

0 Essentials

Matrix/Vector

Range, Kernel, Nullity: $\text{range}(\mathbf{A}) = \{\mathbf{z} | \exists \mathbf{x} : \mathbf{z} = \mathbf{A}\mathbf{x}\} = \text{span}(\text{columns of } \mathbf{A})$
 $\text{rank}(\mathbf{A}) = \dim(\text{range}(\mathbf{A}))$
 $\text{kernel}(\mathbf{A}) = \{\mathbf{x} : \mathbf{A}\mathbf{x} = \mathbf{0}\}$ (spans nullspace)
 $\text{nullity}(\mathbf{A}) = \dim(\text{kernel}(\mathbf{A}))$

Rank-nullity Theorem: $\dim(\text{kernel}(\mathbf{A})) + \dim(\text{range}(\mathbf{A})) = n$

Orthogonal Matrix: $\mathbf{A}^{-1} = \mathbf{A}^\top$, $\mathbf{A}\mathbf{A}^\top = \mathbf{A}^\top\mathbf{A} = \mathbf{I}$, $\det(\mathbf{A}) \in \{+1, -1\}$, $\det(\mathbf{A}^\top\mathbf{A}) = 1$, preserves inner product, norm, distance, angle, rank, matrix orthogonality

Inner Product: $\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^\top \mathbf{y} = \sum_{i=1}^N \mathbf{x}_i \mathbf{y}_i$
• $\langle \mathbf{x} \pm \mathbf{y}, \mathbf{x} \pm \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{x} \rangle \pm 2\langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{y}, \mathbf{y} \rangle$
• $(\mathbf{u}_i^\top \mathbf{v}_j) \mathbf{v}_j = (\mathbf{v}_j \mathbf{v}_j^\top) \mathbf{u}_i$

Outer Product: $\mathbf{u}\mathbf{v}^\top$, $(\mathbf{u}\mathbf{v}^\top)_{i,j} = \mathbf{u}_i \mathbf{v}_j$

Trace: $\text{trace}(\mathbf{XYZ}) = \text{trace}(\mathbf{ZXY})$

Transpose: $(\mathbf{A}^\top)^{-1} = (\mathbf{A}^{-1})^\top$, $(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top$, $(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top$

Cross product: $\vec{a} \times \vec{b} = (a_2 b_3 - a_3 b_2, a_3 b_1 - a_1 b_3, a_1 b_2 - a_2 b_1)^\top$

Cauchy-Schwarz inequality: $|\langle \mathbf{u}, \mathbf{v} \rangle| \leq \|\mathbf{u}\| \|\mathbf{v}\|$

Norms

- $\|\mathbf{x}\|_0 = |\{i | x_i \neq 0\}|$
- $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^N \mathbf{x}_i^2} = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$
- $\|\mathbf{u} - \mathbf{v}\|_2 = \sqrt{(\mathbf{u} - \mathbf{v})^\top (\mathbf{u} - \mathbf{v})}$
- $\|\mathbf{x}\|_p = \left(\sum_{i=1}^N |x_i|^p\right)^{\frac{1}{p}}$

• $\|\mathbf{M}\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n \mathbf{m}_{i,j}^2} = \sqrt{\sum_{i=1}^{\min\{m,n\}} \sigma_i^2} = \|\sigma(\mathbf{A})\|_2 = \sqrt{\text{trace}(\mathbf{M}^\top \mathbf{M})}$

• $\|\mathbf{M}\|_G = \sqrt{\sum_{i,j} g_{ij} x_{ij}^2}$ (weighted Frobenius)

• $\|\mathbf{M}\|_1 = \sum_{i,j} |m_{i,j}|$

• $\|\mathbf{M}\|_2 = \sigma_{\max}(\mathbf{M}) = \|\sigma((\mathbf{M}))\|_\infty$

• $\|\mathbf{M}\|_p = \max_{\mathbf{v} \neq 0} \frac{\|\mathbf{M}\mathbf{v}\|_p}{\|\mathbf{v}\|_p}$

• $\|\mathbf{M}\|_\star = \sum_{i=1}^{\min(m,n)} \sigma_i = \|\sigma(\mathbf{A})\|_1$ (nuclear)

Derivatives

$\frac{\partial}{\partial \mathbf{x}} (\mathbf{b}^\top \mathbf{x}) = \frac{\partial}{\partial \mathbf{x}} (\mathbf{x}^\top \mathbf{b}) = \mathbf{b}$ $\frac{\partial}{\partial \mathbf{x}} (\mathbf{x}^\top \mathbf{x}) = 2\mathbf{x}$
 $\frac{\partial}{\partial \mathbf{x}} (\mathbf{x}^\top \mathbf{A} \mathbf{x}) = (\mathbf{A}^\top + \mathbf{A}) \mathbf{x}$ $\frac{\partial}{\partial \mathbf{x}} (\mathbf{b}^\top \mathbf{A} \mathbf{x}) = \mathbf{A}^\top \mathbf{b}$
 $\frac{\partial}{\partial \mathbf{X}} (\mathbf{c}^\top \mathbf{X} \mathbf{b}) = \mathbf{c} \mathbf{b}^\top$ $\frac{\partial}{\partial \mathbf{X}} (\mathbf{c}^\top \mathbf{X}^\top \mathbf{b}) = \mathbf{b} \mathbf{c}^\top$
 $\frac{\partial}{\partial \mathbf{x}} (\|\mathbf{x} - \mathbf{b}\|_2) = \frac{\mathbf{x} - \mathbf{b}}{\|\mathbf{x} - \mathbf{b}\|_2}$ $\frac{\partial}{\partial \mathbf{x}} (\|\mathbf{x}\|_2^2) = \frac{\partial}{\partial \mathbf{x}} (\mathbf{x}^\top \mathbf{x}) = 2\mathbf{x}$
 $\frac{\partial}{\partial \mathbf{X}} (\|\mathbf{X}\|_F^2) = 2\mathbf{X}$ $\frac{\partial}{\partial \mathbf{x}} \log(x) = \frac{1}{x}$

Eigendecomposition

$\mathbf{A} \in \mathbb{R}^{N \times N}$ then $\mathbf{A} = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^{-1}$ with $\mathbf{Q} \in \mathbb{R}^{N \times N}$.
if fullrank: $\mathbf{A}^{-1} = \mathbf{Q} \mathbf{\Lambda}^{-1} \mathbf{Q}^{-1}$ and $(\mathbf{\Lambda}^{-1})_{i,i} = \frac{1}{\lambda_i}$.

if \mathbf{A} symmetric: $\mathbf{A} = \mathbf{Q} \mathbf{\Lambda} \mathbf{Q}^\top$ (\mathbf{Q} orthogonal).

Probability / Statistics

• $P(x) := \Pr[X = x] := \sum_{y \in Y} P(x, y) \cdot P(x|y) := \Pr[X = x | Y = y] := \frac{P(x, y)}{P(y)}$, if $P(y) > 0$ • $\forall y \in Y : \sum_{x \in X} P(x|y) = 1$ (property for any fixed y) • $P(x, y) = P(x|y)P(y) \cdot P(x|y) = \frac{P(y|x)P(x)}{P(y)}$ (Bayes' rule) • $P(x|y) = P(x) \Leftrightarrow P(y|x) = P(y)$ (iff X, Y independent) • $P(x_1, \dots, x_n) = \prod_{i=1}^n P(x_i)$ (iff IID) • Variance $\text{Var}[X] := E[(X - \mu_x)^2] := \sum_{x \in X} (x - \mu_x)^2 P(x) = E(X^2) - E(X)^2$ • expectation $\mu_x := E[X] := \sum_{x \in X} x P(x)$ • standard deviation $\sigma_x := \sqrt{\text{Var}[X]}$

Lagrangian Multipliers

Minimize $f(\mathbf{x})$ s.t. $g_i(\mathbf{x}) \leq 0$, $i = 1, \dots, m$ (inequality constr.) and $h_i(\mathbf{x}) = \mathbf{a}_i^\top \mathbf{x} - b_i = 0$, $i = 1, \dots, p$ (equality constraint)

$L(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v}) := f(\mathbf{x}) + \sum_{i=1}^m \lambda_i g_i(\mathbf{x}) + \sum_{i=1}^p v_i h_i(\mathbf{x})$

1 Principal Component Analysis

$\mathbf{X} \in \mathbb{R}^{D \times N}$. N observations, K rank.

1. Empirical Mean: $\bar{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n$.
2. Center Data: $\bar{\mathbf{X}} = \mathbf{X} - [\bar{\mathbf{x}}, \dots, \bar{\mathbf{x}}] = \mathbf{X} - \mathbf{M}$.
3. Cov.: $\Sigma = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^\top = \frac{1}{N} \bar{\mathbf{X}} \bar{\mathbf{X}}^\top$.
4. Eigenvalue Decomposition: $\Sigma = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^\top$.
5. Select $K < D$, only keep $\mathbf{U}_K, \boldsymbol{\lambda}_K$.
6. Transform data onto new Basis: $\bar{\mathbf{Z}}_K = \mathbf{U}_K^\top \bar{\mathbf{X}}$.
7. Reconstruct to original Basis: $\tilde{\bar{\mathbf{X}}} = \mathbf{U}_K \bar{\mathbf{Z}}_K$.
8. Reverse centering: $\tilde{\mathbf{X}} = \tilde{\bar{\mathbf{X}}} + \mathbf{M}$.

For compression save $\mathbf{U}_K, \bar{\mathbf{Z}}_K, \bar{\mathbf{x}}$.

$\mathbf{U}_k \in \mathbb{R}^{D \times K}$, $\Sigma \in \mathbb{R}^{D \times D}$, $\bar{\mathbf{Z}}_K \in \mathbb{R}^{K \times N}$, $\bar{\mathbf{X}} \in \mathbb{R}^{D \times N}$

Iterative View

Residual \mathbf{r}_i : $\mathbf{x}_i - \tilde{\mathbf{x}}_i = \mathbf{I} - \mathbf{u} \mathbf{u}^\top \mathbf{x}_i$
Cov of \mathbf{r} : $\frac{1}{n} \sum_{i=1}^n (\mathbf{I} - \mathbf{u} \mathbf{u}^\top) \mathbf{x}_i \mathbf{x}_i^\top (\mathbf{I} - \mathbf{u} \mathbf{u}^\top)^\top = (\mathbf{I} - \mathbf{u} \mathbf{u}^\top) \Sigma (\mathbf{I} - \mathbf{u} \mathbf{u}^\top)^\top = \Sigma - 2 \Sigma \mathbf{u} \mathbf{u}^\top + \mathbf{u} \mathbf{u}^\top \Sigma \mathbf{u} \mathbf{u}^\top = \Sigma - \lambda \mathbf{u} \mathbf{u}^\top$

1. Find principal eigenvector of $(\Sigma - \lambda \mathbf{u} \mathbf{u}^\top)$
2. which is the second eigenvector of Σ
3. iterating to get d principal eigenvector of Σ

Power Method

Power iteration: $\mathbf{v}_{t+1} = \frac{\mathbf{A} \mathbf{v}_t}{\|\mathbf{A} \mathbf{v}_t\|}$, $\lim_{t \rightarrow \infty} \mathbf{v}_t = \mathbf{u}_1$

Assuming $\langle \mathbf{u}_1, \mathbf{v}_0 \rangle \neq 0$ and $|\lambda_1| > |\lambda_j| (\forall j \geq 2)$

2 Singular Value Decomposition

$\mathbf{A} = \mathbf{U} \mathbf{D} \mathbf{V}^\top = \sum_{k=1}^{\text{rank}(\mathbf{A})} d_{k,k} \mathbf{u}_k (\mathbf{v}_k)^\top$
 $\mathbf{A} \in \mathbb{R}^{N \times P}$, $\mathbf{U} \in \mathbb{R}^{N \times N}$, $\mathbf{D} \in \mathbb{R}^{N \times P}$, $\mathbf{V} \in \mathbb{R}^{P \times P}$
 $\mathbf{U}^\top \mathbf{U} = \mathbf{I} = \mathbf{V}^\top \mathbf{V}$ (\mathbf{U}, \mathbf{V} orthonormal)

\mathbf{U} columns are eigenvectors of $\mathbf{A} \mathbf{A}^\top$, \mathbf{V} columns are eigenvectors of $\mathbf{A}^\top \mathbf{A}$, \mathbf{D} diagonal elements are singular values.

$(\mathbf{D}^{-1})_{i,i} = \frac{1}{d_{i,i}}$ (don't forget to transpose)

1. calculate $\mathbf{A}^\top \mathbf{A}$.
2. calculate eigenvalues of $\mathbf{A}^\top \mathbf{A}$, the square root of them, in descending order, are the diagonal elements of \mathbf{D} .
3. calculate eigenvectors of $\mathbf{A}^\top \mathbf{A}$ using the eigenvalues resulting in the columns of \mathbf{V} .
4. calculate the missing matrix: $\mathbf{U} = \mathbf{A} \mathbf{V} \mathbf{D}^{-1}$.
5. normalize each column of \mathbf{U} and \mathbf{V} .

Low-Rank approximation

Using only K largest eigenvalues and corresponding eigenvectors. $\tilde{\mathbf{A}}_{i,j} = \sum_{k=1}^K \mathbf{U}_{i,k} \mathbf{D}_{k,k} \mathbf{V}_{j,k} = \sum_{k=1}^K \mathbf{U}_{i,k} \mathbf{D}_{k,k} (\mathbf{V}^\top)_{k,j}$.

Echart-Young Theorem

$\mathbf{A}_k = \arg \min_{\text{rank}(\mathbf{B})=k} \|\mathbf{A} - \mathbf{B}\|_F^2$ (not convex)

$\min_{\text{rank}(\mathbf{B})=K} \|\mathbf{A} - \mathbf{B}\|_F^2 = \|\mathbf{A} - \mathbf{A}_K\|_F^2 = \sum_{r=k+1}^{\text{rank}(\mathbf{A})} \sigma_r^2$

$\min_{\text{rank}(\mathbf{B})=K} \|\mathbf{A} - \mathbf{B}\|_2 = \|\mathbf{A} - \mathbf{A}_K\|_2 = \sigma_{k+1}$

3 Matrix Approximation & Reconstruction

$\min_{\text{rank}(\mathbf{B})=k} [\sum_{(i,j) \in I} (a_{ij} - b_{ij})^2]$, $I = \{(i, j) : ob.\}$

Alternating Least Squares

$f(\mathbf{U}, \mathbf{v}_i) = \sum_{(i,j) \in I} (a_{i,j} - \langle \mathbf{u}_j, \mathbf{v}_i \rangle)^2$

$f(\mathbf{u}_i, \mathbf{V}) = \sum_{(i,j) \in I} (a_{i,j} - \langle \mathbf{u}_j, \mathbf{v}_i \rangle)^2$

Convex when fixed one.

Convex Optimization

Def.: $\{(x, t) | x \in \text{dom } f, f(x) \leq t\}$, $f : \mathbb{R}^D \rightarrow \mathbb{R}$ is convex, if $\text{dom } f$ is a convex set, and if $\forall \mathbf{x}, \mathbf{y} \in \text{dom } f$, and $\forall \alpha \in [0, 1]$: $f(\alpha \mathbf{x} + (1 - \alpha) \mathbf{y}) \leq \alpha f(\mathbf{x}) + (1 - \alpha) f(\mathbf{y})$. Convex \iff Hessian p.s.d \iff local=global

Positive semi-definite: all principal minors (same-indexed rows and columns) ≥ 0

Positive definite: leading principal minors > 0

Convex Relaxation

Replace non-convex rank constraints by convex norm constraints (superset). Then project optimum back (hopefully still optimal).

$\min_{\mathbf{B} \in P_k} \|\mathbf{A} - \mathbf{B}\|_G^2, P_k = \{\mathbf{B} : \|\mathbf{B}\|_* \leq k\} \supseteq Q_k = \{\mathbf{B} : \text{rank}(\mathbf{B}) \leq k\}$ (in fact tightest convex lower-bound $\text{rank}(\mathbf{B}) \geq \|\mathbf{B}\|_*, \text{ for } \|\mathbf{B}\|_2 \leq 1$)

SVD Thresholding

$\mathbf{B}^* = \text{shrink}_\tau(\mathbf{A}) = \arg \min_{\mathbf{B}} \{\|\mathbf{A} - \mathbf{B}\|_F^2 + \tau \|\mathbf{B}\|_*\}$

Then with SVD $\mathbf{A} = \mathbf{U} \mathbf{D} \mathbf{V}^\top$, $\mathbf{D} = \text{diag}(\sigma_i)$, holds $\mathbf{B}^* = \mathbf{U} \mathbf{D}_\tau \mathbf{V}^\top$, $\mathbf{D}_\tau = \text{diag}(\max\{0, \sigma_i - \tau\})$

Iteration: $\mathbf{B}_{t+1} = \mathbf{B}_t + \eta_t \Pi(\mathbf{A} - \text{shrink}_\tau(\mathbf{B}_t))$

4 Non-Negative Matrix Factorization

$\mathbf{X} \in \mathbb{Z}_{\geq 0}^{N \times M}$, NMF: $\mathbf{X} \approx \mathbf{U}^\top \mathbf{V}$, $x_{ij} = \sum_z u_{zi} v_{zj} = \langle \mathbf{u}_i \mathbf{v}_j \rangle$ Decompose object into features: topics, face parts, etc.. \mathbf{u} weights on parts, \mathbf{v} parts (bases). More interpretable (PCA: holistic repre.).

EM for MLE for pLSA (No global opt guarantee)

Context Model: $p(w|d) = \sum_{z=1}^K p(w|z)p(z|d)$

Conditional independence assumption (*):

$p(w|d) = \sum_z p(w, z|d) = \sum_z p(w|d, z)p(z|d) = \sum_z p(w|z)p(z|d)$

Symmetric parameterization:

$p(w, d) = \sum_z p(z)p(w|z)p(d|z)$

Log-Likelihood: $L(\mathbf{U}, \mathbf{V}) = \sum_{i,j} x_{i,j} \log p(w_j | d_i)$

$= \sum_{(i,j) \in X} \log \sum_{z=1}^K p(w_j | z) p(z | d_i)$

$p(w_j | z) = v_{zj}$, $p(z | d_i) = u_{zi}$, $\sum_j v_{zj} = \sum_z u_{zi} = 1$

E-Step (optimal q: posterior of z over (d_i, w_j)):

$q_{zij} = \frac{p(w_j | z) p(z | d_i)}{\sum_{k=1}^K p(w_j | k) p(k | d_i)} := \frac{v_{zj} u_{zi}}{\sum_{k=1}^K v_{kj} u_{ki}}$, $\sum_z q_{zij} = 1$

M-Steps:

$p(z | d_i) = \frac{\sum_j x_{ij} q_{zij}}{\sum_j x_{ij}}$, $p(w_j | z) = \frac{\sum_i x_{ij} q_{zij}}{\sum_{i,l} x_{il} q_{zil}}$

Latent Dirichlet Allocation

To sample a new document, we need to extend \mathbf{X} and \mathbf{U}^\top with a new row, s.t. $\mathbf{X} = \mathbf{U}^\top \mathbf{V}$. (While pLSA fixes both dimensions)

For each d_i sample topic weights

$\mathbf{u}_i \sim \text{Dirichlet}(\boldsymbol{\alpha})$: $p(u_i | \boldsymbol{\alpha}) = \prod_{z=1}^K u_{zi}^{\alpha_z - 1}$, then

topic $z^t \sim \text{Multi}(u_i)$, word $w^t \sim \text{Multi}(v_{z^t})$

Multinomial obsv. model on wc vec: $p(\mathbf{x} | \mathbf{V}, \mathbf{u}) =$

$\frac{l!}{\prod_j x_j!} \prod_j \pi_j^{x_j}$ where $\pi_j = \sum_z v_{zj} u_{zi}$, $l = \sum_j x_j$

Bayesian averaging over \mathbf{u} : $p(\mathbf{x} | \mathbf{V}, \boldsymbol{\alpha}) =$

$\int p(\mathbf{x} | \mathbf{V}, \mathbf{u}) p(\mathbf{u} | \boldsymbol{\alpha}) d\mathbf{u}$

NMF Algorithm for quadratic cost function

$\min_{\mathbf{U}, \mathbf{V}} J(\mathbf{U}, \mathbf{V}) = \frac{1}{2} \|\mathbf{X} - \mathbf{U}^\top \mathbf{V}\|_F^2$ (non-negativity)

s.t. $\forall i, j, z : u_{zi}, v_{zj} \geq 0$

Comparison with pLSA:

1. sampling model: Gaussian vs multinomial 2. objective: quadratic vs KL divergence 3. constraints: not normalized

Alternating least squares:

1. init: $\mathbf{U}, \mathbf{V} = \text{rand}()$
2. repeat 3~4 for maxIters :
3. upd. $(\mathbf{V} \mathbf{V}^\top) \mathbf{U} = \mathbf{V} \mathbf{X}^\top$, proj. $u_{zi} = \max\{0, u_{zi}\}$
4. update $(\mathbf{U} \mathbf{U}^\top) \mathbf{V} = \mathbf{U} \mathbf{X}$, proj. $v_{zj} = \max\{0, v_{zj}\}$

5 Word Embeddings

Distributional Model:

$p_\theta(w|w') = \Pr[w \text{ occurs in context of } w']$

Log-likelihood:

$L(\theta; \mathbf{w}) = \sum_{t=1}^T \sum_{\Delta \in I} \log p_\theta(w^{(t+\Delta)} | w^{(t)})$

Latent Vector Model: $w \rightarrow (\mathbf{x}_w, b_w) \in \mathbb{R}^{D+1}$

$p_\theta(w|w') = \frac{\exp(\langle \mathbf{x}_w, \mathbf{x}_{w'} \rangle + b_w)}{\sum_{v \in V} \exp(\langle \mathbf{x}_v, \mathbf{x}_{w'} \rangle + b_v)}$ (soft-max).

Modifications:

$\log p_\theta(w|w') = \langle y_w, \mathbf{x}_{w'} \rangle + b_w$, word y_w , c'txt \mathbf{x}_w , use GloVe objective

negative sampling (logistic classification)

GoVe (Weighted Square Loss)

Co-occurrence Matrix:

$\mathbf{N} = (n_{ij}) \in \mathbb{R}^{V \times |C|} = \# \text{ of word } w_i \text{ in context } w_j$

Objective: $H(\theta; \mathbf{N})$

$= \sum_{n_{ij} > 0} f(n_{ij}) (\log n_{ij} - \log \exp[\langle \mathbf{x}_i, \mathbf{y}_j \rangle + b_i + d_j])^2$
with $f(n) = \min\{1, (\frac{n}{n_{\max}})^\alpha\}$, $\alpha \in (0; 1]$.

unnormalized distr. \rightarrow 2-sided loss function

1. sample $(i, j) u.a.r. s.t. n_{ij} > 0$

2. $\mathbf{x}_i^{new} \leftarrow \mathbf{x}_i + 2\eta f(n_{ij})(\log n_{ij} - \langle \mathbf{x}_i, \mathbf{y}_j \rangle) \mathbf{y}_j$

3. $\mathbf{y}_j^{new} \leftarrow \mathbf{y}_j + 2\eta f(n_{ij})(\log n_{ij} - \langle \mathbf{x}_i, \mathbf{y}_j \rangle) \mathbf{x}_i$

Discussion

Word embeddings can model analogies and relatedness, but antonyms are usually not well captured.

6 Data Clustering & Mixture Models

KMeans

Target: $\min_{\mathbf{U}, \mathbf{Z}} J(\mathbf{U}, \mathbf{Z}) = \|\mathbf{X} - \mathbf{UZ}\|_F^2$

$= \sum_{n=1}^N \sum_{k=1}^K z_{k,n} \|\mathbf{x}_n - \mathbf{u}_k\|_2^2$

1. **Initiate:** choose K centroids $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_K]$

2. **Cluster Assign:** data points to clusters.

$k^*(\mathbf{x}_n) = \arg \min_k \{\|\mathbf{x}_n - \mathbf{u}_k\|_2\}$ returns cluster k^* , whose centroid \mathbf{u}_{k^*} is closest to data point \mathbf{x}_n .

Set $\mathbf{z}_{k^*,n} = 1$, and for $l \neq k^*$ $\mathbf{z}_{l,n} = 0$.

3. **Update centroids:** $\mathbf{u}_k = \frac{\sum_{n=1}^N z_{k,n} \mathbf{x}_n}{\sum_{n=1}^N z_{k,n}}$.

4. Repeat from step 2, stops if $\|\mathbf{Z} - \mathbf{Z}^{new}\|_0 = \|\mathbf{Z} - \mathbf{Z}^{new}\|_F^2 = 0$.

Computational cost: $O(k \cdot n \cdot d)$

Gaussian Mixture Models (GMM)

Gaussian $p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$ Multivariate

$p(\mathbf{x}; \mu; \Sigma) = \frac{1}{|\Sigma|^{\frac{1}{2}} (2\pi)^{\frac{D}{2}}} \exp[-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)]$

For GMM let $\theta_k = (\mu_k, \Sigma_k)$; $p_{\theta_k}(\mathbf{x}) = \mathcal{N}(\mathbf{x} | \mu_k, \Sigma_k)$

Mixture Models: $p_{\theta}(\mathbf{x}) = \sum_{k=1}^K \pi_k p_{\theta_k}(\mathbf{x})$

Assignment variable (generative model):

$z_{ij} \in \{0, 1\}$, $\sum_{j=1}^k z_{ij} = 1$

$\Pr(z_k = 1) = \pi_k \Leftrightarrow p(\mathbf{z}) = \prod_{k=1}^K \pi_k^{z_k}$

Complete data distribution:

$p_{\theta}(\mathbf{x}, \mathbf{z}) = \prod_{k=1}^K (\pi_k p_{\theta_k}(\mathbf{x}))^{z_k}$

Posterior Probabilities:

$\Pr(z_k = 1 | \mathbf{x}) = \frac{\Pr(z_k=1)p(\mathbf{x}|z_k=1)}{\sum_{i=1}^K \Pr(z_i=1)p(\mathbf{x}|z_i=1)} = \frac{\pi_k p_{\theta_k}(\mathbf{x})}{\sum_{i=1}^K \pi_i p_{\theta_i}(\mathbf{x})}$

posterior $p(A|B) = \frac{\text{prior } p(A) \times \text{likelihood } p(B|A)}{\text{evidence } p(B)}$

Likelihood of observed data X:

$p_{\theta}(\mathbf{X}) = \prod_{n=1}^N p_{\theta}(\mathbf{x}_n) = \prod_{n=1}^N (\sum_{k=1}^K \pi_k p_{\theta_k}(\mathbf{x}_n))$

Max. Likelihood Estimation (MLE):

$\arg \max_{\theta} \sum_{n=1}^N \log(\sum_{k=1}^K \pi_k p_{\theta_k}(\mathbf{x}_n))$

$\geq \sum_{n=1}^N \sum_{k=1}^K q_k [\log p_{\theta_k}(\mathbf{x}_n) + \log \pi_k - \log q_k]$
with $\sum_{k=1}^K q_k = 1$ by Jensen Inequality.

Generative Model

1. sample cluster index $j \sim \text{Categorical}(\pi)$

2. given j , sample data $\mathbf{x} \sim \text{Normal}(\mu_j, \Sigma_j)$

Expectation-Maximization (EM) for GMM

E-Step: $\Pr[z_{k,n} = 1 | \mathbf{x}_n] = q_{k,n} =$

$$\frac{\pi_k^{(t-1)} \mathcal{N}(\mathbf{x}_n | \mu_k^{(t-1)}, \Sigma_k^{(t-1)})}{\sum_{j=1}^K \pi_j^{(t-1)} \mathcal{N}(\mathbf{x}_n | \mu_j^{(t-1)}, \Sigma_j^{(t-1)})}$$

M-Step: $\mu_k^{(t)} := \frac{\sum_{n=1}^N q_{k,n} \mathbf{x}_n}{\sum_{n=1}^N q_{k,n}}$, $\pi_k^{(t)} := \frac{1}{N} \sum_{n=1}^N q_{k,n}$

$$\Sigma_k^{(t)} = \frac{\sum_{n=1}^N q_{k,n} (\mathbf{x}_n - \mu_k^{(t)}) (\mathbf{x}_n - \mu_k^{(t)})^T}{\sum_{n=1}^N q_{k,n}}$$

Discussion K-means vs. EM

hard assignment vs soft. spherical clusters shapes vs covariance matrix. fast vs slow and more iteration. K-means can be used as initialization for EM.

K-means as a special case of GMM with covariances $\Sigma_j = \sigma^2 I$. in the limit of $\sigma \rightarrow 0$, recover K-means (hard assignments).

Model Order Selection (AIC / BIC for GMM)

Trade-off between data fit (i.e. likelihood $p(\mathbf{X}|\theta)$) and complexity (i.e. # of free parameters $\kappa(\cdot)$). For choosing K :

Akaike Information Criterion: $\text{AIC}(\theta|\mathbf{X}) = -\log p_{\theta}(\mathbf{X}) + \kappa(\theta)$

Bayesian Information Criterion: $\text{BIC}(\theta|\mathbf{X}) = -\log p_{\theta}(\mathbf{X}) + \frac{1}{2} \kappa(\theta) \log N$

of free params, fixed covariance matrix:

$\kappa(\theta) = K \cdot D + (K - 1)$ (K : # clusters, D : dim(data) = dim(μ_i), $K - 1$: π of # free clusters),

full covariance matrix: $\kappa(\theta) = K(D + \frac{D(D+1)}{2}) + (K - 1)$.

Compare AIC/BIC for different K – the smaller the better. BIC penalizes complexity more.

7 Sparse Coding

Orthogonal Basis

Pros: fast inverse; preserves energy. For \mathbf{x} and orthog. mat. \mathbf{U} compute $\mathbf{z} = \mathbf{U}^T \mathbf{x}$. Approx $\hat{\mathbf{x}} = \mathbf{U} \hat{\mathbf{z}}$, $\hat{z}_i = z_i$ if $|z_i| > \epsilon$ else 0. Reconstruction Error $\|\mathbf{x} - \hat{\mathbf{x}}\|^2 = \sum_{d \notin \sigma} \langle \mathbf{x}, \mathbf{u}_d \rangle^2$. Choice of base depends on signal. Fourier for global, wavelet for local support. PCA basis optimal for given Σ . Stripes & check patterns: hi-freq in Fourier.

Haar Wavelets (form orthogonal basis)

scaling fcn $\phi(x) = [1, 1, 1, 1]$, mother $W(x) = [1, 1, -1, -1]$, dilated $W(2x) = [1, -1, 0, 0]$, translated $W(2x - 1) = [0, 0, 1, -1]$

Overcomplete Basis

$\mathbf{U} \in \mathbb{R}^{D \times L}$ for # atoms = $L > D = \text{dim}(\text{data})$.

Decoding involved \rightarrow add constraint $\mathbf{z}^* \in$

$\arg \min_{\mathbf{z}} \|\mathbf{x} - \mathbf{Uz}\|_0$ s.t. $\mathbf{x} = \mathbf{Uz}$. NP-hard \rightarrow approximate with 1-norm (convex) or with MP.

Coherence • $m(\mathbf{U}) = \max_{i,j:i \neq j} |\mathbf{u}_i^T \mathbf{u}_j|$ • $m(\mathbf{B}) = 0$ if \mathbf{B} orthogonal matrix • $m([\mathbf{B}, \mathbf{u}]) \geq \frac{1}{\sqrt{D}}$ if atom \mathbf{u} is added to orthogonal basis \mathbf{B} (o.n.b. = orthonormal base)

Matching Pursuit (MP) approximation of \mathbf{x} onto \mathbf{U} , using K entries. Objective: $\mathbf{z}^* \in \arg \min_{\mathbf{z}} \|\mathbf{x} - \mathbf{Uz}\|_2$, s.t. $\|\mathbf{z}\|_0 \leq K$ 1. init: $\mathbf{z} \leftarrow 0, r \leftarrow \mathbf{x}$ 2. while $\|\mathbf{z}\|_0 < K$ do 3. select atom with smallest angle $i^* = \arg \max_i |\langle \mathbf{u}_i, \mathbf{r} \rangle|$ 4. update coefficients: $z_{i^*} \leftarrow z_{i^*} + \langle \mathbf{u}_{i^*}, \mathbf{r} \rangle$ 5. update residual: $\mathbf{r} \leftarrow \mathbf{r} - \langle \mathbf{u}_{i^*}, \mathbf{r} \rangle \mathbf{u}_{i^*}$.

Exact recovery when: $K < 1/2(1 + 1/m(\mathbf{U}))$

Compressive Sensing: Compress data while gathering: • $\mathbf{x} \in \mathbb{R}^D$, K -sparse in o.n.b. \mathbf{U} . $\mathbf{y} \in \mathbb{R}^M$ with $y_i = \langle \mathbf{w}_i, \mathbf{x} \rangle$: M lin. combinations of signal; $\mathbf{y} = \mathbf{Wx} = \mathbf{WUz} = \mathbf{\Theta z}$, $\mathbf{\Theta} \in \mathbb{R}^{M \times D}$ • Reconstruct $\mathbf{x} \in \mathbb{R}^D$ from \mathbf{y} ; find $\mathbf{z}^* \in \arg \min_{\mathbf{z}} \|\mathbf{z}\|_0$, s.t. $\mathbf{y} = \mathbf{\Theta z}$ (e.g. with MP, or convex it with 1-norm: can be eq!). Given \mathbf{z} , reconstruct $\mathbf{x} = \mathbf{Uz}$ Any orthogonal \mathbf{U} sufficient if: • $\mathbf{W} =$ Gaussian random projection, i.e. $w_{ij} \sim \mathcal{N}(0, \frac{1}{D})$ • $M \geq cK \log(\frac{D}{K})$, where c is some constant

8 Dictionary Learning

Adapt the dictionary to signal characteristics. Objective: $(\mathbf{U}^*, \mathbf{Z}^*) \in \arg \min_{\mathbf{U}, \mathbf{Z}} \|\mathbf{X} - \mathbf{UZ}\|_F^2$ not jointly convex but convex in 1 argument.

Matrix Factorization by Iter Greedy Minimization 1. Coding step: $\mathbf{Z}^{t+1} \in \arg \min_{\mathbf{Z}} \|\mathbf{X} - \mathbf{U}^t \mathbf{Z}\|_F^2$ subject to \mathbf{Z} being sparse ($\mathbf{z}_i^{t+1} \in \arg \min_{\mathbf{z}} \|\mathbf{z}\|_0$ s.t. $\|\mathbf{x}_n - \mathbf{U}^t \mathbf{z}\|_2 \leq \sigma \|\mathbf{x}_n\|_2$)

2. Dict update step: $\mathbf{U}^{t+1} \in \arg \min_{\mathbf{U}} \|\mathbf{X} - \mathbf{UZ}^{t+1}\|_F^2$, subj to $\forall l \in [L] : \|\mathbf{u}_l\|_2 = 1$. (set $\mathbf{U} = [\mathbf{u}_1^t \dots \mathbf{u}_l^t \dots \mathbf{u}_L^t]$, $\min_{\mathbf{u}_l} \|\mathbf{X} - \mathbf{UZ}^{t+1}\|_F^2 = \min_{\mathbf{u}_l} \|\mathbf{R}_l^t - \mathbf{u}_l (\mathbf{z}_l^{t+1})^T\|_F^2$ with $\mathbf{R}_l^t = \tilde{\mathbf{U}} \Sigma \tilde{\mathbf{V}}^T$ by $\mathbf{u}_l^* = \tilde{\mathbf{u}}_1$)

9 Neural Networks

Activation: $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ sigmoid $s(x) = \frac{1}{1 + e^{-x}}$, $s'(x) = s(x)(1 - s(x))$, ReLU $\max(0, x)$

Neurons: $F_{\sigma}(\mathbf{x}; \mathbf{w}) = \sigma(w_0 + \sum_{i=1}^M x_i w_i)$.

Output: linear regression $\mathbf{y} = \mathbf{W}^L \mathbf{x}^{L-1}$, binary (logistic) $y_1 = P[Y = 1 | \mathbf{x}] = \frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{x}^{L-1})}$, multiclass (soft-max) $y_k = P[Y = k | \mathbf{x}] =$

$\frac{\exp(\mathbf{w}_k^T \mathbf{x}^{L-1})}{\sum_{m=1}^K \exp(\mathbf{w}_m^T \mathbf{x}^{L-1})}$. **Loss function** $l(y, \hat{y})$: squared loss $\frac{1}{2}(y - \hat{y})^2$, cross-entropy loss $-y \log \hat{y} - (1 - y) \log(1 - \hat{y})$. **Units and Layers:** layer-to-layer fwd. prop. notati-

on: $\mathbf{x}^l = \sigma^l(\mathbf{W}^{(l)} \mathbf{x}^{(l-1)})$. L-layer network:

$$\mathbf{y} = \sigma^{(L)}(\mathbf{W}^{(L)} \sigma^{(L-1)}(\dots(\sigma^{(1)}(\mathbf{W}^{(1)} \mathbf{x}) \dots)))$$

Backpropagation

Layer-to-layer Jacobian: $\mathbf{x} =$ prev. layer activation, $\mathbf{x}^+ =$ next layer activation. Jacobian matrix $\mathbf{J} = J_{ij}$ of mapping $\mathbf{x} \rightarrow \mathbf{x}^+$, $\mathbf{x}_i^+ = \sigma(\mathbf{w}_i^T \mathbf{x})$,

$J_{ij} = \frac{\partial \mathbf{x}_i^+}{\partial \mathbf{x}_j} = w_{ij} \cdot \sigma'(\mathbf{w}_i^T \mathbf{x})$. Across multiple layers:

$\frac{\partial \mathbf{x}^{(l)}}{\partial \mathbf{x}^{(l-n)}} = \mathbf{J}^{(l)} \cdot \frac{\partial \mathbf{x}^{(l-1)}}{\partial \mathbf{x}^{(l-n)}} = \mathbf{J}^{(l)} \cdot \mathbf{J}^{(l-1)} \dots \mathbf{J}^{(l-n+1)}$ and then back prop. $\nabla_{\mathbf{x}^{(l)}} \ell = \nabla_{\mathbf{y}^T} \ell \cdot \mathbf{J}^{(L)} \dots \mathbf{J}^{(l+1)}$

Weights: $\frac{\partial \ell}{\partial w_{ij}^{(l)}} = \frac{\partial \ell}{\partial x_i^{(l)}} \frac{\partial x_i^{(l)}}{\partial w_{ij}^{(l)}}$, $\frac{\partial x_i^l}{\partial w_{ij}^l} =$

$\sigma'([\mathbf{w}_i^{(l)}]^T \mathbf{x}^{(l-1)}) \cdot x_j^{(l-1)}$ (sensitivity of downstream unit · activation of up-stream unit)

Gradient Descent (or Deepest Descent)

Gradient: $\nabla f(\mathbf{x}) := (\frac{\partial f(\mathbf{x})}{\partial x_1}, \dots, \frac{\partial f(\mathbf{x})}{\partial x_D})^T$

1. init: $\mathbf{x}^{(0)} \in \mathbb{R}^D$

2. for $t = 0$ to maxIter :

3. $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \gamma \nabla f(\mathbf{x}^{(t)})$, usually $\gamma \approx \frac{1}{t}$

Stochastic Gradient Descent (SGD)

Assume **Additive Objective:**

$f(\mathbf{x}) = \frac{1}{N} \sum_{n=1}^N f_n(\mathbf{x})$

1. init: $\mathbf{x}^{(0)} \in \mathbb{R}^D$

2. for $t = 0$ to maxIter :

3. sample $n \in_{u.a.r.} \{1, \dots, N\}$

4. $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \gamma \nabla f_n(\mathbf{x}^{(t)})$, typically $\gamma \approx \frac{1}{t}$.

Neural Networks for Images (CNN)

Translation invariance of images \rightarrow neurons compute same fct, shift invariant filters; weights defined as filter masks, e.g. convolution: $F_{n,m}(\mathbf{x}; \mathbf{w}) = \sigma(b + \sum_{k=-2}^2 \sum_{l=-2}^2 w_{k,l} x_{n+k,m+l})$. To reduce dimension of convolution, use {max, avg}-pooling

10 Deep Unsupervised Learning

Autoregressive

Image $p(\mathbf{x}) = \prod_i^{n^2} p(x_i | x_1, \dots, x_{i-1})$

Variational Autoencoder

$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} = \mathbb{E}_i [\frac{\log P_i}{\log Q_i}]$ (0:similar)

Elbo $\mathbb{E}_{\mathbf{x} \sim P_X} [\mathbb{E}_{\mathbf{z} \sim Q} \log P_g(\mathbf{x} | \mathbf{z}) - D^{KL}(Q(\mathbf{z} | \mathbf{x}) || P(\mathbf{z}))]$
 Q enc. posterior distr., $P(\mathbf{z})$ prior distr. on latent var \mathbf{z} , P_g likelihood of dec. generated \mathbf{x}

Jointly trained: enc. optimize regularizer term, sample $\mathbf{z} \sim Q$, feed to dec., produce $\hat{\mathbf{x}}$ to max. reconstruction quality. Both terms diff'able, can use SGD to train end-to-end.