Implementational Detail: Mean Normalization

If the ranking system for movies is used from the previous lectures, then new users (who have watched no movies), will be assigned new movies incorrectly. Specifically, they will be assigned θ with all components equal to zero due to the minimization of the regularization term. That is, we assume that the new user will rank all movies 0, which does not seem intuitively correct.

- Mean normalization should let us fix this problem

We rectify this problem by normalizing the data relative to the mean. First, we use a matrix Y to store the data from previous ratings, where the i^{th} row of Y is the ratings for the i^{th} movie and the j^{th} column corresponds to the ratings for the j^{th} user.

$$Y = \begin{bmatrix} 5 & 5 & 0 & 0 & ? \\ 5 & ? & ? & 0 & ? \\ ? & 4 & 0 & ? & ? \\ 0 & 0 & 5 & 4 & ? \\ 0 & 0 & 5 & 0 & ? \end{bmatrix}$$

We can now define a vector

$$\mu = [\mu_1, \mu_2, \dots, \mu_{n_m}]$$

such that

$$\mu_i = \frac{\sum_{j:r(i,j)=1} Y_{i,j}}{\sum_j r(i,j)}$$

Which is effectively the mean of the previous ratings for the i^{th} movie (where only movies that have been watched by users are counted). We now can normalize the data by subtracting μ, the mean rating, from the actual ratings for each user (column in matrix Y):

E.g. consider the following matrix Y and mean ratings μ:

$$Y = \begin{bmatrix} 5 & 5 & 0 & 0 \\ 4 & ? & ? & 0 \\ 0 & 0 & 5 & 4 \\ 0 & 0 & 5 & 0 \end{bmatrix}, \quad \mu = \begin{bmatrix} 2.5 \\ 2 \\ 2.25 \\ 1.25 \end{bmatrix}$$

The resulting Y' vector is:

$$Y' = \begin{bmatrix} 2.5 & 2.5 & -2.5 & -2.5 \\ 2 & ? & ? & -2 \\ -.2.25 & -2.25 & 3.75 & 1.25 \\ -1.25 & -1.25 & 3.75 & -1.25 \end{bmatrix}$$

- So then we take the new set of ratings and use it with the collaborative filtering algorithm
  - Learn $\theta^{(j)}$ and $x^{(i)}$ from the mean normalized ratings

Now we must slightly modify the linear regression prediction to include the mean normalization term:

- For user j, on movie i predict: $(\theta^{(j)})^T x^{(i)} + \mu_i$

Now, for a new user, the initial predicted values will be equal to the μ term instead of simply being initialized to zero, which is more accurate.

Video Question: We talked about mean normalization. However, unlike some other applications of feature scaling, we did not scale the movie ratings by dividing by the range (max – min value). This is because:

- This sort of scaling is not useful when the value being predicted is real-valued.

  > All the movie ratings are already comparable (e.g., 0 to 5 stars), so they are already on similar scales.

- Subtracting the mean is mathematically equivalent to dividing by the range.
- This makes the overall algorithm significantly more computationally efficient.