1 Essentials

Norms

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

$\|\mathbf{M}\|_1 = \sum_{i,j}|m_{i,j}|$ $\quad\|\mathbf{M}\|_2 = \sigma_{\max}(\mathbf{M})$

$\|\mathbf{M}\|_p = \max_{\mathbf{v}\neq 0}\frac{\|\mathbf{Mv}\|_p}{\|\mathbf{v}\|_p}$ $\quad\|\mathbf{M}\|_\star = \sum_{i=1}^{\min(m,n)}\sigma_i$

(Nuclear Norm)$\|M\|_* = \sum_i \sigma_i, \sigma_i$: singular value of M.

rank

$rank(XY) \le rank(X)$

$rank(XY) = rank(X), \forall Y \in \mathbb{R}^{n\times n}, rank(Y) = n$

$rank(A) = rank(UDV^T) = rank(D)$

Derivatives

$\frac{\partial}{\partial \mathbf{x}}(\mathbf{b}^\top\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^\top\mathbf{b}) = \mathbf{b}$ $\quad\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^\top\mathbf{x}) = 2\mathbf{x}$

$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^\top\mathbf{A}\mathbf{x}) = (\mathbf{A}^\top + \mathbf{A})\mathbf{x}$ $\quad\frac{\partial}{\partial \mathbf{x}}(\mathbf{b}^\top\mathbf{A}\mathbf{x}) = \mathbf{A}^\top\mathbf{b}$

$\frac{\partial}{\partial \mathbf{X}}(\mathbf{c}^\top\mathbf{X}\mathbf{b}) = \mathbf{c}\mathbf{b}^\top$ $\quad\frac{\partial}{\partial \mathbf{X}}(\mathbf{c}^\top\mathbf{X}^\top\mathbf{b}) = \mathbf{b}\mathbf{c}^\top$

$\frac{\partial}{\partial \mathbf{x}}(\|\mathbf{x}-\mathbf{b}\|_2) = \frac{\mathbf{x}-\mathbf{b}}{\|\mathbf{x}-\mathbf{b}\|_2}$ $\quad\frac{\partial}{\partial \mathbf{x}}(\|\mathbf{x}\|_2^2) = \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^\top\mathbf{x}) = 2\mathbf{x}$

$\frac{\partial}{\partial \mathbf{X}}(\|\mathbf{X}\|_F^2) = 2\mathbf{X}$

$\frac{\partial}{\partial \mathbf{x}}(\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2) = 2(\mathbf{A}^\top\mathbf{A}\mathbf{x} - \mathbf{A}^\top\mathbf{b})$

Eigenvalue / -vectors

Eigenvalue Problem: $\mathbf{Ax} = \lambda\mathbf{x}$

1. solve $\det(\mathbf{A}-\lambda\mathbf{I}) \overset{!}{=} 0$ resulting in $\{\lambda_i\}_i$
2. $\forall\lambda_i$: solve $(\mathbf{A}-\lambda_i\mathbf{I})\mathbf{x}_i = \mathbf{0}$, for $\mathbf{x}_i$.
3. $\mathbf{A}\in\mathbb{R}^{N\times N}$ 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: $A = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$ ($\mathbf{Q}$ orthogonal).

Convexity

$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda)f(y), \forall\lambda\in[0,1]$ if $dom\ f$ and f convex

Distribution

Dirichlet: $p(u_i|\alpha) = \prod_{z=1}^{K} u_{zi}^{\alpha_k - 1}$

Multinomial: $p(x|\pi) = \frac{l!}{\prod_j x_j!}\prod_j \pi_j^{x_j}$

Gaussian: $f(x) = \frac{1}{2\pi\sqrt{|\Sigma|}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}$

KL-Divergence

$D_{KL}(P\|Q) = -\sum_{x\in\mathcal{X}}P(x)\log\frac{P(x)}{Q(x)}$

$x\sim\mathcal{N}(\mu,\sigma^2)\Rightarrow$

$\int_{-\infty}^{\infty}p(x)\log p(x)dx = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}$

Lagrangian Multipliers

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

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

Dual function: $D(\boldsymbol{\lambda},\boldsymbol{\nu}) := \inf_{\mathbf{x}} L(\mathbf{x},\boldsymbol{\lambda},\boldsymbol{\nu}) \in \mathbb{R}$

Dual Problem: $\max_{\boldsymbol{\lambda},\boldsymbol{\nu}} D(\boldsymbol{\lambda},\boldsymbol{\nu})$ s.t. $\boldsymbol{\lambda}\ge\mathbf{0}$. Note: $\max_{\boldsymbol{\lambda},\boldsymbol{\nu}} D(\boldsymbol{\lambda},\boldsymbol{\nu}) \le \min_{\mathbf{x}} f(\mathbf{x})$, equality if $dom\ f$ and f convex

2 Principle Component Analysis

Algorithm Implementation

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

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

Iterative View

Idea: optimal direction = principle eigenvector of the sample Residual r_i: $x_i - \tilde{x}_i = I - uu^Tx_i$

Cov of r: $\frac{1}{n}\sum_{i=1}^{n}(I-uu^T)x_ix_i^T(I-uu^T)^T = (I-uu^T)\Sigma(I-uu^T)^T = \Sigma - \lambda uu^T$

1. Find principal eigenvector of $(\Sigma - \lambda uu^T)$, which is the second eigenvector of Σ
2. 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 $|\lambda_1| > |\lambda_j|(\forall j\ge 2)$

Reconstruction Error

$\min_{rank(B)=K}\frac{1}{N}\|A-B\|_F^2 = \frac{1}{N}\|A-A_k\|_F^2 = \sum_{r=k+1}^{rank(A)}\lambda_r$, λ_r is the eigenvalue of $\Sigma = \frac{1}{N}AA^T$.

In SVD, there is no $\frac{1}{N}$

Interpret eigenvalues as the variance in the dimension specified by the corresponding eigenvector.

Others

1. If $A = BB^T$ then A is semi-positive definite.

Proof: $v^TAv = v^TBB^Tv = \|B^Tv\|_2 \ge 0$

2. Spectral Theorem: Matrix A is diagonalizable by an orthogonal matrix $\Leftrightarrow$ it is symmetric
3. Compared to general linear autoencoder, PCA is unique and interpretable.
4. Compare power methods and SVD.

Power: good for small k. Robust and conceptually easy.

SVD: good for mid-sized problem. Leverage wealth of numerical techniques, e.g. QR decomposition.

3 Support Vector Discriminant

$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top = \sum_{k=1}^{\text{rank}(\mathbf{A})}d_{k,k}u_k(v_k)^\top$

$\mathbf{U}^\top\mathbf{U} = I = \mathbf{V}^\top\mathbf{V}$ ($\mathbf{U}$,$\mathbf{V}$orthonormal)

$\mathbf{U}_{:i}$: eigenvectors of $\mathbf{A}\mathbf{A}^\top$, $\mathbf{V}_{:i}$: eigenvectors of $\mathbf{A}^\top\mathbf{A}$, $\mathbf{D}_{ii}$: singular values.

Algorithm Implementation

1. eigenvalues of $\mathbf{A}^\top\mathbf{A}$,in descending order, are $D_{ii} = \sqrt{A_{ii}}$ 2. eigenvectors of $\mathbf{A}^\top\mathbf{A} \to \mathbf{V}$. 3. $\mathbf{U} = \mathbf{A}\mathbf{V}\mathbf{D}^{-1}$. 4. normalize each column of $\mathbf{U}$ and $\mathbf{V}$.

Low-Rank approximation

$A = \sum_{i=1}^{k}d_iu_iv_i^T = U_kD_kV_k^T$

Echart-Young Theorem

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

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

4 Matrix Reconstruction

Two formalization

1. (Low-rank) $\min_{X:rank(X)\le k}\|A-X\|_{\mathcal{I}}^2$

Regularize$\Rightarrow$ $L(U,V) = \|A-U^TV\|_{\mathcal{I}}^2 + \lambda\|U\|_F^2 + \lambda\|V\|_F^2 = \sum_{(i,j)\in\mathcal{I}}(a_{ij}-u_i^Tv_j)^2 + \lambda\sum_{i=1}^{m}\|u_i\|^2 + \lambda\sum_{j=1}^{n}\|v_j\|^2$

- non-convex w.r.t (U,V) but convex w.r.t U and V.

2.(Exact matrix) $\min_X rank(X)$, s.t. $\|A-X\|_{\mathcal{I}} = 0$

- the rank function (objective) is not convex, it is not smooth
- the constraint is very stringent.

ALS

For the first formalization: low-rank.

$u_i = (\sum_{j:(i,j)\in\mathcal{I}}v_jv_j^T + \lambda I_k)^{-1}\sum_{j:(i,j)\in\mathcal{I}}a_{ij}v_j$

Interpret: given low-dimensional representations of the items, compute independently the best representation of each user

Computational complexity for u_i is $O(n_ik^2 + k^3)$, k features, n_i number of items evaluated by user i.

$v_j = (\sum_{i:(i,j)\in\mathcal{I}}u_iu_i^T + \lambda I_k)^{-1}\sum_{i:(i,j)\in\mathcal{I}}a_{ij}u_i$

Interpret: given low-dimensional representations of the users, compute independently the best representation of each item.

Nuclear Norm

(Exact form) $\min_B\|B\|_*, s.t.\|A-B\|_G = 0$

(Approximate)$\min_B\|A-B\|_G^2$, s.t. $\|B\|_* \le r$

Property1:$\text{rank}(B) \ge \|B\|_*, \forall\|B\|_2 \le 1$

Property2: convexity. Proof:

goal: $\|\lambda A + (1-\lambda)B\|_* \le \lambda\|A\|_* + (1-\lambda)\|B\|_*$

key1: write $\lambda A + (1-\lambda)B = U_\lambda D_\lambda V_\lambda^T$

key2: prove $\text{trace}(U_\lambda^TAV_\lambda) \le \sum_j\sigma_j(A) = \|A\|_*$, using Cauchy-Schwartz inequality.

SVD shrinkage

Idea: SVD thresholding + projection.

Objective: $B^* = \arg_B\min\{\|B\|_* + \frac{1}{2\tau}\|B\|_F^2\}$,s.t. $\Pi_{\mathcal{I}}(A-B) = x_{ij}\forall(i,j)\in\mathcal{I}$

Algorithm: $B_{t+1} = B_t + \eta\Pi_{\mathcal{I}}(A - shrink_\tau(B_t))$

$shrink_\tau$: all singular values are reduced by at least τ.

5 Non-Negative Matrix Factorization

pLSA

Conditional independence: $p(w|d,z) = p(w|z)$

Objective: $L(U,V) = \sum_{i,j}x_{ij}\log p(w_j|d_i) = \sum_{i,j}x_{ij}\log\sum_z p(w_j|z)p(z|d_i) = \sum_{i,j}x_{ij}\log\sum_z v_{zj}u_{zi}$

$v_{jz} := p(w_j|z), u_{zi} := p(z|d_i), \sum_z v_{jz} = \sum_k u_{ki} = 1$

EM

Latent variable: q_{zij} probability of w_j in d_i generated by topic z, $\sum_z q_{zij} = 1$

Lower bound: $\log\sum_{z=1}^{K}q_{zij}\frac{u_{zi}v_{zj}}{q_{zij}} \ge \sum_{z=1}^{K}q_{zij}[\log u_{zi} + \log v_{zj} - \log q_{zij}]$

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

M-Step: $u_{zi} = p(z|d_i) = \frac{\sum_j x_{ij}q_{zij}}{\sum_z\sum_j x_{ij}q_{zij} = \sum_j x_{ij}}$,

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

Latent Dirichlet Allocation

Objective:

$p(x|V,\alpha) = \int p(X|V,u)p(u|\alpha)du$

$p(x|V,u) = \text{Multi}(x|\pi), \pi_j := \sum_z v_{zj}u_z$

Generative model:

1. for d_i, sample $u_i \sim \text{Dirichlet}(\alpha)$
2. for word slot t: (1)sample topic $z^t \sim \text{Multi}(u_i)$; (2) sample word $w^t \sim \text{Multi}(v_{z^t})$

NMF Algorithm for quadratic cost function

NMF: $\mathbf{X} \approx \mathbf{U}^\top\mathbf{V}, x_{ij}$, $\min_{\mathbf{U},\mathbf{V}}J(\mathbf{U},\mathbf{V}) = \frac{1}{2}\|\mathbf{X} - \mathbf{U}^\top\mathbf{V}\|_F^2$, s.t. $\forall i,j,z: u_{zi}, v_{zj} \ge 0$

Constraints: non-negativity, normalization.

Projected ALS 1. randomly init $\mathbf{U}$, $\mathbf{V}$ 2. repeat for enough times:

2.1. update $\mathbf{U}$: $(\mathbf{V}\mathbf{V}^\top)\mathbf{U} = \mathbf{V}\mathbf{X}^\top$, proj. $u_{zi} = \max\{0, u_{zi}\}$

2.2. update $(\mathbf{U}\mathbf{U}^\top)\mathbf{V} = \mathbf{U}\mathbf{X}$, proj. $v_{zj} = \max\{0, v_{zj}\}$

6 Word Embeddings

(Skip-gram model) Log-likelihood:

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

Latent vector model

log-bilinear: $\log p(w|w') = \langle\mathbf{x}_w, \mathbf{z}_{w'}\rangle + b_w + \text{const}$

Negative Sampling We change the objective function as $L(\theta) = \sum_{(i,j)\in\Delta^+}\log\sigma(\langle\mathbf{x}_i,\mathbf{z}_j\rangle) + \sum_{(i,j)\in\Delta^-}\log\sigma(-\langle\mathbf{x}_i,\mathbf{z}_j\rangle)$

distribution: ratio to the appearance frequency. Number: given

GloVe (Weighted Square Loss)

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

$= \sum_{(i,j)}f(n_{ij})(\log n_{ij} - \log\exp[\langle\mathbf{x}_i,\mathbf{y}_j\rangle + b_i + d_j])^2$

weighting: $f(n) = \min\{1, (\frac{n}{n_{max}})^\alpha\}$, $\alpha\in(0;1]$.

SGD solution

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

SGD

GD: $\mathbf{w}_{k+1} = \mathbf{w}_k - \gamma_k \frac{1}{n} \sum_i \nabla f_i(\mathbf{w}_k)$
SGD: $\mathbf{w}_{k+1} = \mathbf{w}_k - \gamma_k \nabla f_{r_k}(\mathbf{w}_k)$, r_k random.

7 K-means

Objective $\min_{\mathbf{U},\mathbf{Z}} J(\mathbf{U},\mathbf{Z}) = \|\mathbf{X} - \mathbf{U}\mathbf{Z}^T\|_F^2$
Constraints $\sum_j z_{ij} = 1, z_{ij} \in \{0,1\}$
Strategy alternating (EM) minimization
- $z_{ij} = 1, j = \arg_k \min \|x_i - u_k\|^2$ (closed centroid)
- $\mathbf{u}_j = \frac{\sum_{i=1}^N z_{ij}\mathbf{x}_i}{\sum_{i=1}^N z_{ij}}$

Computational cost of each iteration: $O(k \cdot n \cdot d)$
Quadratic convergence rate.
K-means++ 1. $\mathcal{U}_1 = \{x_I\}, I \sim \text{Uniform}[1:N]$
2. $\mathcal{U}_{k+1} = \mathcal{U}_{k+1} \bigcup \{x_I\}, I \sim \text{Categorical}(\mathbf{p}), p_i = \frac{D_i^2}{\sum_{j=1}^N D_j^2}$, D_i^2 the closest distance of node i to all the centroids.
more expensive but theoretical $O(\log K)$ guarantee
Core set $p_i = \frac{1}{2N} + \frac{D_i^2}{2\sum_{j=1}^N D_j^2}$, $D_i^2 = \|x_i - \frac{1}{N}\sum_j x_j\|^2$
selected sample weighed by $\frac{1}{mp_i}$
ϵ-approximation guarantee: $m \propto \frac{dk \log k + \log 1/\delta}{\epsilon^2}$

8 Gaussian Mixture Models (GMM)

Mixture model

Complete data distribution:
$p_\theta(\mathbf{x},\mathbf{z}) = \prod_{k=1}^K \left(\pi_k p_{\theta_k}(\mathbf{x})\right)^{z_k}$
- $p_\theta(\mathbf{x}) = \sum_{k=1}^K \pi_k p_{\theta_k}(\mathbf{x})$
- $\Pr(z_k = 1) = \pi_k \Leftrightarrow p(\mathbf{z}) = \prod_{k=1}^K \pi_k^{z_k}$

Generation: given π, generate x
1. sample cluster index $j \sim \text{Categorical}(\pi)$
2. sample x from the j-th component.

Inference: given x, infer z.
$\Pr(z_k = 1|\mathbf{x}) = \frac{\pi_k p_{\theta_k}(\mathbf{x})}{\sum_{l=1}^K \pi_l p_{\theta_l}(\mathbf{x})}$
Objective: find π, $\{\theta_k\}$
MLE estimate: $\hat{\pi}, \hat{\Theta} = \arg_{\pi,\Theta} \max \log p_\theta(\mathbf{X}) = \log \prod_{n=1}^N p_\theta(\mathbf{x}_n) = \sum_{n=1}^N \log\left(\sum_{k=1}^K \pi_k p_{\theta_k}(\mathbf{x}_n)\right)$
$\geq \sum_{n=1}^N \sum_{k=1}^K q_k[\log p_{\theta_k}(\mathbf{x}_n) + \log \pi_k - \log q_k]$
with $\sum_{k=1}^K q_k = 1$ by Jensen Inequality.
EM method Objective Lagrangian
Expectation (Two-side equal): $q_j^* = \frac{\pi_j p(x;\theta_j)}{\sum_{l=1}^K \pi_l p(x;\theta_l)}$
$\pi_j^* := \frac{1}{N}\sum_{i=1}^N q_{ij}^*$
Maximization: Gaussian model: $\theta_j = (\mu_j, \Sigma_j)$
- $\mu_j^* := \frac{\sum_{i=1}^N q_{ij}\mathbf{x}_i}{\sum_{i=1}^N q_{ij}}$, $\Sigma_j^* = \frac{\sum_{i=1}^N q_{ij}(\mathbf{x}_i - \mu_j)(\mathbf{x}_i - \mu_j)^\top}{\sum_{i=1}^N q_{ij}}$

To K-means

With covariances $\Sigma_j = \sigma^2 * I, \sigma \to 0$

Model Selection

$\text{AIC}(\theta|\mathbf{X}) = -\log p_\theta(\mathbf{X}) + \kappa(\theta)$
$\text{BIC}(\theta|\mathbf{X}) = -\log p_\theta(\mathbf{X}) + \frac{1}{2}\kappa(\theta)\log N$
BIC penalizes complexity more.
Example for GMM
- fixed covariance matrix: $\kappa(\theta) = K \cdot D + (K-1)$
- full covariance matrices: $\kappa(\theta) = K \cdot (D + \frac{D(D+1)}{2}) + (K-1)$

9 Convolutional Neural Networks

Neurons:
$\mathbf{x} = \sigma(w_0 + \sum_{i=1}^M x_i w_i)$. $\mathbf{x}^l = \sigma^l\left(\mathbf{W}^{(l)}\mathbf{x}^{(l-1)}\right)$,
$\mathbf{y} = \sigma^{(L)}\left(\mathbf{W}^(L)\sigma^{(L-1)}\left(\cdots\left(\sigma^{(1)}\left(\mathbf{W}^{(1)}\mathbf{x}\right)\cdots\right)\right)\right)$
Activation: ReLu $\max\{x,0\}$, Sigmoid: $s(x) = \frac{1}{1+e^{-x}}$, $s'(x) = s(x)(1-s(x))$
Output:
Regression: Loss: MSE, $\hat{\mathbf{y}} = \mathbf{W}^L \mathbf{x}^{L-1}$
Classification: Loss: cross entropy: $-\sum_{i=1}^N \sum_{j=1}^K y_{ij}\log(\hat{p}_{ij})$
Logistic: $\hat{y_1} = \text{P}[Y = 1|\mathbf{x}] = \frac{1}{1+\exp[-\langle \mathbf{w}_1^L, \mathbf{x}^{L-1}\rangle]}$
Soft-max: $\hat{y_k} = \text{P}[Y = k|\mathbf{x}] = \frac{\exp[\langle \mathbf{w}_k^L, \mathbf{x}^{L-1}\rangle]}{\sum_{m=1}^K \exp[\langle \mathbf{w}_m^L, \mathbf{x}^{L-1}\rangle]}$.
Objective $\mathcal{L}_\lambda(\theta;\mathcal{X}) = \frac{1}{N}\sum_{i=1}^N l(y_i, \hat{y}_i) + \frac{\lambda}{2}\|\theta\|_2^2$

Backpropagation

SGD $\theta \leftarrow (1-\eta\lambda)\theta - \eta\nabla_\theta l(y_t^*; y(x_t;\theta))$
Jacobian $\mathbf{x}_\mathbf{i}^+ = \sigma(\mathbf{w}_i^\top \mathbf{x}), J_{ij} = \frac{\partial \mathbf{x}_i^+}{\partial \mathbf{x}_j} = w_{ij} \cdot \sigma'(\mathbf{w}_i^\top \mathbf{x})$.
$\frac{\partial \mathbf{x}^{(l)}}{\partial \mathbf{x}^{(l-n)}} == \mathbf{J}^{(l)} \cdot \mathbf{J}^{(l-1)} \cdots \mathbf{J}^{(l-n+1)}$
$\nabla_{x^{(l)}}^\top \ell = \nabla_y^\top \ell \cdot \mathbf{J}^{(L)} \cdots \mathbf{J}^{(l+1)}$
$\frac{\partial l}{\partial w_{ij}^{(l)}} = \frac{\partial l}{\partial x_i^{(l)}} \frac{\partial x_i^{(l)}}{\partial w_{ij}^{(l)}} = \frac{\partial l}{\partial x_i^{(l)}} \sigma'([\mathbf{w}_i^{(l)}]^\top x^{(l-1)}) x_j^{(l-1)}$

CNN

Convolution layer $F_{n,m}(\mathbf{x};\mathbf{w}) = \sigma(b + \sum_{k=-2}^2 \sum_{l=-2}^2 w_{k,l} x_{n+k,m+l})$. To reduce dimension of convolution, use {max, avg}-pooling

10 Generative models

VAE

Objective
$\log p_\theta(x) = \mathbb{E}_{h\sim q_\phi}[\log p_\theta(x)] = \mathbb{E}_h[\log p_\theta(x|h)] - D[q_\phi(h|x)\|p_\theta(h)] + D[q_\phi(h|x)\|p_\theta(h|x)]$
$= \text{ELBO} + D[q_\phi(h|x)\|p_\theta(h|x)]$
ELBO $\log p(x;\theta) = \log \int p(x,h)dh \geq \int q(h|x)[\log p(x|h) + \log p(h) - \log q(h|x)]dh = \int q(h|x)\log p(x|h)dh - D[q(h|x)\|p(h)]$
Stochastic approximation: $\nabla_\theta \mathbb{E}_{q_\phi}[\mathcal{L}(\mathbf{x},\mathbf{h})] \simeq \frac{1}{L}\sum_{r=1}^L \nabla_\theta \log p_\theta(\mathbf{x}|\mathbf{h}^{(r)})$, $\mathbf{h}^{(r)} \sim q_\phi(\cdot|\mathbf{x})$, i.i.d
In the log-likelihood context: $\mathbb{E}_{q_\phi}[\mathcal{L}(\mathbf{x},\mathbf{h})] = B(q_\phi, \mathbf{x})$
Re-parametrization trick $q_\phi(h;x) = g_\phi(\zeta;x)$, ζ ~simple distribution

GAN

Objective: $\min_G \max_D \mathbb{E}_{x\sim P_{data}(x)}[\log D(x)] + \mathbf{E}_{z\sim P_z(z)}[\log(1 - D(G(z)))]$ or
$l(\theta,\phi) := \mathbb{E}_{\tilde{p}_\theta}[y \ln q_\phi(\mathbf{x}) + (1-y)\ln(1 - q_\phi(\mathbf{x}))]$
- θ: Generator, $\tilde{p}_\theta(\mathbf{x}, y = 1) = p(y = 1) \cdot p(\mathbf{x}), \tilde{p}_\theta(\mathbf{x}, y = 0) = p(y = 0) \cdot p_\theta(\mathbf{x})$
- ϕ: Discriminator. $q_\phi : \mathbf{x} \mapsto [0;1]$

Saddle-point problem SGD as a heuristic solution (may diverge!)
$\theta^{t+1} = \theta^t - \eta\nabla_\theta l(\theta^t, \phi^t)$, minimize: minus
$\phi^{t+1} = \phi^t + \eta\nabla_\phi l(\theta^{t+1}, \phi^t)$, maximize: plus

Autoregressive model

$p(x_1, x_2, \cdots, x_m) = \prod_{t=1}^m p(x_t|x_{1:t-1})$

11 Sparse Coding

Orthogonal Basis

For $\mathbf{x}$ and o.n.b. $\mathbf{U}$ compute $\mathbf{z} = \mathbf{U}^\top \mathbf{x}$. Approx $\hat{\mathbf{x}} = \mathbf{U}\hat{\mathbf{z}}$, $\hat{z}_i = z_i$ if $|z_i| > \epsilon$ else 0. Reconstruction Error $\|\mathbf{x} - \hat{\mathbf{x}}\|^2 = \sum_{d \notin \sigma}\langle \mathbf{x}, \mathbf{u}_d\rangle^2$.

Wavelet Basis

Mother wavelet: $\psi(t) = 1, 0 \leq t \leq \frac{1}{2}; = 0, \frac{1}{2} \leq t \leq 1.$
$\psi_{n,k}(t) = 2^{\frac{n}{2}}\psi(2^n t - k), 0 \leq k < 2^n$
Coherence • $m(\mathbf{U}) = \max_{i,j:i\neq j}|\mathbf{u}_i^\top \mathbf{u}_j|$
• $m([\mathbf{B},\mathbf{u}]) \geq \frac{1}{\sqrt{D}}$ if atom $\mathbf{u}$ is added to orthogonal basis $\mathbf{B}$

Matching Pursuit (MP)

1. init: $\hat{\mathbf{x}}_0 < - \leftarrow 0, \mathbf{r}_0 \leftarrow \mathbf{x}$ 2. select the basis $j^\star = \arg\max_j |\langle \mathbf{u}_j, \mathbf{r}_i\rangle|$ 3. update coefficients: $\mathbf{x}_{i+1} \leftarrow \mathbf{x}_i + \langle \mathbf{u}_{j^\star}, \mathbf{r}_i\rangle \mathbf{u}_{j^\star}$ 4. update residual: $\mathbf{r}_{i+1} \leftarrow \mathbf{r}_i - \langle \mathbf{u}_{j^\star}, \mathbf{r}_i\rangle \mathbf{u}_{j^\star}$.
Convergence $\frac{\|\mathbf{r}_{i+1}\|_2^2}{\|\mathbf{r}_i\|_2^2} = 1 - |\langle \frac{\mathbf{r}_i}{\|\mathbf{r}_i\|_2}, \mathbf{u}_{j^\star}\rangle|^2$
$\exists \mu_m in \in (0,1]$ (for overcompelete dictionary) $\|\mathbf{r}_i\|_2^2 \leq (1-\mu_{min})^i \|\mathbf{r}_0\|_2^2$
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; $\mathbf{y} = \mathbf{W}\mathbf{x} = \mathbf{W}\mathbf{U}\mathbf{z} = \theta\mathbf{z}$, $\theta \in \mathbb{R}^{M\times D}$ • Reconstruct $\mathbf{x} \in \mathbb{R}^D$ from $\mathbf{y}$; find $\mathbf{z}^\star \in \arg\min_\mathbf{z} \|\mathbf{z}\|_0$, s.t. $\mathbf{y} = \theta\mathbf{z}$ (e.g. with MP). Given $\mathbf{z}$, reconstruct $\mathbf{x}$ via $\mathbf{x} = \mathbf{U}\mathbf{z}$
Sufficient conditions: • $\mathbf{W}$ = Gaussian random projection, i.e. $w_{ij} \sim \mathcal{N}(0, \frac{1}{D})$ • $\text{M} \geq cKlog(\frac{D}{K})$, where c is some constant

12 Dictionary Learning

Objective: $(\mathbf{U}^\star, \mathbf{Z}^\star) \in \arg\min_{\mathbf{U},\mathbf{Z}} \|\mathbf{X} - \mathbf{U}\cdot\mathbf{Z}\|_F^2$

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 ($\mathbf{z}_n^{t+1} \in \arg\min_\mathbf{z} \|\mathbf{z}\|_0$ s.t.$\|\mathbf{x}_n - \mathbf{U}^t\mathbf{z}\|_2 \leq \sigma\|\mathbf{x}_n\|_2$)
2. Dict update step: $\mathbf{U}^{t+1} \in \arg\min_\mathbf{U} \|\mathbf{X} - \mathbf{U}\mathbf{Z}^{t+1}\|_F^2$. $\forall l \in 1, 2\cdots, L$:
set $\mathbf{U} = [\mathbf{u}_1^t \cdots \mathbf{u}_l \cdots \mathbf{u}_L^t]$
$\min_{u_l} \|\mathbf{X} - \mathbf{U}\mathbf{Z}^{t+1}\|_F^2 = \|(\mathbf{X} - \sum_{e\neq l}\mathbf{u}_e^t(\mathbf{z}_e^{t+1})^\top) - \mathbf{u}_l\mathbf{z}_l^{t+1}\|_F^2 = \|\mathbf{R}_l^t - \mathbf{u}_l(\mathbf{z}_l^{t+1})^\top\|_F^2$
Doing 1-SVD on $\mathbf{R}_l' = \sum_i \sigma_i \tilde{\mathbf{u}}_i \tilde{\mathbf{v}}_i^\top$
update $\mathbf{u}_l^{t+1} = \tilde{\mathbf{u}}_1$, $z_l^{t+1} = \sigma_1 \tilde{\mathbf{v}}_i$