ראשית שללנו את אפשרות השימוש בפרדיקציה נאיבית המבוססת על ממוצע שאותו נשים בכל מקום פנוי במטריצה.

הפתרון בפועל: ניצור מטריצה R המורכבת משורות המיצגות את המשתמשים. ומעמודות המייצגות את האמנים. נעבור על כל הדאטא ונמלא את ערכי המטריצה בכמות ההשמעות של כל משתמש ביחס לכל אמן.(log 10 של כמות ההשמעות)

ניצור פרדיקציה המורכבת מהממוצע שראינו בפתרון הנאיבי + ההטיה של המשתמש + ההטיה של האמן. פונקציית העלות LOSS מורכבת מערך הפרדיקציה פחות log10 של הערך האמיתי.

$$\min_{\{b_u,b_i\}} \sum_{(u,i)} \left(r_{avg} + b_u + b_i - r_{ui}\right)^2$$

פונקציית loss, נרצה שהיא תהיה מינימלית

כעת נרצה למצוא את ההטיות הנ"ל ואלו יהפכו להיות המשקולות. כל החישובים של הפרדיקציה והערך האמיתי הם ב log10 משום שיש פער של כמה סדרי גודל בין מספר הצפיות הגבוה ביותר לנמוך ביותר. וכדי שלא תהיה הטייה של המודל לטובת הדגימות של מספר הצפיות הגבוה אז נעבוד על ערכי הלוג (בסקלה מעריכית)

נסדר את b_i ו b_u כווקטור אחד ארוך b המכיל את ההטיה של המשתמשים ואת ההטיה של האמנים.

נבנה ביטוי המורכב מ *A* – המטריצה. *b* – הוקטור הנ"ל. *c* – הערך האמיתי פחות הממוצע.

אם נעשה נורמה על הביטוי, נשווה לאפס ונבצע *min* על כל הביטוי, נגיע לפתרון.

בניית המטריצה *A*

ניצור 2 רשימות, רשימה המכילה את כל המשתמשים לפי סדר ייחודי ובאותו אופן ניצור גם רשימה עבור האמנים.

נבנה מטריצה שארכה (שורות) יהיה כאורך הדאטא סט, ורוחבה (עמודות) יהיה כמספר המשתמשים + מס' האמנים, המאותחלת באפסים.

נעבור בלולאה על כל רשומה בקובץ האקסל ונמצא את המיקום של המשתמש ושל האמן בתוך הרשימות שיצרנו, ונסמן את העמודות המתאימות ב 1. וכך בעצם סמן את כל הרשומות בהן משתמש כלשהו האזין לאמן כלשהו.

כעת נוכל לפתור את המשוואה השקולה וכך למצוא את הוקטור *b*.

$$(\mathbf{A}^T\mathbf{A})\mathrm{b} = \mathbf{A}^T\mathbf{c}$$

כפי שלמדנו בהרצאה, על מנת למנוע מצב של *"overfitting"* נוסיף גורם הכללה לביטוי ונקבל:

$$\min_{\{b_u,b_i\}} \sum_{(u,i)} (\widehat{r}_{ui} - r_{ui})^2 + \lambda \left(\sum_u b_u^2 + \sum_i b_i^2 \right)$$

אנו מוסיפים את הביטוי באדום עם ריסון של סקלר λ כך שככל שנקטין את הסקלר המודל ייתן שגיאה קטנה בסט האימון ושגיאה גדולה בסט המבחן, וככל שנגדיל את הסקלר – להיפך.

מצאנו כי הערך של λ במודל האופטימלי הוא 12.

נבנה וקטור של אינדקסים באורך של כל סט המבחן ונחלק אותו באופן רנדומלי, $\frac{2}{3}$ לסט אימון ו $\frac{1}{3}$ סט מבחן על מנת לבדוק עליהם את הביצועים של המודל.

השתמשנו בספריה *sklearn* על מנת לחלק את הדאטה ולהתאים מודל לניארי לסט האימון .

$$\underbrace{A^T A}_{X}\ \underbrace{b}_{w} = \underbrace{A^T c}_{y}$$

את האופטימיזציה עשינו על ערכי λ:

הרצנו בלולאה וכל פעם העלנו את ערך X וביצענו אימון ובדקנו את ערך פונקציית העלות על הסט הבדיקה . ערך פונקציית העלות היה הנמוך ביותר עבור $\lambda = 12$

שמרנו את המודל הליניארי האופטימאלי והשתמשנו בו כדי לבצע בהמשך פרדיקציה על האקסל של הבדיקה .

לבסוף חישבנו את מטריצת "הקורלציה " בין האמנים השונים (שיטת השכנים)

$$d_{ij} = \frac{\tilde{\boldsymbol{r}}_i^T \tilde{\boldsymbol{r}}_j}{\|\tilde{\boldsymbol{r}}_i\|_2 \|\tilde{\boldsymbol{r}}_j\|_2} = \frac{\sum_u \tilde{r}_{ui}\tilde{r}_{uj}}{\sqrt{\sum_u (\tilde{r}_{ui})^2 \sum_u (\tilde{r}_{uj})^2}}$$

מודל הפרדיקציה הסופי שלנו מורכב מהמודל הליניארי + ההטיה של הקורלציה בין האומנים :

$$\hat{r}_{ui}^N = \left(r_{avg} + b_u + b_i\right) + \frac{d_{ik}\tilde{r}_{uk} + d_{il}\tilde{r}_{ul}}{|d_{ik}| + |d_{il}|}$$

את הערך שהתקבל עבור כל צמד בקובץ הבדיקה המרנו ל $10^\wedge(\hat{r}_{iu}^N)$ כדי לקבל את מספר ההשמעות הצפוי לכל צמד.

בשורות בהם הופיע משתמש או אמן שלא היה באקסל של הדאטה , שמנו את הערך הממוצע של ההשמעות (חיזוי נאיבי)