CCA scribe

Vicky

August 11, 2015

Canonical correlation analysis

A method to identify and measure the associations between two sets of variables. Its biggest difference from PCA is that it operates on 2 axes instead of 1.

Two Examples

1. two types (sets) of measurements on students:
   - Academic: maths, reading, etc
   - Psychological: motivation, self concept, etc
2. mouse
   - Genetic: set of genes
   - Physiological: level of lipid expression

Notations X = feature matrix 1 $\in \mathbb{R}^{N*D_1}$
Y = feature matrix 2 $\in \mathbb{R}^{N*D_2}$
N = # of observations

The Problem The first pair of canonical variates
$v_1 \in \mathbb{R}^{D_1}$
$w_1 \in \mathbb{R}^{D_2}$
are defined so that
$cor(X_i^T v_1, Y_i^T w_1)$ is as large as possible

Large negative correlation is just as good (aka indicative) as large positive correlation, because they are the same thing once you flip the direction of the vector.