# Models of Human Representation

## 1 Introduction

The problem of categorisation is one that determines how an intelligent agent can group stimuli into discrete concepts and then furthermore deals with the ability of humans to generalise learned concepts to others that are innately similar along some abstract dimension.

This problem i.e., the categorical grouping of concepts and stimuli, has clear interpretations in Marr's levels of analysis (read Marr, 1892) and a multitude of other classical psychological works; and all previous attempts at categorisation have been commended and then duly condemned over the years. In particular, the development of high-precision statistical models of human behaviour that rely on decision boundaries and mathematical formulations to form distinct clusters of concepts and categories have been widely used in category theory.

## 2 Models of Categorisation

There exists extensive psychological literature of variants of models that try to accurately categorise concepts using both parametric and non-parametric methods. Given a novel stimulus that corresponds to a concept belonging to a category, it seems intuitive that we categorise it based on similarity to previously learned concepts and categories. Categorisation models therefore, can fall under a common framework where the elements of interst to us are a stimulus $y$ that we wish to categorise, a category $C$ belonging to a set of categories, a function $S(y, t)$ that determines by some similarity metric, the distance between $y$ and existing category members, expressed in a summary statistic $t_C = f(x : x \in C)$. In existing psychological literature, there are two canonical strategies that have gained prominence with regards to categorisation; a *prototype* model and an *exemplar* model. In a prototype model, a category protoype (typically the average of the existing category members) is used as a means of representation of the category for comparison. Here $t_C$ becomes the central tendency of members of category $C$. In an exemplar model, all existing category members are used as representation of a category and as a means of comparison. Therefore $t_C$ represents all existing members of "exemplars". For similarity calculation, we can use a n exponentially decreasing function to take similarity to mean distance in stimulus feature space, and $S$ is also typically an additive function to allow

summation of similarities of members of a category. We can then determine the likelihood of a single categorisation in the following way:

$$p(category = C_i|y) = \frac{S(y, t_{C_i})^\gamma}{\sum_{C_i} S(y, t_{C_i})^\gamma} \quad (1)$$

Given human judgements for representations of categories (e.g., images and text), we can reduce the likelihood of each judgement in the above equation.

$$p(category = C_j|y) = \frac{1}{1 + \sum_{C_i} e^{\gamma \log\left(\frac{S(y,t_{C_i})}{S(y,t_{C_j})}\right)}} \quad (2)$$

This therefore defines a sigmoid function around the classification boundary where $\gamma$ is a freely-estimated response-scaling parameter that controls the degree of determinism (by controlling the slope). Essentially, this kind of prototype model is equivalent to a multivariate Gaussian classifier and the exemplar model to a $k$ nearest neighbour classifier with weighted distance.

## 2.1 Variants of Prototype Models

Consider, for simplicity, a categorisation problem with only two categories.

$$S(y, t_c) = e^{-d(y)} \quad (3)$$

Here $d(y)$ can be some approximation of a distance function and this leads to the following general log-likelihood