BITS F464: Machine Learning

08 Expectation Maximization SVD

Dr. Kamlesh Tiwari
Assistant Professor, Department of CSIS,
BITS Pilani, Pilani Campus, Rajasthan-333031 INDIA

Feb 03, 2021 ONLINE (Campus @ BITS-Pilani Jan-May 2021)

http://ktiwari.in/ml

Basics

- Vector: amplitude, addition, scaler multiplication, dot product and angle between two vectors

$$\cos\theta = \frac{\vec{a}.\vec{b}}{|\vec{a}|.|\vec{b}|}$$

- Matrix, transpose, multiplication and inverse

$$A^{-1} = \frac{adjoint(A)}{|A|}$$

Adjoint is transpose of co-factor matrix of A

- Non singular matrix has $|A| \neq 0$

Expectation Maximization (EM)

Mixture of Gaussian

Data may come from multiple distributions. How to separate?

Example: Consider two coins A and B with biases θ_A, θ_B (Probabilities of getting head). Let we have two vectors $x = (x_1, x_2, .., x_m)$ and $z = (z_1, z_2, .., z_m)$ specifying number of heads in 10 flips and identity of coin (where $z_i \in \{A, B\}$). Then

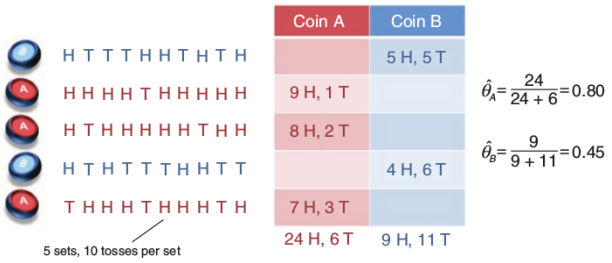
$$\hat{\theta}_A = \frac{\text{Number of heads using A}}{\text{Total coin flips using A}}$$

$$\hat{\theta}_B = \frac{\text{Number of heads using B}}{\text{Total coin flips using B}}$$

Example: let $x = (5, 9, 8, 4, 7)$ and $z = (B, A, A, B, A)$

Expectation Maximization (EM)

$x = (5, 9, 8, 4, 7)$ and $z = (B, A, A, B, A)$



What if z is not provided ? Refer z as latent (hidden) variable.

Expectation Maximization (EM)

Algorithm 1: Expectation Maximization

1. Begin with initial guess of parameters $\theta^{(0)} = (\hat{\theta}_A^{(0)}, \hat{\theta}_B^{(0)})$
2. **repeat**
3. For each of the m data point estimate z (which coin A or B, has generated this observation)
4. Assume this coin assignment to be correct and apply maximum likelyhood estimation to get $\theta^{(t+1)} = (\hat{\theta}_A^{(t+1)}, \hat{\theta}_B^{(t+1)})$
5. **until** *until converge*;
6. **return** $\theta^{(t)}$

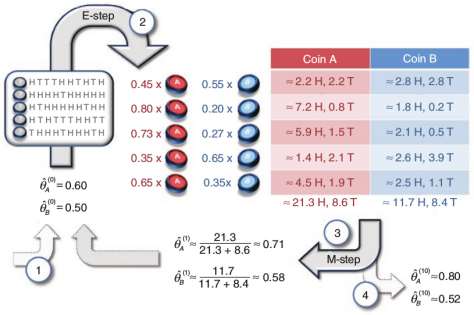
Recall (Example)

If you have a coin with $p(H) = 0.3$ what is the probability of getting 4 heads in 12 trials?

$$^{12}C_4 \times p(H)^4 \times (1 - p(H))^8$$

EM in Action

1. Compute likelihood that it is from coin A or B. Using the binomial distribution with probability θ of head on n trials with k success $p(k) = {}^n C_k \theta^k (1-\theta)^{n-k}$. Likelihood of A and B for first trial is 0.00079 and 0.00097 (prob 0.45, 0.55)



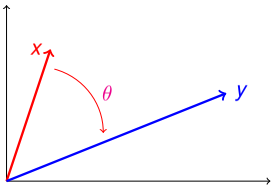
Vector and Matrix multiplication

Let

$$x = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \qquad A = \begin{bmatrix} 2 & 1 \\ -1 & 1 \end{bmatrix}$$

Consider

$$y = Ax = \begin{bmatrix} 2 & 1 \\ -1 & 1 \end{bmatrix} \times \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} 5 \\ 2 \end{bmatrix}$$

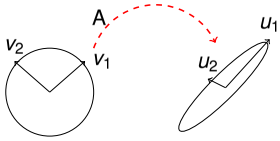


Matrix multiplication has two effect

$$\textit{Rotation} \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

$$\textit{Scaling} \begin{bmatrix} \alpha & 0 \\ 0 & \alpha \end{bmatrix}$$

What happens to a circle under Matrix multiplication



- Circle becomes ellipse
- If unit vectors along major and minor axis be u_1 and u_2 then orthogonal vectors v_1 and v_2 on the circle becomes $\sigma_1 u_1$ and $\sigma_2 u_2$
- Essentially it is a transformation from one **vector space** with v_1, v_2, ...v_n to new vector space u_1, u_2, ...u_n along with stretching factor $\sigma_1, \sigma_2, ...\sigma_n$. Such that $A \times v_j = \sigma_j u_j \quad j \in \{1, ..., n\}$.

$$[A]\,[v_1, v_2, ..., v_n] = [u_1, u_2, ..., u_n] \begin{bmatrix} \sigma_1 & 0 & ... & 0 \\ 0 & \sigma_2 & ... & 0 \\ 0 & 0 & ... & 0 \\ 0 & 0 & ... & \sigma_n \end{bmatrix}$$

$$AV = U\Sigma \tag{1}$$

What happens to a circle under Matrix multiplication

$$AV = U\Sigma \tag{2}$$

When V being a *orthogonal matrix* (*i.e.* it's transpose is inverse). Multiply V^T to both side of Equation.2
$AV \times V^T = U\Sigma V^T$

$$A = U\Sigma V^T$$

- One matrix is being represented as a product of three matrices
- V^T and U are orthogonal matrices that are related to rotation
- Σ is related to scaling or stretching

1. Every matrix $A_{m\times n}$ has SVD decomposition
2. Singular values $\{\sigma_i\}$ are positive and are uniquely determined. Also $\sigma_i \geq \sigma_j \geq 0 \quad \forall i \leq j$
3. $\{u_j\}$ and $\{v_j\}$ are also unique.

How to find SVD

$$A = U\Sigma V^T$$

$$\begin{aligned} A^TA &= (U\Sigma V^T)^T U\Sigma V^T \\ &= (V\Sigma^T U^T)U\Sigma V^T \\ &= V\Sigma^T(U^TU)\Sigma V^T \\ &= V\Sigma(I)\Sigma V^T \\ &= V\Sigma^2 V^T \\ (A^TA)V &= (V\Sigma^2 V^T)V \\ &= V\Sigma^2 \end{aligned}$$

$$\begin{aligned} AA^T &= U\Sigma V^T(U\Sigma V^T)^T \\ &= U\Sigma V^T(V\Sigma^T U^T) \\ &= U\Sigma(V^TV)\Sigma^T U^T \\ &= U\Sigma(I)\Sigma U^T \\ &= U\Sigma^2 U^T \\ (AA^T)U &= (U\Sigma^2 U^T)U \\ &= U\Sigma^2 \end{aligned}$$

1. Both the branches lead to a eigne value problem like $A \times x = \lambda \times x$
2. Therefore U and V are eigen vectors

Example

$$A = \begin{bmatrix} 5 & 5 \\ -1 & 7 \end{bmatrix}$$

$$A^TA = \begin{bmatrix} 26 & 18 \\ 18 & 74 \end{bmatrix}$$

$$A^TA = \begin{bmatrix} 5 & -1 \\ 5 & 7 \end{bmatrix} \times \begin{bmatrix} 5 & 5 \\ -1 & 7 \end{bmatrix}$$

$$det(A^TA - \lambda I) = \begin{vmatrix} 26-\lambda & 18 \\ 18 & 74-\lambda \end{vmatrix}$$

$$\lambda^2 - 100\lambda + 1600 = 0$$

$$(\lambda - 20)(\lambda - 80) = 0$$

$$\lambda = 20, 80$$

Eigen vector for $\lambda = 80$

$$(A^TA - 80I) = 0$$

$$\begin{bmatrix} -54 & 18 \\ 18 & -6 \end{bmatrix} \times \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = 0$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{10}} \\ \frac{3}{\sqrt{10}} \end{bmatrix}$$

Example

Eigen vector for $\lambda = 20$

$$(A^TA - 20I) = 0$$

$$\begin{bmatrix} 6 & 18 \\ 18 & 54 \end{bmatrix} \times \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = 0$$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \frac{-3}{\sqrt{10}} \\ \frac{1}{\sqrt{10}} \end{bmatrix}$$

Therefore,

$$V = \begin{bmatrix} \frac{1}{\sqrt{10}} & \frac{-3}{\sqrt{10}} \\ \frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}} \end{bmatrix}$$

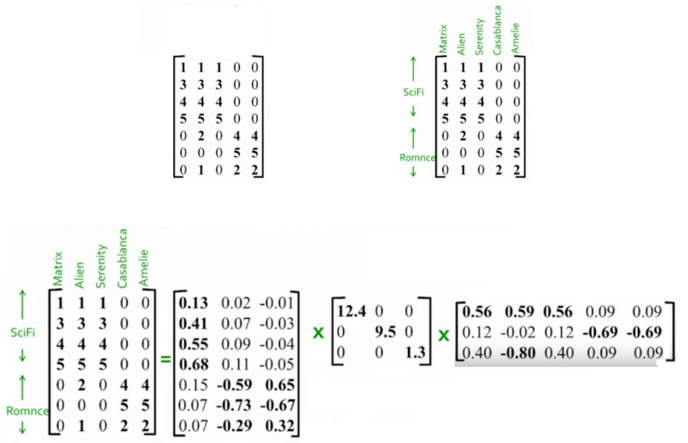
$$\Sigma = \begin{bmatrix} 4\sqrt{5} & 0 \\ 0 & 2\sqrt{5} \end{bmatrix}$$

Using $AV = U\Sigma$

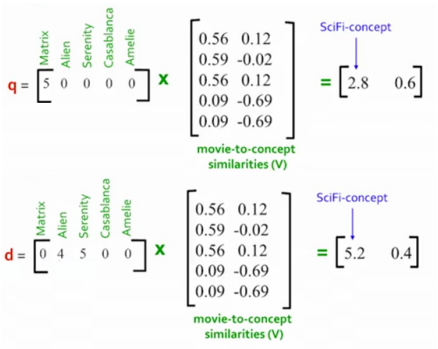
$$\begin{bmatrix} 5 & 5 \\ -1 & 7 \end{bmatrix} \times \begin{bmatrix} \frac{1}{\sqrt{10}} & \frac{-3}{\sqrt{10}} \\ \frac{3}{\sqrt{10}} & \frac{1}{\sqrt{10}} \end{bmatrix} = U \times \begin{bmatrix} 4\sqrt{5} & 0 \\ 0 & 2\sqrt{5} \end{bmatrix}$$

$$U = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

A Case Study



A Case Study



Conclusion

- SVD: every matrix can be decomposed in three components

$$\underset{m\times n}{A} = \underset{m\times r}{U} \times \underset{r\times r}{\Sigma} \times \underset{r\times n}{V^T}$$

- Time complexity is $O(m^2n)$ or $O(mn^2)$ as $AA^T = U\Sigma^2U^T$
- SVD provides best possible projection and is an optimal low rank approximation in terms of **Frobenius norm** $\sqrt{\sum(a_{ij} - b_{ij})^2}$
- For dimensionality reduction, retain 80-90% of energy ($\sum \sigma_i^2$)
- Interpretation is hard. a singular vector specifies a linear combination of all input columns or rows.
- Lack of sparsity. singular vectors are dense.

Thank You!

Thank you very much for your attention!

Queries ? Ref[1]

[1] Victor Lavrenko: curse of dimensionality [2] Jon Shlens A TUTORIAL ON PRINCIPAL COMPONENT ANALYSIS [3] Victor Lavrenko eigen-faces [4] **Face Recognition Using Eigenfaces**, by Turk, CVPR 91.