Essentials

Matrix/Vector

Multiplication: $\mathbf{C}=\mathbf{AB} \Leftrightarrow c_{ik}=\sum_{j=1}^{m} a_{ij}\cdot b_{jk}$

Orthogonal Matrix: (full rank square matrix with orthonormal columns) $\mathbf{A}^{-1}=\mathbf{A}^\top$, $\mathbf{A}\mathbf{A}^\top=\mathbf{A}^\top\mathbf{A}=\mathbf{I}$, $\det(\mathbf{A})\in\{+1,-1\}$, $\det(\mathbf{A}^\top\mathbf{A})=1$, preserves: inner product, norm, distance, angle, rank, mat. orthogon.

Inner Product: $\langle \mathbf{x},\mathbf{y}\rangle=\mathbf{x}^\top\mathbf{y}=\sum_{i=1}^{N}\mathbf{x}_i\mathbf{y}_i$.
$\langle \mathbf{x}\pm\mathbf{y},\mathbf{x}\pm\mathbf{y}\rangle=\langle \mathbf{x},\mathbf{x}\rangle\pm 2\langle \mathbf{x},\mathbf{y}\rangle+\langle \mathbf{y},\mathbf{y}\rangle$
$\langle \mathbf{x},\mathbf{y}+\mathbf{z}\rangle=\langle \mathbf{x},\mathbf{y}\rangle+\langle \mathbf{x},\mathbf{z}\rangle$
$\langle \mathbf{x},\mathbf{y}\rangle=\|\mathbf{x}\|_2\cdot\|\mathbf{y}\|_2\cdot\cos(\theta)$
If $\mathbf{y}$ is a unit vector then $\langle \mathbf{x},\mathbf{y}\rangle$ projects $\mathbf{x}$ onto $\mathbf{y}$
$(\mathbf{u}_i^T\mathbf{v}_j)\mathbf{v}_j=(\mathbf{v}_j\mathbf{v}_j^T)\mathbf{u}_i$

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

Transpose: $(\mathbf{A}^\top)^{-1}=(\mathbf{A}^{-1})^\top$

Determinant: $|\mathbf{A}|=\sum_i\lambda_i$, $\quad |\mathbf{A}^{-1}|=1/|\mathbf{A}|$

Norms

$\|\mathbf{x}\|_0=|\{i|x_i\neq 0\}|\qquad \|\mathbf{x}\|_2=\sqrt{\sum_{i=1}^{N}\mathbf{x}_i^2}=\sqrt{\langle\mathbf{x},\mathbf{x}\rangle}$
$\|\mathbf{x}\|_p=\left(\sum_{i=1}^{N}|x_i|^p\right)^{\frac{1}{p}}\qquad \|\mathbf{M}\|_F=\sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbf{m}_{i,j}^2}$
$=\sqrt{\sum_{i=1}^{\min\{m,n\}}\sigma_i^2}=\|\sigma(\mathbf{A})\|_2=\sqrt{trace(\mathbf{M}^T\mathbf{M})}$
$\|\mathbf{M}\|_G=\sqrt{\sum_{ij}g_{ij}x_{ij}^2}$ (weighted Frobenius)
$\|\mathbf{M}\|_1=\sum_{i,j}|m_{i,j}|\qquad \|\mathbf{M}\|_p=\max_{\mathbf{v}\neq 0}\frac{\|\mathbf{Mv}\|_p}{\|\mathbf{v}\|_p}$
$\|\mathbf{M}\|_2=\sigma_{\max}(\mathbf{M})=\|\sigma((M))\|_\infty$
$\|\mathbf{M}\|_*=\sum_{i=1}^{\min(m,n)}\sigma_i=\|\sigma(\mathbf{A})\|_1$ (nuclear norm)
$\mathrm{rank}(\mathbf{B})\geq\|\mathbf{B}\|_*$ for $\|B\|_2\leq 1$

Derivatives

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

Eigenvalues & Eigenvectors

Eigenvalue problem: $\mathbf{Ax}=\lambda\mathbf{x}$
1. solve $\det(\mathbf{A}-\lambda\mathbf{I})=0$ resulting in $\{\lambda_i\}$
2. $\forall\lambda_i$ solve $(\mathbf{A}-\lambda_i\mathbf{I})x_i=\mathbf{0}$ for $\mathbf{x}_i$

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 $(\mathbf{\Lambda}^{-1})_{i,i}=1/\lambda_i$
if $\mathbf{A}$ symmetric: $A=\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$ ($\mathbf{Q}$ orthogonal)

Probability / Statistics

$P(x)=\sum_{y\in Y}P(x,y)\qquad P(x,y)=P(x|y)P(y)$
$\forall y\in Y:\sum_{x\in X}P(x|y)=1$ (property for any fixed y)
$P(x|y)=\frac{P(x,y)}{P(y)}$, $\quad$ if $P(y)>0\qquad P(x|y)=\frac{P(y|x)P(x)}{P(y)}$
(Bayes' rule) $\quad P(x_1,\ldots,x_n)=\prod_{i=1}^{n}P(x_i)$ (iff i.i.d)
$P(x|y)=P(x)\Leftrightarrow P(y|x)=P(y)$ (iff X, Y indep.)
$E[X]:=\sum_{x\in X}x\cdot P(x)\qquad Var[X]:=E[(X-\mu_x)^2]:=\sum_{x\in X}(x-\mu_x)^2P(x)=E(X^2)-E(X)^2$
standard deviation $\sigma_x:=\sqrt{Var[X]}$

Lagrangian Multipliers

Problem $\min_Q g(Q)$ with constraint $\forall j\ \sum_i Q_{ij}=1$ turn into $L(Q,\alpha)=g(Q)+\sum_j\alpha_j(1-\sum_i Q_{ij})$ and find $\max_\alpha\min_Q L(Q,\alpha)$ (can use constraint form.).

Convex Function

$\forall x_1,x_2\in X,\forall t\in[0,1]:\quad$ (also iff $\forall x:\ f''(x)\geq 0$)
$f(tx_1+(1-t)x_2)\leq tf(x_1)+(1-t)f(x_2)$
Sum of convex functions is convex, **log** is convex
Exercise: Show that if f is convex, any local optimum is global. Assume there is a local optimum $\hat{x}$ that is not the global optimum x^*, then if we choose t to be in the ball of the local optimum, so we know that $f(t\hat{x}+(1-t)x^*)\geq f(\hat{x})$. Since $f(x^*)<f(\hat{x})$, we have $t\cdot f(\hat{x})+(1-t)f(x^*)<f(\hat{x})$. So we get $f(t\hat{x}+(1-t)x^*)\geq f(\hat{x})>t\cdot f(\hat{x})+(1-t)f(x^*)$, which contradicts the convexity of f.

Jensen Inequality:

for convex ϕ: $\phi(\sum_{i=1}^{n}\lambda_i x_i)\leq\sum_{i=1}^{n}\lambda_i f(x_i)$ if $\sum_{i=1}^{n}\lambda_i=1$. Also $\phi(E[X])\leq E[\phi(X)]$.

Singular Value Decomposition

$\mathbf{A}=\mathbf{UDV}^\top=\sum_{k=1}^{\mathrm{rank}(\mathbf{A})}d_{k,k}u_k(v_k)^\top$
$\mathbf{A}\in\mathbb{R}^{N\times P},\mathbf{U}\in\mathbb{R}^{N\times N},\mathbf{D}\in\mathbb{R}^{N\times P},\mathbf{V}\in\mathbb{R}^{P\times P}$
$\mathbf{U}^\top\mathbf{U}=I=\mathbf{V}^\top\mathbf{V}$ ($\mathbf{U},\mathbf{V}$ orthonormal)
$\mathbf{U}$: cols are eigenvectors of $\mathbf{A}\mathbf{A}^\top$, $\mathbf{V}$: cols are eigenvectors of $\mathbf{A}^\top\mathbf{A}$, $\mathbf{D}$ diag. el. are singular values.
1. calculate $\mathbf{A}^\top\mathbf{A}$.
2. calculate eigenvalues of $\mathbf{A}^\top\mathbf{A}$, the square root of them, in desc. 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: $\mathbf{U}=\mathbf{AVD}^{-1}$.
5. normalize each column of $\mathbf{U}$ and $\mathbf{V}$.
Complexity: $O(\min(mn^2,nm^3))$

Eckart-Young Theorem

$\min_{rank(B)=K}\|A-B\|_F^2=\|A-A_k\|_F^2=\sum_{r=k+1}^{rank(A)}\sigma_r^2$
$\min_{rank(B)=K}\|A-B\|_2=\|A-A_k\|_2=\sigma_{k+1}$

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: $\overline{\mathbf{X}}=\mathbf{X}-[\bar{\mathbf{x}},\ldots,\bar{\mathbf{x}}]=\mathbf{X}-\mathbf{M}$.
3. Cov.: $\Sigma=\frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n-\bar{\mathbf{x}})(\mathbf{x}_n-\bar{\mathbf{x}})^\top=\frac{1}{N}\overline{\mathbf{X}}\overline{\mathbf{X}}^\top$.
4. Eigenvalue Decomposition: $\Sigma=\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$.
5. Select $K<D$, only keep $\mathbf{U}_K,\lambda_K$.
6. Transform data onto new Basis: $\overline{\mathbf{Z}}_K=\mathbf{U}_K^\top\overline{\mathbf{X}}$.
7. Reconstruct to original Basis: $\tilde{\overline{\mathbf{X}}}=\mathbf{U}_k\overline{\mathbf{Z}}_K$.
8. Reverse centering: $\tilde{\mathbf{X}}=\tilde{\overline{\mathbf{X}}}+\mathbf{M}$.
For compression save $\mathbf{U}_k,\overline{\mathbf{Z}}_K,\bar{\mathbf{x}}$.
$\mathbf{U}_k\in\mathbb{R}^{D\times K},\Sigma\in\mathbb{R}^{D\times D},\overline{\mathbf{Z}}_K\in\mathbb{R}^{K\times N},\overline{\mathbf{X}}\in\mathbb{R}^{D\times N}$

Reconstruction Error Exercise

$\tilde{\mathbf{X}}=\mathbf{U}_K\mathbf{U}_K^\top\overline{\mathbf{X}}$, the error is $\frac{1}{N}\sum_{i=1}^{N}\|\tilde{x}_i-\bar{x}_i\|_2^2$
$=\frac{1}{N}\|\tilde{\mathbf{X}}-\overline{\mathbf{X}}\|_F^2=\frac{1}{N}\|(\mathbf{U}_K\mathbf{U}_K^\top-\mathbf{I}_d)\overline{\mathbf{X}}\|_F^2$
$=\frac{1}{N}\mathrm{trace}((\mathbf{U}_K\mathbf{U}_K^\top-\mathbf{I}_d)\overline{\mathbf{X}}\overline{\mathbf{X}}^\top(\mathbf{U}_K\mathbf{U}_K^\top-\mathbf{I}_d)^\top)$
$=\mathrm{trace}((\mathbf{U}_K\mathbf{U}_K^\top-\mathbf{I}_d)\Sigma(\mathbf{U}_K\mathbf{U}_K^\top-\mathbf{I}_d))\quad\Sigma=\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$
$=\mathrm{trace}((\mathbf{U}_K\mathbf{U}_K^\top\mathbf{U}-\mathbf{U})\mathbf{\Lambda}(\mathbf{U}^\top\mathbf{U}_K\mathbf{U}_K^\top-\mathbf{U}^\top))$
$=\mathrm{trace}(([\mathbf{U}_K;\mathbf{0}]-\mathbf{U})\mathbf{\Lambda}([\mathbf{U}_K;\mathbf{0}]-\mathbf{U}^\top))$
$=\mathrm{trace}(\sum_{i=K+1}^{D}\lambda_i u_i u_i^\top)=\sum_{i=K+1}^{D}\lambda_i\cdot\mathrm{trace}(u_i u_i^\top)$
$=\sum_{i=K+1}^{D}\lambda_i$ since $\mathrm{trace}(u_i u_i^\top)=\|u_i\|_2^2=1$

Iterative View

Residual r_i: $x_i-\tilde{x}_i=I-uu^Tx_i$
Cov of r: $\frac{1}{n}\sum_{i=1}^{n}(I-uu^T)x_ix_i^T(I-uu^T)^T=(I-uu^T)\Sigma(I-uu^T)^T=\Sigma-2\Sigma uu^T+uu^T\Sigma uu^T=\Sigma-\lambda uu^T$
1. Find principal eigenvector of $(\Sigma-\lambda uu^T)$
2. Which is the second eigenvector of Σ
3. Iterating to get d principal eigenvector of Σ

Power Method

Power iteration: $v_{t+1}=\frac{Av_t}{\|Av_t\|}$, $\lim_{t\to\infty}v_t=u_1$
Assuming $\langle u_1,v_0\rangle\neq 0$ and $|\lambda_1|>|\lambda_j|(\forall j\geq 2)$ Then $\lambda_1=\lim_{t\to\infty}\|\mathbf{A}\mathbf{v}_t\|/\|\mathbf{v}_t\|$

Matrix Reconstruction

Alternating Least Squares

Beyond SVD: unobserved entries! $f(\mathbf{U},v_i)=\sum_{(i,j)\in I}(a_{i,j}-\langle\mathbf{u}_j,\mathbf{v}_i\rangle)^2$, Fix one, alternate other:
$\mathbf{U}\leftarrow\arg\min_{\mathbf{U}}f(\mathbf{U},\mathbf{V})$, $\mathbf{V}\leftarrow\arg\min_{\mathbf{V}}f(\mathbf{U},\mathbf{V})$
Can decompose (solve independently)
$f(\mathbf{U},v_i)=\sum_i\left[\sum_{(i,j)\in I}(a_{i,j}-\langle\mathbf{u}_j,\mathbf{v}_i\rangle)^2\right]$
Can add regularization $\mu(\|U\|_F^2+\|V\|_F^2)\quad\mu>0$
Upd: $u_i=\left(\sum_{(i,j)\in\mathscr{I}}v_jv_j^\top+I_k\lambda\right)^{-1}\left(\sum_{(i,j)\in\mathscr{I}}a_{ij}v_j\right)$

SVD Thresholding

$\mathbf{B}^*=shrink_\tau(\mathbf{A}):=\arg\min_{\mathbf{B}}\{\|\mathbf{A}-\mathbf{B}\|_F^2+\tau\|\mathbf{B}\|_*\}$
then with SVD holds $\mathbf{B}^*=\mathbf{U}\mathbf{D}_\tau\mathbf{V}^\mathbf{T}$, $\mathbf{D}_\tau=diag(\max\{0,\sigma_i-\tau\})$, $\Pi(\mathbf{X})=x_{ij}$ if $(i,j)\in\mathscr{I}$ el. 0
Iteration: $\mathbf{B}_{t+1}=\mathbf{B}_t+\eta_t\Pi(\mathbf{A}-shrink_\tau(\mathbf{B}_t))$

Non-Negative Matrix Factorization

Want to learn words w in a document d. Use topic latent variable: $\mathbf{X}\in\mathbb{Z}_{\geq 0}^{N\times M}$, NMF: $\mathbf{X}\approx\mathbf{U}^\top\mathbf{V},x_{ij}=\sum_z u_{zi}v_{zj}=\langle\mathbf{u}_i\mathbf{v}_j\rangle$ $\mathbf{U},\mathbf{V}$ are non-neg., L_1 col normal.

Probabilistic LSA

Context Model: $p(w|d)=\sum_{i=1}^{K}p(w|i)p(i|d)$
Conditional independence assumption $(*)$**:**
$p(w|d)=\sum_i p(w,i|d)=\sum_i p(w|d,i)p(i|d)\overset{*}{=}\sum_i p(w|i)p(i|d)\qquad$ Note $p(w|i):=p(w|z=i)$
Symmetric parameterization:
$p(w,d)=\sum_z p(z)p(w|z)p(d|z)$

EM for MLE for pLSA (NO global opt guarantee)

Log-Likelihood: $L(\mathbf{U},\mathbf{V})=\sum_{w,d}X_{w,d}\log p(w|d)$
$p(w|i)=u_{wi}$, $p(i|d)=v_{id}$, $\sum_w u_{wi}=\sum_i v_{di}=1$
$q_{iwd}\in\{0,1\}$: w in d generated via $z=i$.
Lower bound from Jensen: $\sum_{w,d}\log\sum_{z=1}^{K}q_{iwd}\frac{u_{wi}v_{id}}{q_{iwd}}\geq\sum_{w,d,i}q_{iwd}[\log u_{wi}+\log v_{id}-\log q_{iwd}]$
Don't forget to add the sum over i,j and X_{ij} again.
E-Step (optimal q: posterior $p(z=i|w,d)$):
$q_{iwd}=\frac{p(w|i)p(i|d)}{\sum_{k=1}^{K}p(w|k)p(k|d)}:=\frac{u_{wi}v_{id}}{\sum_{k=1}^{K}u_{wk}v_{kd}}$, $\sum_i q_{iwd}=1$
M-Steps:
$p(w|i)=u_{wi}=\frac{\sum_d q_{iwd}X_{wd}}{\sum_{w,d}q_{iwd}X_{wd}}$, $p(i|d)=v_{id}=\frac{\sum_w q_{iwd}X_{wd}}{\sum_w X_{wd}}$
Derivations: maximize lower bound w.r.t. $\sum_w u_{wi}=1$ and $\sum_i v_{id}=1$ (no $>=0$ constraint bc. log). $\min_{U,V}\max_{\alpha,\beta}\mathscr{L}$ where $\mathscr{L}=-g(X;U,V)+\sum_i\alpha_i(\sum_w u_{wi}-1)+\sum_d\beta_d(\sum_i v_{id}-1)$. Set $\partial\mathscr{L}/\partial u_{wi}=0$ and get $u_{wi}=\sum_d X_{wd}q_{iwd}/\alpha_i$. Setting $\partial\mathscr{L}/\partial\alpha_i$ gives $\sum_w u_{wi}=1\to\sum_{w,d}X_{wd}q_{iwd}/\alpha_i=1\to\alpha_i=1/(\sum_{w,d}X_{wd}q_{iwd})$ then plug it in. Similar for v_{id} but with extra step: $\sum_w X_{wd}\sum_i q_{iwd}/\beta_d=1\to\beta_d=1/(\sum_w X_{wd})$ since $\sum_i q_{iwd}=1$.

Latent Dirichlet Allocation

To sample new d, need to extend X and U^T (in pLSA matrix dims fixed). For each d_i sample topic weights $\mathbf{u}_i$~Dirichlet(α): $p(u_i|\alpha)=\prod_{z=1}^{K}u_{zi}^{\alpha_k-1}$, then topic z^t~Multi(u_i), word w^t~Multi(v_{z^t})
LDA Model: $p(\mathbf{x}|V,u)=\frac{l!}{\prod_j\mathbf{x}_j!}\prod_j\pi_j^{\mathbf{x}_j}$ where $\pi_j=\sum_z v_{zj}u_z$, $l=\sum_j x_j$ and $l=\sum_j x_j$
Bayesian averaging over $\mathbf{u}$:
$p(\mathbf{x}|\mathbf{V},\alpha)=\int p(\mathbf{x}|\mathbf{V},\mathbf{u})p(\mathbf{u}|\alpha)d\mathbf{u}$

NMF Algorithm for Quadratic Cost Function

$\min_{\mathbf{U},\mathbf{V}}J(\mathbf{U},\mathbf{V})=\frac{1}{2}\|\mathbf{X}-\mathbf{U}^\top\mathbf{V}\|_F^2$ s.t. $\forall i,j,z:\ u_{zi},v_{zj}\geq 0$ (non-negativity)
Comparison with pLSA: different sampling model (Gaussian not multinomial), different objective (quadratic not KL divergence), not normalized
ALS (not joint convex over (U,V)):
1. init: $\mathbf{U},\mathbf{V}=rand()$ 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. $v_{zj}=\max\{0,v_{zj}\}$

Word Embeddings

Distributional Model:
$p_\theta(w|w')$ = Pr[w 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: $w\to(\mathbf{x}_w,b_w)\in\mathbb{R}^{D+1}$
$p_\theta(w|w')=\frac{\exp[\langle\mathbf{x}_w,\mathbf{x}_{w'}\rangle+b_w]}{\sum_{v\in V}\exp[\langle\mathbf{x}_v,\mathbf{x}_{w'}\rangle+b_v]}$ (soft-max).
Skip Gram Model:
$\mathscr{L}(\theta;\mathbf{w})=\sum_t\sum_{\Delta\in\mathscr{I}}b_{w^{(t+\Delta)}}+\langle\mathbf{x}_{w^{(t+\Delta)}},\mathbf{x}_{w^{(t)}}\rangle-\log\sum_{v\in\mathscr{V}}\exp[\langle\mathbf{x}_v,\mathbf{x}_{w^{(t)}}\rangle+b_v]$
Modifications:
$\log p_\theta(w|w')=\langle y_w,x_{w'}\rangle+b_w$, word embedding y_w, context embeddings $x_{w'}$. Alternative to MLE (partition calculation is hard): negative sampling (modify objective into logistic classification), PMI

GloVe (Weighted Square Loss)

Co-occurence Matrix:
$\mathbf{N}=(n_{ij})\in\mathbb{N}^{|V|\times|C|}$ = #occ. of w_i in context w_j
Obj: $\mathcal{H}(\theta;\mathbf{N})=\sum_{i,j} f(n_{ij})(\log n_{ij}-\log\tilde{p}_\theta(w_i|w_j))^2$
Unnorm. dist: $\tilde{p}_\theta(w_i|w_j)=\exp[\langle \mathbf{x}_i,\mathbf{y}_j\rangle+b_i+c_j]$
with $f(n)=\min\{1,(\frac{n}{n_{max}})^\alpha\}$, $\alpha\in(0;1]$.
normalized: need to compute partition function, but cannot be large everywhere.
unnormalized: can use two-sided loss function
Perform SGD to find local minimum
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: can model analogies and relatedness, but antonyms are usually not well captured.

Data Clustering & Mixture Models

K-Means

$\mathbf{Z}\in\{0,1\}^{N\times K}$ (if point i assigned to cluster j)
Target: $\min_{\mathbf{U},\mathbf{Z}} J(\mathbf{U},\mathbf{Z})=\|\mathbf{X}-\mathbf{U}\mathbf{Z}\|_F^2$
$=\sum_{n=1}^N\sum_{k=1}^K \mathbf{z}_{k,n}\|\mathbf{x}_n-\mathbf{u}_k\|_2^2$
1. **Initiate:** choose K centroids $\mathbf{U}=[\mathbf{u}_1,\dots,\mathbf{u}_K]$
2. **Cluster Assign:** assign data points to closest cluster $z_{ij}^*=1$ if $j=\arg\min_k\|\mathbf{x}_i-\mathbf{u}_k\|^2$ else 0
3. **Update centroids**: $\mathbf{u}_k=\frac{\sum_{n=1}^N z_{k,n}\mathbf{x}_n}{\sum_{n=1}^N z_{k,n}}$. Repeat 2

Stop if $\|\mathbf{Z}-\mathbf{Z}_{\mathbf{new}}\|_F^2=0$. Guaranteed to converge to local optimum. Computational cost: $O(k\cdot n\cdot d)$

Gaussian Mixture Models (GMM)

$p(x;\mu;\Sigma)=\frac{1}{|\Sigma|^{\frac{1}{2}}(2\pi)^{\frac{D}{2}}}exp[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)]$
$\int\mathcal{N}(z;\mu,\Sigma)\log(\mathcal{N}(z;0,I))dz=E_z[\mathcal{N}(z;0,I)]$
$=-D/2\log(2\pi)-1/2\sum_{i=1}^D(\mu_i^2+\sigma_i^2)$
For GMM let $\theta_k=(\mu_k,\Sigma_k)$; $p_{\theta_k}(\mathbf{x})=\mathcal{N}(\mathbf{x}|\mu_k,\Sigma_k)$
Mixture Models: $p_\theta(\mathbf{x})=\sum_{k=1}^K\pi_k p_{\theta_k}(\mathbf{x})$
Generate: sample cluster $j\sim Categorical(\pi)$, sample data from j-th cluster: $x\sim\mathcal{N}(\mu_j,\Sigma_j)$
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:
$\Pr(z_k=1|\mathbf{x})=\frac{\Pr(z_k=1)p(\mathbf{x}|z_k=1)}{\sum_{l=1}^K\Pr(z_l=1)p(\mathbf{x}|z_l=1)}=\frac{\pi_k p_{\theta_k}(\mathbf{x})}{\sum_{l=1}^K\pi_l p_{\theta_l}(\mathbf{x})}$
posterior $p(A|B)=\frac{\text{prior } p(A)\,\times\,\text{likelihood } p(B|A)}{\text{evidence } p(B)}$
Likelihood of observed data X:
$p_\theta(\mathbf{X})=\prod_{n=1}^N p_\theta(\mathbf{x}_n)=\prod_{n=1}^N\left(\sum_{k=1}^K\pi_k p_{\theta_k}(\mathbf{x}_n)\right)$
Max. Likelihood Estimation (MLE):
$\arg\max_\theta\sum_{n=1}^N\log\left(\sum_{k=1}^K\pi_k p_{\theta_k}(\mathbf{x}_n)\right)$ (mult. 1)
$\geq\sum_{n=1}^N\sum_{k=1}^K q_{nk}[\log p_{\theta_k}(\mathbf{x}_n)+\log\pi_k-\log q_{nk}]$
with $\sum_{k=1}^K q_{nk}=1$ by Jensen Inequality.

Expectation-Maximization (EM) for GMM

E-Step: $Pr[z_j=1|\mathbf{x}_i]=q_{ij}=\frac{\pi_j p(\mathbf{x}_i;\theta_j)}{\sum_{l=1}^K\pi_l p(\mathbf{x}_i;\theta_l)}$
Derivation: Can maximize independent of i. Take derivative of lower bound w.r.t. q_{ij} with Lagrangian: $\lambda(\sum_j q_{ij}-1)$ to get: $\log\pi_j+\log p(x_i|\mu_j,\Sigma_j)-\log q_{ij}-1+\lambda=0\rightarrow q_{ij}=\pi_j p(x_i|\mu_j,\Sigma_j)e^{\lambda-1}$. Now use $\sum_j q_{ij}=\sum_j\pi_j p(x_i|\mu_j,\Sigma_j)e^{\lambda-1}=1\rightarrow e^{\lambda-1}=1/(\sum_j pi_j p(x_i|\mu_j,\Sigma_j))$ and plug in.
M-Step: $\mu_j^*:=\frac{\sum_{i=1}^N q_{ij}\mathbf{x}_i}{\sum_{i=1}^N q_{ij}}$, $\pi_j^*:=\frac{1}{N}\sum_{i=1}^N q_{ij}$
$\Sigma_k^*=\frac{\sum_{i=1}^N q_{ij}(\mathbf{x}_i-\mu_j)(\mathbf{x}_i-\mu_j)^\top}{\sum_{i=1}^N q_{ij}}$
Derivations: Take derivative of lower bound w.r.t. π_k with Lagrangian: $\lambda(\sum_k\pi_k-1)$ to get $\sum_i q_{ik}(1/\pi_k)+\lambda=0\rightarrow\pi_k=(\sum_i q_{ik})/\lambda$. Then use $\sum_k\pi_k=1\rightarrow(\sum_k\sum_i q_{ik})/\lambda=N/\lambda=1\rightarrow\lambda=1/N$ and plug in.
For μ_k: $-\sum_i q_{ik}\Sigma_k^{-1}(\mathbf{x}_i-\mu_i)=\Sigma_k^{-1}\sum_i q_{ik}(\mathbf{x}_i-\mu_k)$ (note: Σ is symmetric & invertible so $Ax=0$ iff $x=0$), so $\sum_i q_{ik}(\mathbf{x}_i-\mu_k)=0\rightarrow\mu_k=(\sum_i q_{ik}\mathbf{x}_i)/(\sum_i q_{ik})$
Guaranteed to converge to local optimum.
Comparison to K-Means
Soft assignments (not hard), learn cov. matrix (not spherical clusters), slow (not fast), more (not less) iterations, use K-means as initialization (use sample coveriance as matrix, use fraction of datapoints as mixing weights). K-means as a special case of GMM with covariances $\Sigma_j=\sigma^2 I$. in the limit of $\sigma\rightarrow0$, 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 $\kappa(\cdot)$). Compare for different parameters and take smallest:
$\mathrm{AIC}(\theta|\mathbf{X})=-\log p_\theta(\mathbf{X})+\kappa(\theta)$
$\mathrm{BIC}(\theta|\mathbf{X})=-\log p_\theta(\mathbf{X})+\frac{1}{2}\kappa(\theta)\log N$
(penalizes complexity more)
Example: #free params, fixed cov. matrix: $\kappa(\theta)=K\cdot D+(K-1)$ (K: # clusters, D: $\dim(\text{data})=\dim(\mu_i)$), full cov. matrix: $\kappa(\theta)=K(D+\frac{D(D+1)}{2})+(K-1)$.

Neural Networks

Activation: ReLU: $\max(0,x)$
$\tanh(x)=\frac{e^x-e^{-x}}{e^x+e^{-x}}$, $\tanh'(x)=1-\tanh^2(x)$
sigmoid $s(x)=\frac{1}{1+e^{-x}}$, $s'(x)=s(x)(1-s(x))$
Output: linear regression $\mathbf{y}=\mathbf{W}^L\mathbf{x}^{L-1}$,
binary (logistic) $y_1=\frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x}^{L-1})}$,
multiclass (soft-max) $y_k=\frac{\exp(\mathbf{w}_k^T\mathbf{x}^{L-1})}{\sum_{m=1}^K\exp(\mathbf{w}^T\mathbf{x}^{L-1})}$.
Loss function: $l(y,\hat{y})$: squared loss $\frac{1}{2}(y-\hat{y})^2$,
cross-entropy loss $-y\log\hat{y}-(1-y)\log(1-\hat{y})$
$\mathrm{Conv}_{n,m}^{k\times k}(\mathbf{x};\mathbf{w})=\sigma\left(b+\sum_{i=-k}^k\sum_{j=-k}^k w_{i,j}x_{n+i,m+j}\right)$
CNN: weight sharing (<< param), shift invar. filters

Backpropagation

$J_{ij}=\frac{\partial\mathbf{x}_i^{out}}{\partial\mathbf{x}_j^{in}}=w_{ij}\cdot\sigma'(\mathbf{w}_i^\top\mathbf{x}^{in})$. Across multiple layers: $\frac{\partial\mathbf{x}^{(l)}}{\partial\mathbf{x}^{(l-n)}}=\mathbf{J}^{(l)}\cdot\frac{\partial\mathbf{x}^{(l-1)}}{\partial\mathbf{x}^{(l-n)}}=\mathbf{J}^{(l)}\cdot\mathbf{J}^{(l-1)}\cdots\mathbf{J}^{(l-n+1)}$ and then back prop. $\nabla_{\mathbf{x}^{(l)}}^\top\ell=\nabla_{\mathbf{y}}^\top\ell\cdot\mathbf{J}^{(L)}\cdots\mathbf{J}^{(l+1)}$
$\frac{\partial l}{\partial w_{ij}^{(l)}}=\frac{\partial l}{\partial x_i^{(l)}}\frac{\partial x_i^{(l)}}{\partial w_{ij}^{(l)}}$, $\frac{\partial x_i^l}{\partial w_{ij}^l}=\sigma'([\mathbf{w}_i^{(l)}]^T\mathbf{x}^{(l-1)})\cdot x_j^{(l-1)}$

Generative Models

Variational Autoencoder (VAE)

Sample random vector $\mathbf{z}\sim\mathcal{N}(0,\mathbf{I})$. Transform through (deterministic) DNN $F_\theta:\mathbb{R}^m\rightarrow\mathbb{R}^n$. Note: expectations $\mathbb{E}_x[f(x)]=\mathbb{E}_z[f(F_\theta(z))]$ (law of the unconscious statistician). Infeasible, would need to find inv. Jacobian determinant
$D_{KL}(P\|Q)=\sum_i P(i)\log\frac{P(i)}{Q(i)}=\mathbb{E}_i[\frac{\log P_i}{\log Q_i}]$
More general: $p_\theta(\mathbf{x}|\mathbf{z})$ instead of F_θ: ELBO
$\log p_\theta(x^{(i)})=\mathbb{E}_{z\sim q_\phi(z|x)}\left[\log\frac{p_\theta(z)\cdot p_\theta(x^{(i)}|z)}{p_\theta(z|x^{(i)})}\frac{q_\Phi(z|x^{(i)})}{q_\Phi(z|x^{(i)})}\right]$
$=\mathbb{E}_z[\log p_\theta(x^{(i)}|z)]-D_{KL}(q_\Phi(z|x^{(i)})||p_\theta(z))$
$+D_{KL}(q_\Phi(z|x^{(i)})||p_\theta(z|x^{(i)})$ (drop last part)
1st: reconstr. quality, 2nd: posterior close to prior.
Update: $\nabla_\theta\mathbb{E}_{q_\phi}[\log p_\theta(x|z)]=\mathbb{E}[\nabla_\theta\log p_\theta(x|z)]\approx\frac{1}{L}\sum_{r=1}^L\nabla_\theta\log p_\theta(x|z^{(r)}), z^{(r)}\sim^{iid}q_\theta(\cdot|x)$
Reinforce trick:
$\nabla_\theta\mathbb{E}_{q_\theta}[\mathcal{L}(\mathbf{x},\mathbf{z})]=\mathbb{E}_{q_\theta}[\mathcal{L}(\mathbf{x},\mathbf{z})\nabla_\theta\log q_\theta(\mathbf{z};\mathbf{x})]$
Re-parameterization trick: use variational distribution $q_\phi(\mathbf{z};\mathbf{x})=g_\phi(\zeta;\mathbf{x})$ for ζ simple
(e.g. $\zeta\sim\mathcal{N}(\mathbf{0},\mathbf{I}),\mathbf{z}=\mu+\mathbf{U}\zeta\Rightarrow\mathbf{z}\sim\mathbf{N}(\mu,\mathbf{U}\mathbf{U}^T)$)
Stochastic Backprop: for $\zeta^{(r)}\sim^{iid}$ simple
$\mathbb{E}_{q_\phi}[\nabla_\phi\mathcal{L}(\mathbf{x},\mathbf{z})]\approx\frac{1}{L}\sum_{r=1}^L[\nabla_\phi\mathcal{L}(\mathbf{x},g_\phi(\zeta^{(r)}))]$

Generative Adversarial Network (GAN)

$\min_G\max_D V(D,G)=\mathbb{E}_{\mathbf{x}\sim p_{data}(\mathbf{x})}[\log D(\mathbf{x})]$
$+\mathbb{E}_{z\sim p_z(\mathbf{z})}[\log(1-D(G(\mathbf{x}))]$
$\theta^*:=\arg\min_{\theta\in\Theta}\{\sup_{\phi\in\Phi}l(\theta,\phi)\}$ **SGD:** $\theta^{t+1}=\theta^t-\eta\nabla_\theta l(\theta^t,\phi^t)$; $\phi^{t+1}=\phi^t+\eta\nabla_\phi l(\theta^{t+1},\phi^t)$

Autoregressive Models

Generate output one variable at a time: $p(x_1,\dots,x_m)=\prod_{t=1}^m p(x_t|x_{1:t-1})$
PixelCNN: uses exactly that over a window to predict the next pixel (slow process).

Sparse Coding

Orthogonal Basis

Transform: For $\mathbf{x}$ and orthog. mat. $\mathbf{U}$ compute $\mathbf{z}=\mathbf{U}^\top\mathbf{x}$. For compression, can drop small values: $\hat{\mathbf{x}}=\mathbf{U}\hat{\mathbf{z}}$, $\hat{z}_i=z_i$ if $|z_i|>\varepsilon$ else 0. Pros: fast inverse; preserves energy. or $\mathbf{x}$ and orthog. mat. $\mathbf{U}$ compute $\mathbf{z}=\mathbf{U}^\top\mathbf{x}$ else 0. Reconstr. Error $\|\mathbf{x}-\hat{\mathbf{x}}\|^2=\sum_{d\notin\sigma}\langle\mathbf{x},\mathbf{u}_d\rangle^2$ for a subset of basis σ.
Haar Wavelets: scaling function $\phi(x)=[1,1,1,1]$, mother $W(x)=[1,1,-1,-1]$, dilated $W(2x)=[1,-1,0,0]$, translated $W(2x-1)=[0,0,1,-1]$. Do not forget to normalize.
Comparison to Fourier basis: local (not global) support, good for localized (not sin like, repeating) signals. **PCA basis:** data-dependent, but optimal for given Σ. $\hat{\mathbf{x}}=\mathbf{U}_K\mathbf{z}_{[1:K]}$

Overcomplete Dictionaries

Use more atoms than dimensions, then choose the best representation. (e.g.Gabor wavelets use Fourier like features in a localized Gaussian window).
Linear dependency measure: **Coherence**
- $m(\mathbf{U})=\max_{i,j:i\neq j}|\mathbf{u}_i^\top\mathbf{u}_j|$
- $m(\mathbf{B})=0$ if $\mathbf{B}$ orth. mat.
- $m([\mathbf{B},\mathbf{u}])\geq\frac{1}{\sqrt{D}}$ if atom $\mathbf{u}$ is added to $\mathbf{B}$

Signal Reconstruction: orthonormal: $\mathbf{x}=\mathbf{U}\mathbf{z}$, spanning basis (linearly independent): $\mathbf{x}=(\mathbf{U}^T)^{-1}\mathbf{z}$ (can be ill-conditioned), overcomplete: solve $\mathbf{z}^*\in\arg\min_{\mathbf{z}}\|\mathbf{z}\|_0$ s.t. $\mathbf{x}=\mathbf{U}\mathbf{z}$. (NP hard). Can convexify with L_1 norm or greedy approx.:
Matching Pursuit (MP) 1. init: $\hat{z}\leftarrow0,r\leftarrow x$ 2. while $\|\mathbf{z}\|_0<K$ do 3. select atom with smallest angle $j^*=\arg\max_j|\langle\mathbf{u}_j,\mathbf{r}\rangle|$ 4. update coefficients: $\hat{z}\leftarrow\hat{z}+\langle\mathbf{u}_{i^*},\mathbf{r}\rangle\mathbf{u}_{\mathbf{j}^*}$ 5. update residual: $\mathbf{r}\leftarrow\mathbf{r}-\langle\mathbf{u}_{i^*},\mathbf{r}\rangle\mathbf{u}_{i^*}$.
Exact recovery when: $K<1/2(1+1/m(\mathbf{U}))$
Instance when MP never exactly match x: Idea: if we always just half residual and start with a pos. number, we will never arrive at 0. Start with (0,1). u1 = (1,0) u2 = ($\sqrt{2}$/2, $\sqrt{2}$/2), u3 = ($\sqrt{3}$/2, 1/2).
Instance where MP does not have best solution: u1 = (1,0), u2 = (0,1), u3 = (sqrt(2)/2, sqrt(2)/2). x = (2,1), will have largest correlation with u3, but residual then cannot be expressed with only u1 or u2, so we will use all 3 vectors instead of only u1 & u2.

Compressive Sensing

Aquire set $\mathbf{y}$ of M linear combinations of signal, then reconstruct from it. $y_k=\langle\mathbf{w}_k,\mathbf{x}\rangle,k=1,\dots,M$
$\mathbf{y}=\mathbf{W}\mathbf{x}=\mathbf{W}\mathbf{U}\mathbf{z}=:\Theta\mathbf{z}$ with $\Theta=\mathbf{W}\mathbf{U}\in\mathbb{R}^{M\times D}$.
Any orthonormal basis $\mathbf{U}$ can obtain a stable reconstr. for any K-sparse compressible signal if:
- $\mathbf{W}$ is Gaussian random projection, i.e. $w_{ij}\sim\mathcal{N}(0,\frac{1}{D})$
- $M\geq cK\log\frac{D}{K}$(some constant c). Reconstruct as before: $\mathbf{z}^\star\in\arg\min_{\mathbf{z}}\|\mathbf{z}\|_0$, s.t. $\mathbf{y}=\Theta\mathbf{z}$

Dictionary Learning

Adapt the dict. to signal charact.$(\mathbf{U}^\star,\mathbf{Z}^\star)\in\arg\min_{\mathbf{U},\mathbf{Z}}\|\mathbf{X}-\mathbf{U}\mathbf{Z}\|_F^2$ not jointly convex (just 1 arg)
Matrix Factorization by Iter Greedy Minim. 1. Coding step: $\mathbf{Z}^{t+1}\in\arg\min_{\mathbf{Z}}\|\mathbf{X}-\mathbf{U}^t\mathbf{Z}\|_F^2$ subj. to $\mathbf{Z}$ being sparse ($\mathbf{z}_n^{t+1}\in\arg\min_{\mathbf{z}}\|\mathbf{z}\|_0$ s.t. $\|\mathbf{x}_n-\mathbf{U}^t\mathbf{z}\|_2\leq\sigma\|\mathbf{x}_n\|_2$) 2. Init: random, samples from X or fixed overcomplete dictionary. 3. Dict update step: $\mathbf{U}^{t+1}\in\arg\min_{\mathbf{U}}\|\mathbf{X}-\mathbf{U}\mathbf{Z}^{t+1}\|_F^2$, subj. to $\forall l\in[L]:\|\mathbf{u}_l\|_2=1$ One at a time: set $\mathbf{U}=[\mathbf{u}_1^t\cdots\mathbf{u}_l\cdots\mathbf{u}_L^t]$ (fix all except $\mathbf{u}_l$), isolate $\mathbf{R}_l^t$ (residual due to atom $\mathbf{u}_l$), find $\mathbf{u}_l^*$ that minimizes $\mathbf{R}_l^t$ s.t. $\|\mathbf{u}_l^*\|_2=1$: $\min_{\mathbf{u}_l}\|\mathbf{R}_l^t-\mathbf{u}_l(\mathbf{z}_l^{t+1})^\top\|_F^2$ using SVD (first left-singular vector of $\mathbf{R}_l^t$).