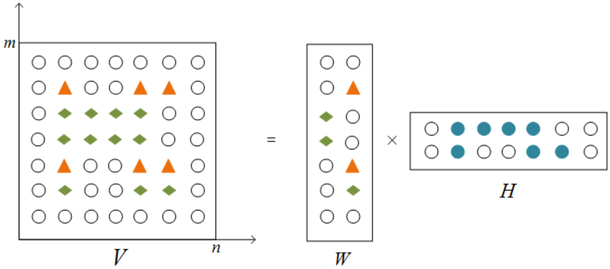
Non-negative Matrix Factorization (NMF)

Présentation des algorithmes



Principe de la NMF - A Review of Face Recognition Technology, Lixiang Li & Al

Sommaire

But

Parametres

- $V \in \mathbb{R}_+^{n\times m}$: matrice de données
- $W \in \mathbb{R}_+^{n\times r}$: matrice de base
- $H \in \mathbb{R}_+^{r\times m}$: matrice de coefficients
- Fonction de cout $L(V, WH)$: qualité de l'approximation
- Fonction de régularisation $R(V, WH)$: contrainte sur les paramètres pour avoir propriétés souhaitées (lisses, creuses, etc.)

Probleme d'optimisation non linéaire avec contraintes

$$\min_{W,H\geq 0} \left(L(V, WH) + R(V, WH)\right)$$

Probleme

- Probleme d'optimisation non convexe : NMF NP-difficile
- Algorithmes heuristiques : convergence vers minimum global non garantis, mais possiblement vers des points stationnaires (conditions de KKT)
- Mal définie : Non unicité de W et H : $V = WH = (WQ)(Q^{-1}H)$ avec Q inversible et $WQ \geq 0$ et $Q^{-1}H \geq 0$

Non unicité

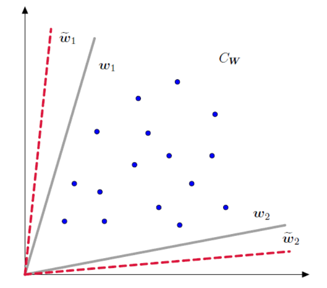


Figure – Non unicité de la NMF

Fonction de cout

Distance euclidienne (norme de Frobenius de l'erreur)

- $L(V, WH) = \frac{1}{2}||V - WH||_F^2 = \sum_{i=1}^n \sum_{j=1}^n (V_{ij} - (WH)_{ij})^2$
- Pour les cas : $V = WH + N$ où N est une matrice gaussienne de bruit
- Pas le choix idéal pour des matrices non-négatives creuses

divergence Kullback-Leibler généralisée de V depuis WH

- $L(V, WH) = D(V||WH) = \sum_{i=1}^n \sum_{j=1}^n (V_{ij} \log(\frac{V_{ij}}{(WH)_{ij}}) - V_{ij} + (WH)_{ij})^2$
- V et WH sont vues comme des distributions de probabilités normalisées

Algorithme générale

Algorithm Algorithme générale de NMF

Entree : Matrice $V \in {\mathbb{R}_+}^{n \times m}$; rang r
Sortie : NMF de rang r de V : $W \in {\mathbb{R}_+}^{n \times r}$, $H \in {\mathbb{R}_+}^{r \times m}, V \approx WH$
Initialisation : $W \in {\mathbb{R}_+}^{n \times r}$, $H \in {\mathbb{R}_+}^{r \times m}$
repeat
    $W = maj(V, H, W)$
    $H = maj(V^T, W^T, H^T)$
until arret

Raison

S'appuie sur le fait que $||V - WH||_F^2 = ||V^T - H^T W^T||_F^2$

KKT

KKT pour L fonction distance

- $W \geq 0$ et $H \geq 0$
- $\nabla_W L(V, W, H) = (WH - V)H^T \geq 0$
- $\nabla_H L(V, W, H) = W^T(WH - V) \geq 0$
- $\nabla_W L(V, W, H) \odot W = 0$
- $\nabla_H L(V, W, H) \odot H = 0$

Règle de mise à jour additives

Pour fonction de cout : distance euclidienne

- $W_{ij} \leftarrow W_{ij} - \eta_{ij}((WH - V)H^T)_{ij} = W_{ij} - \eta_{ij}\nabla_W L(V, W, H)$
- $H_{ij} \leftarrow H_{ij} - \mu_{ij}(W^T(WH - V))_{ij} = H_{ij} - \mu_{ij}\nabla_H L(V, W, H)$
- μ_{ij} et η_{ij} sont des pas positifs et petits.

Details

- Convergence lente si pas trop petit
- Facile à implémenter
- Décroissance monotone de la fonction de cout si pas suffisamment petit

Règle de mise à jour multiplicatives

Pour fonction de cout : distance euclidienne

- $W_{ij} \leftarrow W_{ij} \frac{(VH^T)_{ij}}{(WHH^T)_{ij}}$
- $H_{ij} \leftarrow H_{ij} \frac{(W^TV)_{ij}}{(W^TWH)_{ij}}$

Pour fonction de cout : divergence de Kullback-Leibler

- $W_{ij} \leftarrow W_{ij} \frac{\sum_{k=1}^{m} H_{jk} \frac{V_{ik}}{(WH)_{ik}}}{\sum_{k=1}^{m} H_{jk}}$
- $H_{ij} \leftarrow H_{ij} \frac{\sum_{k=1}^{n} W_{ki} \frac{V_{kj}}{(WH)_{kj}}}{\sum_{k=1}^{n} W_{ki}}$

Details

- Multiplication par un terme qui vaut 1 lorsque $V = WH$ (point fixe)
- Decroissance monotone de la fonction de cout
- Pas de garantie de convergence mais garantit $W, H \geq 0$
- Algorithme de gradient descendant avec $\mu = \frac{H}{W^T W H}$ pour la distance euclidienne
- Algorithme de gradient descendant avec $\mu_{ij} = \frac{H_{ij}}{\sum_k W_{ki}}$ pour la divergence de Kullback-Leibler
- Si un terme de W ou H est nul, il le reste : blocage sur des minimums locaux pauvres
- Augmentation de W_{ij} si $\nabla_W L(V, W, H)_{ij} \leq 0$ et diminution sinon
- Sauf si $W_{ij} = 0$ et $\nabla_W L(V, W, H)_{ij} \leq 0$: Ne converge pas forcement vers point stationnaire

Details

- Possiblité d'ajout $\epsilon > 0$ au numérateur et dénominateur : évite division par 0 et termes négatifs dus aux erreurs numériques
- Convergence lente
- Possibilité de mettre à jour W plusieurs fois avant de mettre à jour H et vice versa

Algorithme des moindres carrés alternées

Regles de mise a jour

- $W \leftarrow \max((argmin_{W \in \mathbb{R}^{n \times r}}(V - WH)), 0)$, max terme à terme.
- $H \leftarrow \max((argmin_{H \in \mathbb{R}^{n \times r}}(V - WH)), 0)$

Details

- Exploite la convexité de la fonction de cout par rapport à W ou à H lorsque l'autre est fixé : méthode des moindres carrés simples
- projection pour assurer $W \geq 0$ et $H \geq 0$
- Ne converge pas forcement : la fonction de cout peut osciller
- Utile pour initialiser d'autres algorithmes (par exemple multiplicatif)

Variantes

Regles de mise a jour

- ANLS : $W \leftarrow (argmin_{W \geq 0}(V - WH))$
- HALS :
  $W(:,j) \leftarrow argmin_{W(:,j) \geq 0}(V - \sum_{i \neq j} W(:,i)H(i,:) - W(:,j)H(j,:))$

Details

- ANLS : convergence garantie vers point stationnaires et reduction la plus importante de l'erreur à chaque itération
- ANLS : couteuse et plus utile pour raffiner d'autres algorithmes
- ANLS : Probleme de mise à l'echelle
- HALS : convergence garantie vers points stationnaires, plus rapide que regles multiplicatives avec presque le meme cout

Comparaisons

Comparaisons

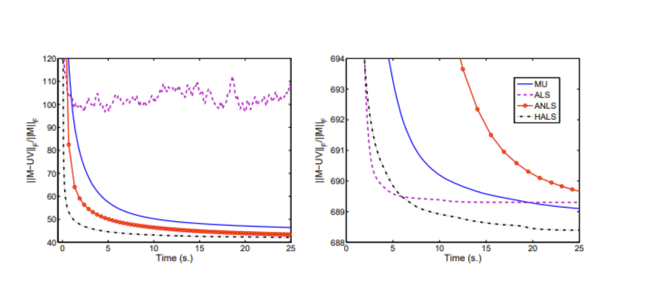


Figure – Comparaisons des algorithmes (V dense à gauche, et creuse à droite) - The Why and How of Nonnegative Matrix Factorization, Nicolas Gillis

Choix des parametres

Choix de r

- Tester plusieurs valeurs de r et choisir la meilleure en fonction des performances voulus
- Estimation à l'aide de la décroissance des valeurs singulières de V

Initialisation

- Initialisation aléatoire (avec des valeurs positives)
- Classification : k-voisins les plus proches pour calculer les centroides pour initialiser W, H à partir de la matrice indicatrice des classes : ($H_{ij} \neq 0$ si $X(:, j)$ est dans la classe i)
- À partir de l'approximation de meilleur rang de V par SVD
- Permet d'améliorer la vitesse de convergence et la qualité (meilleur point stationnaire)

Couts

Couts

- $O(nmr)$ pour la plupart des algorithmes heuristiques
- Existence d'une solution polynomiale pour les matrices quasi séparable : $V = V(:, K)H + N$ avec K ensemble d'indices, $H \geq 0$
- Existence d'un algorithme pour le calcul exact en $O((nm)^{r^2})$