More consistent photographs would be required.

Statistical Model

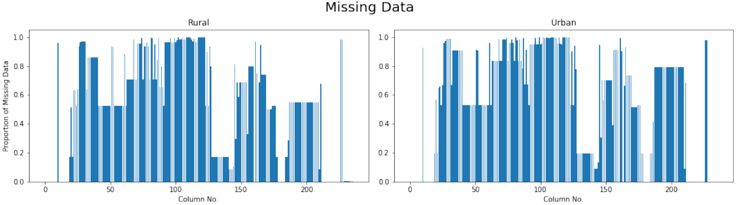
Previously poverty has been estimated with 60 variables using a linear model.

So given our data sets is there anyway to improve the model or learn from the data.

We fit a linear model to a selection of data, and perform AIC and BIC to not over-fit the data and achieve some core predictive variables. We also consider applying a nonlinear model.

Clean the Data

We looked at the sample data set with 30,000 people and 250 variables obtained by a questionnaire.



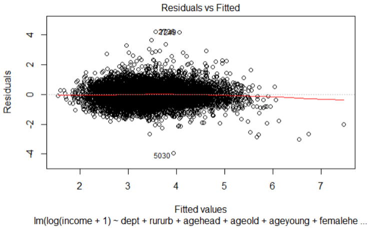
We removed variables where more than 25% of data was missing, and then rows where data responses were missing.

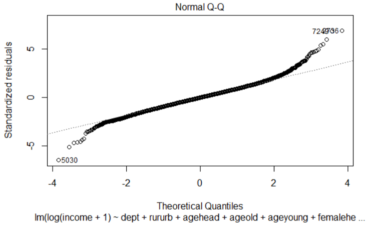
Clean the Data

This left us with 36 variables. These were mostly categorical, such as University (Yes or No?), Rural or Urban, but also 3 age categories.
Then the data was then aggregated into 8,000 households.

Generalized Linear model

We fitted a generalized linear model to the data, attempting to predict the income per household from the other variables. Even with the removal of variables using AIC selection and BIC, the adjusted R-Squared value of 0.5816, with a mean squared error 0.350.

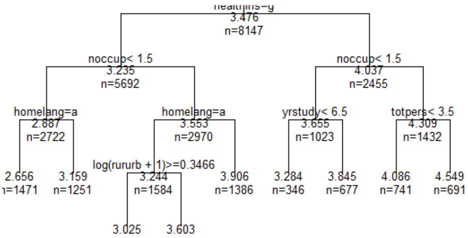




Extreme Poverty: 81%
Poverty: 71%

Decision tree modelling

Due to the categorical nature of the data we attempted a fit a decision tree model.



We find that the adjusted R-squared value is 0.402, but when a Complexity Parameter (cp=0.01) is added we find that the adjusted R-squared value is 0.546.

Extreme Poverty: 77%
Poverty: 71%

K. Olding, T. Deveney, E. Barry, J. Taylor, J. Faraway, J. Molinas

Limitations:

- May never be an appropriate method for Urban houses. However an approach could be to separate types of houses, i.e flat, terrace, room.
- Housing alone is not a true indicator of income.
- No knowledge of number of residents when using image recognition.
- Location is not taken into account.
- We performed this analysis with our own judgment of houses. Using actual income might improve it, but may also be more inaccurate.
- Variable size (33).

Further work:

- Set-up a method of taking pictures (for example what angle, no people or cars) which might mean image recognition improves in the future.
- Use satellite images of rooves in future.
- Improve image recognition with geographical locations.
- Perform more analysis on the questionnaire data, with more than 25% of the data.
- Use image recognition as a variable in statistical model.