『AI・データサイエンスのための数学入門』学習補助資料

空欄の評価を、内積で予測する

評価行列・潜在因子・学習と検証

1　空欄を 0 と区別する

推薦の出発点は、「誰が、どの商品を、どのように評価したか」という関係です。次の表では評価尺度を 1〜5 とし、記号「?」を未観測とします。

	商品 A	商品 B	商品 C
利用者 1	5	?	3
利用者 2	?	4	?
利用者 3	1	?	2

観測された位置の集合は

$$\Omega = \{(1, A), (1, C), (2, B), (3, A), (3, C)\}$$

です。空欄については、好まないのか、知らないのか、まだ使っていないのか、この表だけでは分かりません。

「評価が低い」と「情報がない」は違う

空欄を 0 に置き換えて普通の二乗誤差を計算すると、「その商品への評価は 0 だった」という架空の観測を加えることになります。

小さな数値例：損失に入れるのはどれか

利用者 1 の観測が $(5, ?, 3)$、予測が $(4, 2, 3)$ なら、観測済み成分だけの二乗誤差は

$$(5-4)^2 + (3-3)^2 = 1.$$

空欄を 0 として計算すると、さらに $(0-2)^2 = 4$ が加わり、合計 5 になります。これは**別の学習問題**です。

観測マスクという表し方

観測済みなら $m_{ij} = 1$、未観測なら $m_{ij} = 0$ とする行列を観測マスクといいます。未観測位置に計算用の仮値を入れた行列 $\widetilde{R}$ を用意すれば、

$$L = \sum_{i,j} m_{ij} (\widetilde{r}_{ij} - \hat{r}_{ij})^2$$

と書けます。マスクが 0 の位置の仮値は、損失に影響しません。

2　潜在ベクトルから評価行列を作る

U の各行は利用者、V の各行は商品のベクトルです。ベクトル自体を列ベクトルで書くと、行列の行は $\boldsymbol{u}_i^\top, \boldsymbol{v}_j^\top$ となります。

$$\underbrace{U}_{N\times K}\underbrace{V^\top}_{K\times M} = \underbrace{\widehat{R}}_{N\times M}, \qquad \hat{r}_{ij} = \sum_{\ell=1}^{K} u_{i\ell} v_{j\ell}.$$

例：2 人・3 商品・2 因子

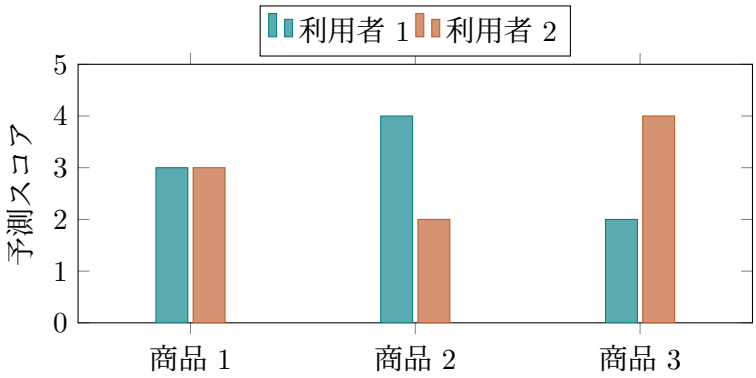
計算の仕組みを確認するため、次の因子が既に与えられているとします。

$$U = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \qquad V = \begin{pmatrix} 1 & 1 \\ 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

このとき、

$$\widehat{R} = UV^\top = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 2 & 0 \\ 1 & 0 & 2 \end{pmatrix} = \boxed{\begin{pmatrix} 3 & 4 & 2 \\ 3 & 2 & 4 \end{pmatrix}}.$$

利用者 1 では商品 2 が、利用者 2 では商品 3 が最高得点です。**同じ商品でも、利用者のベクトルが異なれば順位が変わります。**



内積は「角度」だけで決まらない

$\boldsymbol{u}^\top \boldsymbol{v} = \|\boldsymbol{u}\|\|\boldsymbol{v}\| \cos\theta$ なので、ベクトルの長さも影響します。内積の大小と、方向の近さだけを測るコサイン類似度の大小は、一般には一致しません。

3　観測データに合う因子を求める

前頁では因子を与えましたが、実際の学習では U, V が未知です。

$$L(U,V) = \sum_{(i,j)\in\Omega} (r_{ij} - \boldsymbol{u}_i^\top \boldsymbol{v}_j)^2$$

を小さくするように両者を調整します。**予測の計算と、因子を学習する計算を区別しましょう。**

1 因子の場合を手で解く

商品側の因子を $v_1 = 1, v_2 = 2$ に固定し、ある利用者の観測評価を $r_1 = 2, r_2 = 3$ とします。利用者側の未知因子を u とおくと、

$$L(u) = (2-u)^2 + (3-2u)^2 = 5u^2 - 16u + 13.$$

微分すると $L'(u) = 10u - 16$ なので、最小値を与えるのは

$$\boxed{u = \frac{8}{5} = 1.6}.$$

予測値は $(1.6, 3.2)$、二乗誤差は $0.4^2 + (-0.2)^2 = 0.20$ です。1 つの因子だけでは、両評価に完全には一致しません。

交互に最小二乗問題を解く【発展】

V を固定すれば、各利用者の $\boldsymbol{u}_i$ について最小二乗問題になります。次に U を固定して各 $\boldsymbol{v}_j$ を更新します。これを交互に繰り返す考え方があります。ただし、**片方ずつの最適化ができても、同時の大域最適解が必ず得られるとは限りません。**

因子が大きくなりすぎることを抑える

観測値への当てはめだけを追求すると、未知の評価の予測が不安定になることがあります。そこで、例えば

$$L_\lambda = L + \lambda \left(\|U\|_F^2 + \|V\|_F^2 \right), \qquad \lambda > 0$$

のような**正則化**を加えます。上の例で V を固定すると、u に依存する部分は $L(u) + \lambda u^2$ なので、

$$u = \frac{8}{5 + \lambda}$$

となります。λ を大きくすると因子は小さくなりますが、大きすぎれば必要な特徴まで弱めます。

学習に使わなかった評価で確かめる

一部の観測評価を検証用に取り分け、学習には使わず予測誤差を測ります。因子数 K や正則化の強さは、訓練誤差だけでなく検証結果も見て選びます。

4　低ランク近似との関係と、相違点

$UV^\top$ の各列は U の列の線形結合なので、

$$\mathrm{rank}(UV^\top) \le K.$$

少数の因子を使うことは、予測行列を低ランクに制限することです。

すべての評価が観測されている場合

全成分の二乗誤差は $\|R - UV^\top\|_F^2$ です。正則化がなく、因子に追加の制約もなければ、打ち切り SVD による最良ランク K 近似と結び付きます。

SVD を $R = P\Sigma Q^\top$ と書くと、例えば

$$U = P_K \Sigma_K^{1/2}, \qquad V = Q_K \Sigma_K^{1/2}$$

とすれば $UV^\top = R_K$ です。推薦の U, V は一般に直交行列ではなく、SVD の記号と役割を混同しないようにします。

空欄がある場合

観測位置だけの誤差を小さくする問題では、**空欄を 0 で埋めて SVD を適用しても、元の問題の最適解になるとは限りません。**前頁の学習は、空欄に誤差を課さない点が重要です。

低ランクという条件だけで、空欄は決まるか？

対角成分だけが観測された

$$R = \begin{pmatrix} 1 & ? \\ ? & 1 \end{pmatrix}$$

を考えます。任意の $t \neq 0$ に対して

$$B_t = \begin{pmatrix} 1 & t \\ 1/t & 1 \end{pmatrix}$$

は行列式 0、階数 1 で、観測値に完全に一致します。それでも空欄の値は t によって変わります。

少数のパラメータで表す意味

$N = 1000, M = 500, K = 10$ なら、全成分は 500000 個ですが、因子は

$$(N + M)K = 15000 \text{ 個}$$

です。ただし、これは**密な予測行列との比較**です。観測データを位置と評価だけで保存する疎な表現より、必ず小さくなるという意味ではありません。

5　潜在因子の意味と、推薦の読み方

「第 1 因子はアクション性」といった名前は直感を助けますが、学習結果の座標軸にその意味が自動的に付くわけではありません。

同じ予測をする、異なる因子

任意の直交行列 Q に対して、

$$U' = UQ, \qquad V' = VQ$$

とすると、

$$U'V'^{\top} = UQQ^{\top}V^{\top} = UV^{\top}.$$

したがって、全因子の軸を同時に回転しても予測値は同じです。また、$\|UQ\|_F = \|U\|_F$ なので、前頁までの二乗ノルム正則化もこの回転を区別しません。

因子の座標より、内積による関係を見る

因子の各成分を一意の意味として解釈するには、追加の情報や制約が必要です。数学的に直接決まるのは、モデルが作る予測値や利用者・商品間の関係です。

予測スコア 8 は「8 点の評価」とは限らない

内積には上限・下限がありません。1〜5 点の尺度でも、モデルの出力が範囲を外れることはあります。本文の 8 という値は、まずは**推薦順位を比較するスコア**として理解できます。実際の評価値として使うなら、尺度との整合も確認します。

利用者の採点傾向と、商品の人気を分ける

発展的には、全体平均 μ、利用者の偏り b_i、商品の偏り c_j を加えた

$$\hat{r}_{ij} = \mu + b_i + c_j + \boldsymbol{u}_i^{\top}\boldsymbol{v}_j$$

も考えられます。例えば $\mu = 3, b_i = 0.5, c_j = -0.2$、内積が 0.8 なら、予測は 4.1 です。常に高めの点を付ける傾向を、好みの因子だけで説明せずに済みます。

予測できる範囲を意識する

- **新規利用者・新商品：**観測がなければ、この損失だけから個別の好みや特徴は分かりません。属性などの追加情報が必要になります。
- **評価の集まり方：**よく見られる商品や、評価を投稿する利用者に観測が偏ることがあります。空欄を無視することと、偏りがなくなることは別です。
- **推薦候補：**候補集合を決めてから、その中でスコアを比較します。既に購入済みの商品などを除く場合もあります。

6　本文の練習問題を解く

問 1：内積による推薦スコア

利用者ベクトルは $\boldsymbol{u} = (2, 1)^\top$ なので、

$$\hat{r}_1 = 2 \cdot 1 + 1 \cdot 3 = \boxed{5},$$
$$\hat{r}_2 = 2 \cdot 4 + 1 \cdot 0 = \boxed{8},$$
$$\hat{r}_3 = 2 \cdot 0 + 1 \cdot 2 = \boxed{2}.$$

順位は商品 2、商品 1、商品 3 です。各因子について利用者側と商品側の成分を掛け、その寄与を足しています。この例では第 1 因子の重みが大きい商品 2 が高いスコアになります。

問 2：行列分解と損失関数

予測評価値と観測済み評価の二乗誤差は、

$$\boxed{\hat{r}_{ij} = \boldsymbol{u}_i^\top \boldsymbol{v}_j}, \qquad \boxed{L = \sum_{(i,j)\in\Omega} (r_{ij} - \boldsymbol{u}_i^\top \boldsymbol{v}_j)^2}.$$

U に NK 個、V に MK 個、合計 $(N+M)K$ 個の成分を持ちます。全成分の個数 NM より少なくなる条件は

$$\boxed{K < \frac{NM}{N+M}}.$$

共有する少数の因子で多数の評価値を作るため、成分ごとに別のパラメータを持つ必要がありません。ただし、少ない因子で十分よく予測できるかはデータによります。

問 3：欠損値の予測

利用者ベクトルは $\boldsymbol{u} = (1, 2)^\top$ なので、

$$\hat{r}_1 = 1 \cdot 3 + 2 \cdot 1 = \boxed{5}, \quad \hat{r}_2 = 1 \cdot 0 + 2 \cdot 4 = \boxed{8}, \quad \hat{r}_3 = 1 \cdot 2 + 2 \cdot 2 = \boxed{6}.$$

最高得点は商品 2 です。順位は商品 2、商品 3、商品 1 です。

埋めた値は、観測した事実ではない

$\hat{r}_{ij}$ は、空欄の位置にモデルが与える予測値です。実際に利用者が付けた評価 r_{ij} と区別します。予測した空欄を真の観測として扱うと、モデル自身の予測に合わせる学習になってしまいます。

7　追加の練習問題

問 4：空欄と損失

観測評価が $(4, ?, 2)$、予測が $(3, 5, 2)$ である。観測済み成分だけの二乗誤差を求めよ。空欄を 0 と誤って扱った場合の二乗誤差も求め、違いを説明せよ。

問 5：行列積でまとめて予測

$$U = \begin{pmatrix} 1 & 2 \\ 2 & 0 \end{pmatrix}, \qquad V = \begin{pmatrix} 2 & 1 \\ 0 & 2 \\ 1 & 1 \end{pmatrix}$$

とする。$UV^{\top}$ を求め、各利用者について 3 商品をスコア順に並べよ。すべての商品が推薦候補であるとする。

問 6：1 因子の学習と正則化

商品側の因子を $v_1 = 1, v_2 = 2$、観測評価を 3 と 4 とする。

$$L(u) = (3 - u)^2 + (4 - 2u)^2$$

を最小にする u を求めよ。さらに $L(u) + u^2$ を最小にする u を求めよ。

問 7：パラメータ数

利用者 200 人、商品 300 個、因子数 5 とする。密な予測行列の成分数と因子の成分数を求めよ。また、因子の成分数の方が少ない最大の整数 K を求めよ。

問 8：内積とコサイン類似度

$\boldsymbol{u} = (1, 0)^{\top}$、$\boldsymbol{v} = (1, 0)^{\top}$、$\boldsymbol{w} = (2, 2)^{\top}$ とする。内積とコサイン類似度で、それぞれどちらの商品が高く評価されるか。

問 9：観測に完全一致しても空欄は一意でない

$R = \left(\begin{smallmatrix} 1 & ? \\ ? & 1 \end{smallmatrix}\right)$ に対し、4 頁の B_t の $t = 1, 2$ を書き下せ。それぞれの階数と観測済み成分の誤差を求めよ。

問 10：検証用データで比較

検証用の実測値が $(4, 2)$、モデル A の予測が $(4, 4)$、モデル B の予測が $(3, 2)$ である。平均二乗誤差（MSE）を求め、この指標でどちらを選ぶか答えよ。

発展：問 11　因子の非一意性

$\boldsymbol{u} = (2, 1)^{\top}, \boldsymbol{v} = (1, 3)^{\top}$ の成分を両方とも入れ替える。元と入れ替え後の内積を比較し、潜在因子の名前が一意でないこととの関係を説明せよ。

8　追加問題の解答と考え方

問 4

正しくは $(4-3)^2+(2-2)^2=\boxed{1}$。空欄を 0 とすると $(0-5)^2=25$ を余計に加え、$\boxed{26}$ になります。未観測の商品を低評価だったと扱う誤りです。

問 5

$$UV^\top=\boxed{\begin{pmatrix}4 & 4 & 3\\ 4 & 0 & 2\end{pmatrix}}.$$

利用者 1 は商品 1 と商品 2 が同点で、その次が商品 3。利用者 2 は商品 1、商品 3、商品 2 の順です。同点の場合はスコアだけでは順序が決まりません。

問 6

$L(u)=5u^2-22u+25$ なので $L'(u)=10u-22$。最小点は $\boxed{u=11/5}$ です。正則化を加えると $6u^2-22u+25$ となり、最小点は $\boxed{u=11/6}$ です。

問 7

密な予測行列は $200\cdot 300=\boxed{60000}$ 個。因子は $(200+300)5=\boxed{2500}$ 個です。少なくなる条件は $500K<60000$、すなわち $K<120$ なので、最大整数は $\boxed{119}$ です。

問 8

内積は $\boldsymbol{u}^\top\boldsymbol{v}=1, \boldsymbol{u}^\top\boldsymbol{w}=2$ なので $\boldsymbol{w}$。コサイン類似度はそれぞれ $1, 1/\sqrt{2}$ なので $\boldsymbol{v}$ です。長さを考慮するかどうかで順位が変わります。

問 9

$B_1=\left(\begin{smallmatrix}1&1\\1&1\end{smallmatrix}\right)$、$B_2=\left(\begin{smallmatrix}1&2\\1/2&1\end{smallmatrix}\right)$。どちらも非零で行列式 0 なので階数 1。観測した対角成分の二乗誤差は両方 0 ですが、空欄の予測は異なります。

問 10

A の MSE は $(0^2+(-2)^2)/2=\boxed{2}$、B は $(1^2+0^2)/2=\boxed{0.5}$。この検証データと指標では B を選びます。これだけであらゆる利用者や将来のデータでも B が優れると断定はできません。

問 11

元は $2\cdot 1+1\cdot 3=5$。入れ替え後は $1\cdot 3+2\cdot 1=5$ です。全利用者・全商品の因子を同時に入れ替えても予測は変わらず、「第 1 因子」という位置だけから固定の意味は決まりません。