CS 229 Fall 2015 Vani Khosla

# Predicting a Student's Performance

### I. Problem Statement

This project aims to predicts a student's performance on a given concept, based on similar student's and their performances.

#### II. Data

The dataset for this project was provided by CK-12 Foundation, a non-profit organization whose stated mission is to reduce the cost of, and increase access to, K-12 education. The data in it's raw form (see Table below) was filtered to only include concepts in the Biology Family (SCI.BIO), which was mainly selected due to the fact that the most users have participated in concepts under Biology, and to exclude student's who have only practiced one concept, all practice attempts with less than 3 questions answered, and all concepts that were attempted by less than 100 unique students. For students that have answered some practice quizzes many times, the score taken was the average of all their attempts.

Feature	Description of Feature
Test Score ID	Unique ID for practice/quiz
Encoded Ids	Unique ID for a CK-12 concept (EID)
Student ID	Unique ID for a user
Question ID	Unique ID for a question
Level	Question difficultly level ('very easy', 'easy', 'medium', 'hard','very hard')
Correct	Set to 'true' if the user answered the question correctly. Otherwise 'false'
Time Spent	Time spent by the user to answer the question
Duration	Time spend by the user on the entire practice
Created	Timestamp denoting when the practice was started

#### IV. Predictor

The predictor built was based off of a recommendation system. For each student, the five most similar students were found using three different similarity measurements: Pearson Correlation Similarity (also known as the cosine measure similarity), Euclidean Distance Similarity, and Tanimoto Coefficient Similarity (also known as the Jaccard coefficient). These three similarity measurements were then used to predict the outcome for the same test data, and the score is reported for evaluation purposes.

## Pearson Correlation Similarity Coefficient:

$$\rho_{X,Y} = \text{cov}(X,Y)/\sigma_X\sigma_Y$$

**Euclidean Distance Similarity:** 

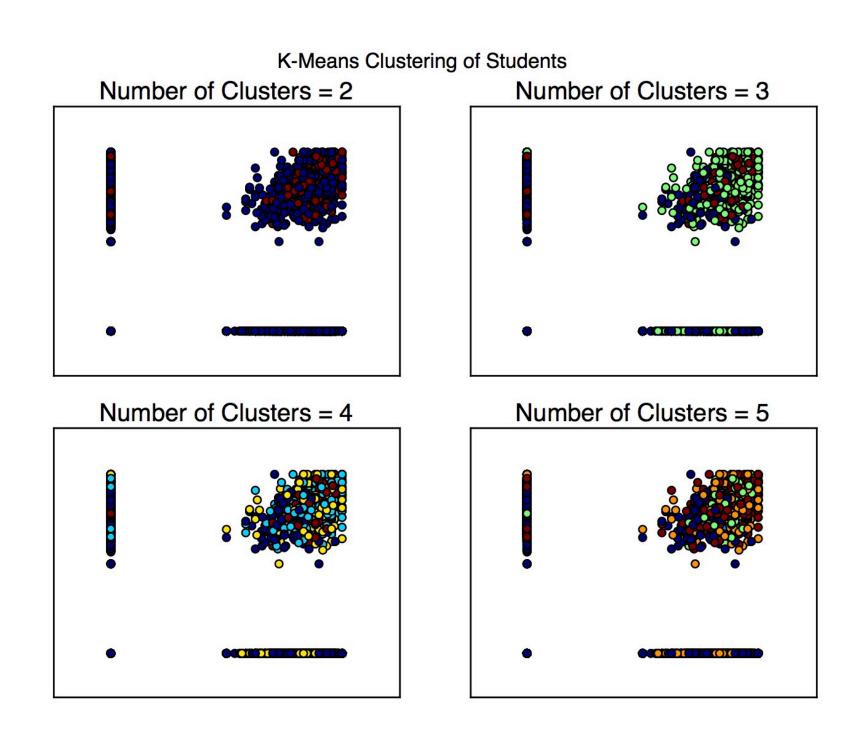
$$d(p,q) = V(\Sigma(q^i - p^i)^2)$$

**Tanimoto Coefficient Similarity:** 

$$T_{S}(X,Y) = \Sigma_{i}(X_{i} \wedge Y_{i}) / \Sigma_{i}(X_{i} \vee Y_{i})$$

$$T_{d}(X,Y) = -\log_{2}(T_{S}(X,Y))$$

### III. K-Means Clustering



Because the goal of this project is to predict a student's performance based on their similarity to other students, clustering was applied to the data to see if there were any obvious clusters of students. The clusters indicate that there is some clustering (there seems to be about three different groups within the visualization), there is no obvious clustering within students who took practice quizzes in the Science-Biology concept.

#### V. Results

Implementation	Score
Pearson Correlation Similarity	0.17587
Pearson Correlation Similarity (weighted)	0.17587
Euclidean Distance Similarity	0.17510
Euclidean Distance Similarity (weighted)	0.17510
Tanimoto Coefficient Similarity	0.13866

Results from the predictor show that the most accurate similarity measurement method is the Pearson Correlation Similarity.

The score on all implementations are not that high, which indicates that while there is a better choice of implementations here, this is not quite the best implementation for this predictor. As seen in the *k*-means clustering results, there aren't obvious clusters of students within the Biology concept. After some further investigation, it is apparent that the reason there aren't obvious clusters and the predictor doesn't perform so well is because the data is very sparse. In addition, the behavior of the predictor after applying weights, mainly that there was no difference, indicates that the data is quite sparse, and not enough students are determined to be similar to one another for weighting to have an effect.

While there are many students (51,167 in this segment of the dataset), there are many concepts, and most students have not taken the practice tests for more than two concepts. Thus finding the similarity between students is difficult in this predictor, as there is not a strong similarity between a majority of the students.