Pearson Correlation Coefficient

77B Recommender Systems
Collaborative Filtering

• Find the nearest neighbors of a particular user \( u \) by some measure of similarity.

• Average the ratings of these nearest neighbors for some item to predict its rating for user \( u \).
Pearson Correlation Coefficient

\[ PCC(\bar{x}, \bar{y}) = \frac{\sum_{i=1}^{n} (x_i - m_x)(y_i - m_y)}{\sqrt{\sum_{i=1}^{n} (x_i - m_x)^2} \sqrt{\sum_{i=1}^{n} (y_i - m_y)^2}} \]

- If there are many data-cases for which \( x_i \) is approximately equal to \( y_i \) then the value is large.
- If \( x_i \) is approximately equal to \( -y_i \) then it is large and negative.
- If the values are unrelated then it is close to 0.
- We subtract the mean so that if someone is rating consistently higher it would have no influence.
- We divide by the scale (standard deviation) so that a wider range of values has no influence.
Correlated Data

Figure 1: Comparison of NALS item difficulties

R = .94
Prediction

\[
x_{iu} = m_u + \frac{\sum_{v=1}^{U} pcc(u, v) \ast (y_{iv} - m_v)}{\sum_{v=1}^{U} pcc(u, v)}
\]

- We assume we have all ratings (relax later)
- \(Pcc\) is the similarity between two users.
- We first remove user "v" her mean rating behavior
- Then we add her rating to the pool weighted by similarity to the user \(u\).
- Divide by total weight so that sum of weights = 1
- Add user \(u\)'s mean back.