REGRESSION FOR PROPORTION DATA

Julian Center
Creative Research Corporation
385 High Plain Rd., Andover, MA 01810 USA, jcenter@ieee.org

Abstract

Proportion data, also called compositional data [1][2], is a type of categorical data. Each observation consists of a vector **r** whose components are nonnegative and sum to one. A component of **r** represents the proportion of something that matches one of a chosen number of categories. For example, in geology, the observation may be the proportion of each of five types of minerals that are present in a rock sample. In medicine, the response to a specific treatment may be characterized by the proportions of patients whose conditions either improve, stay the same, or worsen.

Proportion data is often derived by categorizing a number of items and counting the number in each category. For example, in climate research, fossil pollen in lake sediment samples is examined to infer climate history [3]. Present day pollen samples from known climates are used as training data to calibrate a model relating climate to the proportion of plant species. For each sample, proportion data is typically derived by sorting 400 pollen grains into 14 categories and counting the number in each category. For problems such as this, a multinomial distribution is often used to model the distribution of counts.

In this type of problem, we are trying to model the probability relationship between a vector of conditions **c** (e.g. climate variables) to a proportion response vector **r** (e.g. pollen proportions). To do this we have a set of training data $\mathcal{T} = \{(\mathbf{r}_n, \mathbf{c}_n) : n = 1, \cdots, n\}$. Our objective is to approximate the function $p(\mathbf{r}|\mathbf{c}, \mathcal{T})$.

In this paper, we show how to use a multidimensional log-normal distribution to model proportion data and, if desired, to approximate closely a multinomial distribution. Using this approach, we can adapt a variety of regression techniques to approximate $p(\mathbf{r}|\mathbf{c}, \mathcal{T})$. In particular, we show how to implement Gaussian process regression and a form of kernel regression known as a Nadaraya-Watson model [4].

References:

[1] J. Aitchison, *The Statistical Analysis of Compositional Data*, Chapman & Hall, Ltd. (1986).

[2] V. Pawlowsky-Glahn and R Olea, *Geostatistical Analysis of Compositional Data*, Oxford U. Press (2004).

[3] M. Haslett, et.al., "Bayesian Palaeoclimate Reconstruction," Journal of the Royal Statistical Society, Series A, vol 169, Number 3, 1-36 (2005).

[4] C. Bishop, *Pattern Recognition and Machine Learning*, Springer (2006)

Key Words: compositional data, proportion data, categorical data, log-normal.