1 Essentials

1.1 Matrix/Vector

Orthogonal: (i.e. columns are orthonormal!) $\mathbf{A}^{-1}=\mathbf{A}^\top$, $\mathbf{A}\mathbf{A}^\top=\mathbf{A}^\top\mathbf{A}=\mathbf{I}$, $\det(\mathbf{A})\in\{+1,-1\}$, $\det(\mathbf{A}^\top\mathbf{A})=1$

Inner Product: (in $\mathbb{R}^D$) $\langle\mathbf{x},\mathbf{y}\rangle=\mathbf{x}^\top\mathbf{y}=\sum_{i=1}^N\mathbf{x}_i\mathbf{y}_i$. • $\langle\mathbf{x}\pm\mathbf{y},\mathbf{x}\pm\mathbf{y}\rangle=\langle\mathbf{x},\mathbf{x}\rangle\pm2\langle\mathbf{x},\mathbf{y}\rangle+\langle\mathbf{y},\mathbf{y}\rangle$ • $\langle\mathbf{x},\mathbf{y}+\mathbf{z}\rangle=\langle\mathbf{x},\mathbf{y}\rangle+\langle\mathbf{x},\mathbf{z}\rangle$ • $\langle\mathbf{x}+\mathbf{y},\mathbf{z}\rangle=\langle\mathbf{x},\mathbf{z}\rangle+\langle\mathbf{y},\mathbf{z}\rangle$ • $\langle\mathbf{x},\mathbf{y}\rangle=\|\mathbf{x}\|_2\cdot\|\mathbf{y}\|_2\cdot\cos(\theta)$ • If $\mathbf{y}$ is a unit vector then $\langle\mathbf{x},\mathbf{y}\rangle$ projects $\mathbf{x}$ onto $\mathbf{y}$

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

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

Gram-Schmidt: $\{\mathbf{w}_i\}_i$ non-orthogonal basis. $\mathbf{v}_n=\mathbf{w}_n-\sum_{i=1}^{n-1}\frac{\langle\mathbf{v}_i,\mathbf{w}_n\rangle}{\langle\mathbf{v}_i,\mathbf{v}_i\rangle}\mathbf{v}_i$ results in $\{\mathbf{v}_i\}_i$ an orthogonal basis

1.2 Norms

• $\|\mathbf{x}\|_0=|\{i|x_i\neq0\}|$ • $\|\mathbf{x}\|_2=\sqrt{\sum_{i=1}^N\mathbf{x}_i^2}=\sqrt{\langle\mathbf{x},\mathbf{x}\rangle}$ • $\|\mathbf{x}\|_p=\left(\sum_{i=1}^N|x_i|^p\right)^{\frac{1}{p}}$ • $\mathbf{M}\in\mathbb{R}^{m\times n}$, $\|\mathbf{M}\|_F=\sqrt{\sum_{i=1}^m\sum_{j=1}^n\mathbf{m}_{i,j}^2}=\sqrt{\sum_{i=1}^{\min\{m,n\}}\sigma_i^2}$ • $\|\mathbf{M}\|_1=\sum_{i,j}|m_{i,j}|$ • $\|\mathbf{M}\|_2=\sigma_{\max}(\mathbf{M})$ • $\|\mathbf{M}\|_p=\max_{\mathbf{v}\neq0}\frac{\|\mathbf{M}\mathbf{v}\|_p}{\|\mathbf{v}\|_p}$ • $\|\mathbf{M}\|_\star=\sum_{i=1}^{\min(m,n)}\sigma_i$

1.3 Derivatives

• $\frac{\partial}{\partial\mathbf{x}}(\mathbf{b}^\top\mathbf{x})=\frac{\partial}{\partial\mathbf{x}}(\mathbf{x}^\top\mathbf{b})=\mathbf{b}$ • $\frac{\partial}{\partial\mathbf{x}}(\mathbf{x}^\top\mathbf{x})=2\mathbf{x}$ • $\frac{\partial}{\partial\mathbf{x}}(\mathbf{x}^\top\mathbf{A}\mathbf{x})=(\mathbf{A}^\top+\mathbf{A})\mathbf{x}\overset{\text{if }\mathbf{A}\text{ sym.}}{=}2\mathbf{A}\mathbf{x}$ • $\frac{\partial}{\partial\mathbf{x}}(\mathbf{b}^\top\mathbf{A}\mathbf{x})=\mathbf{A}^\top\mathbf{b}$ • $\frac{\partial}{\partial\mathbf{X}}(\mathbf{c}^\top\mathbf{X}\mathbf{b})=\mathbf{c}\mathbf{b}^\top$ • $\frac{\partial}{\partial\mathbf{X}}(\mathbf{c}^\top\mathbf{X}^\top\mathbf{b})=\mathbf{b}\mathbf{c}^\top$ • $\frac{\partial}{\partial\mathbf{x}}(\|\mathbf{x}-\mathbf{b}\|_2)=\frac{\mathbf{x}-\mathbf{b}}{\|\mathbf{x}-\mathbf{b}\|_2}$ • $\frac{\partial}{\partial\mathbf{x}}(\|\mathbf{x}\|_2^2)=\frac{\partial}{\partial\mathbf{x}}(\mathbf{x}^\top\mathbf{x})=2\mathbf{x}$ • $\frac{\partial}{\partial\mathbf{X}}(\|\mathbf{X}\|_F^2)=2\mathbf{X}$

1.4 Eigenvalue / -vectors

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

1. solve $\det(\mathbf{A}-\lambda\mathbf{I})\overset{!}{=}0$ resulting in $\{\lambda_i\}_i$
2. $\forall\lambda_i$: solve $(\mathbf{A}-\lambda_i\mathbf{I})\mathbf{x}_i=\mathbf{0}$, $\mathbf{x}_i$ is the i-th eigenvector.
3. (opt.) normalize eigenvector q_i: $q_i^{\text{norm}}=\frac{1}{\|q_i\|_2}q_i$.

1.5 Eigendecomposition

• $\mathbf{A}\in\mathbb{R}^{N\times N}$ then $\mathbf{A}=\mathbf{Q}\Lambda\mathbf{Q}^{-1}$ with $\mathbf{Q}\in\mathbb{R}^{N\times N}$. • if all eigenvalues nonzero: $\mathbf{A}^{-1}=\mathbf{Q}\Lambda^{-1}\mathbf{Q}^{-1}$ and $(\Lambda^{-1})_{i,i}=\frac{1}{\lambda_i}$ • if $\mathbf{A}$ symmetric: $A=\mathbf{Q}\Lambda\mathbf{Q}^\top$ (and $\mathbf{Q}$ is orthogonal).

1.6 Probability / Statistics

• $P(x):=Pr[X=x]:=\sum_{y\in Y}P(x,y)$ • $P(x|y):=Pr[X=x|Y=y]:=\frac{P(x,y)}{P(y)}$, if $P(y)>0$ • $\forall y\in Y:\sum_{x\in X}P(x|y)=1$ (property for any fixed y) • $P(x,y)=P(x|y)P(y)$ • $P(x|y)=\frac{P(y|x)P(x)}{P(y)}$ (Bayes' rule) • $P(x|y)=P(x)\Leftrightarrow P(y|x)=P(y)$ (iff X, Y independent) • $P(x_1,\ldots,x_n)=\prod_{i=1}^nP(x_i)$ (iff IID)

2 Dimensionality Reduction / PCA

$\mathbf{X}\in\mathbb{R}^{D\times N}$. N observations, K properties. Target: $\tilde{\mathbf{X}}\in\mathbb{R}^{K\times N}$. 1. Empirical Mean: $\overline{\mathbf{x}}=\frac{1}{N}\sum_{n=1}^N\mathbf{x}_n$ 2. Center Data: $\overline{\mathbf{X}}=\mathbf{X}-[\overline{\mathbf{x}},\ldots,\overline{\mathbf{x}}]=\mathbf{X}-\mathbf{M}$ 3. Cov. Matrix: $\Sigma=\frac{1}{N}\sum_{n=1}^N(\mathbf{x}_n-\overline{\mathbf{x}})(\mathbf{x}_n-\overline{\mathbf{x}})^\top=\frac{1}{N}\overline{\mathbf{X}}\overline{\mathbf{X}}^\top$ 4. Eigenvalue Decomposition: $\Sigma=\mathbf{U}\Lambda\mathbf{U}^\top$, sort eigenvalues (and eigenvectors) in descending order 5. Select $K<D$, keep only the first K eigenvalues and corresponding eigenvectors $\Rightarrow\mathbf{U}_K,\lambda_K$ 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 $\mathbf{U}_k,\overline{\mathbf{Z}}_K,\overline{\mathbf{x}}$.

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

3 SVD

• $\mathbf{A}=\mathbf{U}\mathbf{D}\mathbf{V}^\top=\sum_{k=1}^{\text{rank}(\mathbf{A})}d_{k,k}u_k(v_k)^\top$ • $\mathbf{A}\in\mathbb{R}^{N\times P},\mathbf{U}\in\mathbb{R}^{N\times N},\mathbf{D}\in\mathbb{R}^{N\times P},\mathbf{V}\in\mathbb{R}^{P\times P}$ • $\mathbf{U}^\top\mathbf{U}=I=\mathbf{V}^\top\mathbf{V}$ ($\mathbf{U},\mathbf{V}$ columns are 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, i.e. the square roots of the eigenvalues ($\mathbf{A}^\top\mathbf{A}$ and $\mathbf{A}\mathbf{A}^\top$ have the same eigenvalues) • $(\mathbf{D}^{-1})_{i,i}=\frac{1}{\mathbf{D}_{i,i}}$ ($\mathbf{D}\in\mathbb{R}^{N\times P}\to\mathbf{D}^{-1}\in\mathbb{R}^{P\times N}$, i.e. don't forget to transpose) • Missing columns in $\mathbf{U}$ are basis of null$(A^\top)$ and in $\mathbf{V}$ are basis of null(A). Calculate: $\mathbf{A}^\top\mathbf{u}=\mathbf{0}$ or $\mathbf{A}\mathbf{v}=\mathbf{0}$ for $\mathbf{u}$ or $\mathbf{v}$.

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: $\mathbf{U}=\mathbf{A}\mathbf{V}\mathbf{D}^{-1}$. Can be checked by calculating the eigenvectors of $\mathbf{A}\mathbf{A}^\top$. 5. normalize each column of $\mathbf{U}$ and $\mathbf{V}$.

3.1 Low-Rank approximation

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

$\|\mathbf{A}-\tilde{\mathbf{A}}\|_F=\sqrt{\sum_{i>K}\sigma_i^2}=\sqrt{\sum_{i>K}\lambda_i}$, $\|\mathbf{A}-\tilde{\mathbf{A}}\|_2=\sigma_{K+1}$

4 *K*-means Algorithm

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 $\mathbf{U}=[\mathbf{u}_1,\ldots,\mathbf{u}_K]$ (usually u.a.r.) 2. Assign data points to clusters. $k^\star(\mathbf{x}_n)=\arg\min_k\{\|\mathbf{x}_n-\mathbf{u}_k\|_2\}$ returns cluster $k^\star$, whose centroid $\mathbf{u}_{k^\star}$ is closest to data point $\mathbf{x}_n$. Set $\mathbf{z}_{k^\star,n}=1$, and for $l\neq k^\star$ $\mathbf{z}_{l,n}=0$. 3. Update centroids: $\mathbf{u}_k=\frac{\sum_{n=1}^Nz_{k,n}\mathbf{x}_n}{\sum_{n=1}^Nz_{k,n}}$. 4. Repeat from step 2, stops if $\|\mathbf{Z}-\mathbf{Z}^{\text{new}}\|_0=\|\mathbf{Z}-\mathbf{Z}^{\text{new}}\|_F^2=0$.

4.1 Clustering Stability

• dist. between clust. (same data): $d(C,C'):=\min_\Pi\frac{1}{2}\|Z-\Pi(Z')\|_F^2$, $\Pi(Z')$ = row perm. of Z' • arbitrary sets $\mathbf{X},\mathbf{X}'$ of size N,N': $r:=\frac{1}{N'}\min_\Pi\{\sum_{n=1}^{N'}\mathbb{I}_{\{\Pi(\phi(x_n'))\neq z_n'\}}\}$ (ϕ: multi-class classifier trained on $(\mathbf{X},\mathbf{Z})$) • for K clusters: stability $:=1-\frac{r}{r_{rand}}$ (1 good, 0 bad), rand. clust. of equal size: $r_{rand}=\frac{K-1}{K}$.

5 Gaussian Mixture Models (GMM)

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

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

Assignment variable (generative model): $z_k\in\{0,1\}$, $\sum_{k=1}^Kz_k=1$, $\Pr(z_k=1)=\pi_k\Leftrightarrow p(\mathbf{z})=\prod_{k=1}^K\pi_k^{z_k}$

Complete data distribution: $p_\theta(\mathbf{x},\mathbf{z})=\prod_{k=1}^K(\pi_kp_{\theta_k}(\mathbf{x}))^{z_k}$

Posterior Probabilities:

$\Pr(z_k=1|\mathbf{x})=\frac{\Pr(z_k=1)p(\mathbf{x}|z_k=1)}{\sum_{l=1}^K\Pr(z_l=1)p(\mathbf{x}|z_l=1)}=\frac{\pi_kp_{\theta_k}(\mathbf{x})}{\sum_{l=1}^K\pi_lp_{\theta_l}(\mathbf{x})}$

Likelihood of observed data X: $p_\theta(\mathbf{X})=\prod_{n=1}^Np_\theta(\mathbf{x}_n)=\prod_{n=1}^N\left(\sum_{k=1}^K\pi_kp_{\theta_k}(\mathbf{x}_n)\right)$

MLE: $\arg\max_\theta\sum_{n=1}^N\log\left(\sum_{k=1}^K\pi_kp_{\theta_k}(\mathbf{x}_n)\right)$

$\log\left(\sum_{k=1}^K\frac{q_k\pi_kp_{\theta_k}(\mathbf{x}_n)}{q_k}\right)\geq\sum_{k=1}^Kq_k[\log p_{\theta_k}(\mathbf{x}_n)+\log\pi_k-\log q_k]$ with $\sum_{k=1}^Kq_k=1$ by Jensen. Lagrangian and get q_k as below.

5.1 Expectation-Maximization (EM) for GMM

1. Initialize $\pi_k^{(0)},\mu_k^{(0)},\Sigma_k^{(0)}$ for $k=1,\ldots,K$ and $t=1$.
2. E-Step: $\Pr[z_{k,n}=1|\mathbf{x}_n]=q_{k,n}=\frac{\pi_k^{(t-1)}\mathcal{N}(\mathbf{x}_n|\mu_k^{(t-1)},\Sigma_k^{(t-1)})}{\sum_{j=1}^K\pi_j^{(t-1)}\mathcal{N}(\mathbf{x}_n|\mu_j^{(t-1)},\Sigma_j^{(t-1)})}$
3. M-Step: $\mu_k^{(t)}:=\frac{\sum_{n=1}^Nq_{k,n}\mathbf{x}_n}{\sum_{n=1}^Nq_{k,n}}$ & $\pi_k^{(t)}:=\frac{1}{N}\sum_{n=1}^Nq_{k,n}$ & $\Sigma_k^{(t)}=\frac{\sum_{n=1}^Nq_{k,n}(\mathbf{x}_n-\mu_k^{(t)})(\mathbf{x}_n-\mu_k^{(t)})^\top}{\sum_{n=1}^Nq_{k,n}}$
4. Repeat from (2.) with $t=t+1$ if not $\|\log p(\mathbf{X}|\pi^{(t)},\mu^{(t)},\Sigma^{(t)})-\log p(\mathbf{X}|\pi^{(t-1)},\mu^{(t-1)},\Sigma^{(t-1)})\|<\varepsilon$

5.2 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), $K-1$: # 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.

6 Word Embeddings

Distributional Model: $p_\theta(w|w')=\Pr[w\text{ occurs close to }w']$

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: $w \mapsto (\mathbf{x}_w, b_w) \in \mathbb{R}^{D+1}$
$p_\theta(w|w') = \frac{\exp[\langle \mathbf{x}_w,\mathbf{x}_{w'}\rangle + b_w]}{\sum_{v\in V}\exp[\langle \mathbf{x}_v,\mathbf{x}_{w'}\rangle + b_v]}$. Modifications: • split vocab in main vocab V, context vocab C: $\log p_\theta(w|w') = \langle y_w, x_{w'}\rangle + b_w$, word embed. y_w, context embed. $x_{w'}$ • use GloVe objective

6.1 GloVe (Weighted Square Loss)

Co-occurence Matrix: $\mathbf{N} = (n_{ij}) \in \mathbb{R}^{|V|\cdot|C|} \leftrightarrow \#w_i$ in c'txt 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 distribution $\rightarrow$ two-sided loss function
SGD: 1. $\mathbf{x}_i^{new} \leftarrow \mathbf{x}_i + 2\eta f(n_{ij})(\log n_{ij} - \langle \mathbf{x}_i,\mathbf{y}_j\rangle)\mathbf{y}_j$
2. $\mathbf{y}_j^{new} \leftarrow \mathbf{y}_j + 2\eta f(n_{ij})(\log n_{ij} - \langle \mathbf{x}_i,\mathbf{y}_j\rangle)\mathbf{x}_i$

7 Non-Negative Matrix Factorization (NMF) / pLSA

Context Model: $p(w|d) = \sum_{z=1}^{K} p(w|z)p(z|d)$
Conditional independence assumption (∗): $p(w|d) = \sum_z p(w,z|d) = \sum_z p(w|d,z)p(z|d) \stackrel{*}{=} \sum_z p(w|z)p(z|d)$
Symmetric parameterization: $p(w,d) = \sum_z p(z)p(w|z)p(d|z)$

7.1 EM for pLSA:

1. Log-Likelihood: $L(\mathbf{U},\mathbf{V}) = \sum_{i,j} x_{i,j}\log p(w_j|d_i) = \sum_{(i,j)\in X}\log\sum_{z=1}^{K} p(w_j|z)p(z|d_i)$
2. E-Step (optimal q): $q_{zij} = \frac{p(w_j|z)p(z|d_i)}{\sum_{k=1}^{K} p(w_j|k)p(k|d_i)} := \frac{v_{zj}u_{zi}}{\sum_{k=1}^{K} v_{kj}u_{ki}}$
3. M-Steps: $p(z|d_i) = \frac{\sum_j x_{ij}q_{zij}}{\sum_j x_{ij}}$ & $p(w_j|z) = \frac{\sum_i x_{ij}q_{zij}}{\sum_{i,l} x_{il}q_{zil}}$

7.2 NMF Algorithm for quadratic cost function

• $\mathbf{X} \in \mathbb{Z}_{\geq 0}^{N\times M}$ • NMF: $\mathbf{X} \approx \mathbf{U}^\top\mathbf{V}, x_{ij} = \sum_z u_{zi}v_{zj} = \langle \mathbf{u}_i, \mathbf{v}_j\rangle$
$\min_{\mathbf{U},\mathbf{V}} J(\mathbf{U},\mathbf{V}) = \frac{1}{2}\|\mathbf{X} - \mathbf{U}^\top\mathbf{V}\|_F^2$ s.t. $\forall i,j,z\ u_{zi}, v_{zj} \geq 0$
1. init: $\mathbf{U},\mathbf{V} = rand()$ 2. repeat for *maxIters*: 3. update $\mathbf{U}$: $(\mathbf{V}\mathbf{V}^\top)\mathbf{U} = \mathbf{V}\mathbf{X}^\top$ 4. project $u_{zi} = \max\{0, u_{zi}\}$ 5. update $\mathbf{V}$: $(\mathbf{U}\mathbf{U}^\top)\mathbf{V} = \mathbf{U}\mathbf{X}$ 6. project $v_{zj} = \max\{0, v_{zj}\}$

8 Convolutional Neural Networks

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 classification; $y_1 = \mathrm{P}[Y=1|\mathbf{x}] = \frac{1}{1+\exp[-\langle \mathbf{w}_1^L, \mathbf{x}^{L-1}\rangle]}$, multiclass; $y_k = \mathrm{P}[Y=k|\mathbf{x}] = \frac{\exp[\langle \mathbf{w}_k^L,\mathbf{x}^{L-1}\rangle]}{\sum_{m=1}^{K}\exp[\langle \mathbf{w}_m^L,\mathbf{x}^{L-1}\rangle]}$. **Loss function** $l(y,\hat{y})$: squared loss; $\frac{1}{2}(y-\hat{y})^2$, cross-entropy loss; $-y\log\hat{y} - (1-y)\log(1-\hat{y})$.

8.1 Neural Networks for Images

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

9 Optimization

9.1 Coordinate Descent (update the d-th coord. per step)

1. init: $\mathbf{x}^{(0)} \in \mathbb{R}^D$ 2. for $t = 0$ to *maxIter*: 3. sample u.a.r. $d \sim \{1,\ldots,D\}$ 4. $u^\star = \arg\min_{u\in\mathbb{R}} f(x_1^{(t)},..,x_{d-1}^{(t)},u,x_{d+1}^{(t)},..,x_D^{(t)})$ 5. $\mathbf{x}_d^{(t+1)} = u^\star$ and $\mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t)}$ for $i \neq d$

9.2 Gradient Descent (or Deepest Descent)

Gradient: $\nabla f(\mathbf{x}) := \left(\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_1},\ldots,\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}_D}\right)^\top$ 1. init: $\mathbf{x}^{(0)} \in \mathbb{R}^D$ 2. for $t = 0$ to *maxIter*: $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \gamma\nabla f(\mathbf{x}^{(t)})$, usually $\gamma \approx \frac{1}{t}$

9.3 Stochastic Gradient Descent (SGD)

Assume **Additive Objective**; $f(x) = \frac{1}{N}\sum_{n=1}^{N} f_n(x)$ 1. init: $\mathbf{x}^{(0)} \in \mathbb{R}^D$ 2. for $t = 0$ to *maxIter*: 3. sample u.a.r. $n \sim \{1,\ldots,N\}$ 4. $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \gamma\nabla f_n(\mathbf{x}^{(t)})$, usually stepsize $\gamma \approx \frac{1}{t}$.

9.4 Projected Gradient Descent (Constrained Opt.)

minimize $f(x), x \in Q$ (constraint). **Project** x onto Q: $P_Q(\mathbf{x}) = \arg\min_{y\in Q}\|\mathbf{y}-\mathbf{x}\|$, **Projected Gradient Update**: $\mathbf{x}^{(t+1)} = P_Q[\mathbf{x}^{(t)} - \gamma\nabla f(\mathbf{x}^{(t)})]$, $\mathbf{x}^{(t+1)}$ is unique if Q convex.

9.5 Lagrangian Multipliers

Minimize $f(\mathbf{x})$ s.t. $g_i(\mathbf{x}) \leq 0,\ i = 1,..,m$ (**inequality constr.**) and $h_i(\mathbf{x}) = \mathbf{a}_i^\top\mathbf{x} - b_i = 0,\ i = 1,..,p$ (**equality constraint**)
Lagrangian: $L(\mathbf{x},\lambda,\nu) := f(\mathbf{x}) + \sum_{i=1}^{m}\lambda_i g_i(\mathbf{x}) + \sum_{i=1}^{p}\nu_i h_i(\mathbf{x})$
Dual function: $D(\lambda,\nu) := \inf_\mathbf{x} L(\mathbf{x},\lambda,\nu) \in \mathbb{R}$
Dual Problem: $\max_{\lambda,\nu} D(\lambda,\nu)$ s.t. $\lambda \geq \mathbf{0}$. Note: $\max_{\lambda,\nu} D(\lambda,\nu) \leq \min_\mathbf{x} f(\mathbf{x})$, equality if *dom f* and f convex

9.6 Convex Optimization

$f : \mathbb{R}^D \rightarrow \mathbb{R}$ is convex, if *dom f* is a convex set, and if $\forall \mathbf{x},\mathbf{y} \in$ *dom f*, and for $0 \leq \alpha \leq 1$: $f(\alpha\mathbf{x} + (1-\alpha)\mathbf{y}) \leq \alpha f(\mathbf{x}) + (1-\alpha)f(\mathbf{y})$. local=global min, **Convergence**: $f(\mathbf{x}^{(t)}) - f(\mathbf{x}^*) \leq \frac{c}{t}$.
Subgradient $g \in \mathbb{R}^D$ of f at $\mathbf{x}$: $f(\mathbf{y}) \geq f(\mathbf{x}) + g^\top(\mathbf{y}-\mathbf{x})\ \forall\mathbf{y}$

10 Sparse Coding

10.1 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 $|z_i| > \varepsilon$ 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.

10.2 Overcomplete Basis

$\mathbf{U} \in \mathbb{R}^{D\times L}$ for # atoms $= L > D = \dim(\text{data})$. Decoding involved $\rightarrow$ add constraint $\mathbf{z}^\star \in \arg\min_\mathbf{z}\|\mathbf{z}\|_0$ s.t. $\mathbf{x} = \mathbf{U}\mathbf{z}$. NP-hard $\rightarrow$ approximate with 1-norm (convex) or with MP.
Coherence • $m(\mathbf{U}) = \max_{i,j:i\neq j}|\mathbf{u}_i^\top\mathbf{u}_j|$ • $m(\mathbf{B}) = 0$ 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: $\mathbf{z}^\star \in \arg\min_\mathbf{z}\|\mathbf{x} - \mathbf{U}\mathbf{z}\|_2$, s.t. $\|\mathbf{z}\|_0 \leq K$ 1. init: $z \leftarrow 0, r \leftarrow x$ 2. while $\|\mathbf{z}\|_0 < K$ do 3. select atom with smallest angle $i^\star = \arg\max_i|\langle \mathbf{u}_i,\mathbf{r}\rangle|$ 4. update coefficients: $z_{i^\star} \leftarrow z_{i^\star} + \langle \mathbf{u}_{i^\star},\mathbf{r}\rangle$ 5. update residual: $\mathbf{r} \leftarrow \mathbf{r} - \langle \mathbf{u}_{i^\star},\mathbf{r}\rangle\mathbf{u}_{i^\star}$.
Exact recovery when: $K < 1/2(1 + 1/m(\mathbf{U}))$
Compressive Sensing • $\mathbf{x} \in \mathbb{R}^D$, 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 $\mathbf{x} = \mathbf{U}\mathbf{z}$

10.3 Dictionary Learning

Adapt the dictionary to signal characteristics. Objective: $(\mathbf{U}^\star,\mathbf{Z}^\star) \in \arg\min_{\mathbf{U},\mathbf{Z}}\|\mathbf{X} - \mathbf{U}\cdot\mathbf{Z}\|_F^2$ 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 ($\mathbf{z}_n^{t+1} \in \arg\min_\mathbf{z}\|\mathbf{z}\|_0$ s.t. $\|\mathbf{x}_n - \mathbf{U}^t\mathbf{z}\|_2 \leq \sigma\|\mathbf{x}_n\|_2$) 2. Dict update step: $\mathbf{U}^{t+1} \in \arg\min_\mathbf{U}\|\mathbf{X} - \mathbf{U}\mathbf{Z}^{t+1}\|_F^2$, subj to $\forall l \in [L]$: $\|\mathbf{u}_l\|_2 = 1$. (set $\mathbf{U} = [\mathbf{u}_1^t\cdots\mathbf{u}_l\cdots\mathbf{u}_L^t]$, $\min_{\mathbf{u}_l}\|\mathbf{X} - \mathbf{U}\mathbf{Z}^{t+1}\|_F^2 = \min_{\mathbf{u}_l}\|\mathbf{R}_l^t - \mathbf{u}_l(\mathbf{z}_l^{t+1})^\top\|_F^2$ with $\mathbf{R}_l^t = \tilde{\mathbf{U}}\Sigma\tilde{\mathbf{V}}^\top$ by $\mathbf{u}_l^* = \tilde{\mathbf{u}}_1$)

11 Robust PCA

• Idea: Approximate $\mathbf{X}$ with $\mathbf{L}+\mathbf{S}$, $\mathbf{L}$ is low-rank, $\mathbf{S}$ is sparse.
• $\min_{\mathbf{L},\mathbf{S}}\operatorname{rank}(\mathbf{L}) + \mu\|\mathbf{S}\|_0$, s. t. $\mathbf{L}+\mathbf{S} = \mathbf{X}$. As non-convex, change to $\min_{\mathbf{L},\mathbf{S}}\|\mathbf{L}\|_\star + \lambda\|\mathbf{S}\|_1$ (*not* the same in general)
• Perfect reconstruction is *not* possible if $\mathbf{S}$ is low-rank, $\mathbf{L}$ is sparse, or $\mathbf{X}$ is low-rank *and* sparse. Formally coherence: $\|\mathbf{U}^\top\mathbf{e}_i\|^2 \leq \frac{\nu r}{n}$, $\|\mathbf{V}^\top\mathbf{e}_i\|^2 \leq \frac{\nu r}{n}$, $\|\mathbf{U}\mathbf{V}^\top\|_{ij}^2 \leq \frac{\nu r}{n^2}$: $\mathbf{L} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$

11.1 Dual Ascent (Gradient Method for Dual Problem)

$\lambda^{t+1} = \lambda^t + \eta\nabla D(\lambda^t)$, $\nabla D(\lambda) = \mathbf{A}\mathbf{x}^* - \mathbf{b}$ for $\mathbf{x}^* \in \arg\min_\mathbf{x}\mathscr{L}(\mathbf{x},\lambda)$ **Dual Decomposition for Dual Ascent**: $\mathbf{x}_i^{t+1} := \arg\min_{\mathbf{x}_i}\mathscr{L}_i(\mathbf{x}_i,\lambda^t)$; $\lambda^{t+1} := \lambda^t + \eta^t\left(\sum_{i=1}^{N}\mathbf{A}_i\mathbf{x}_i^{t+1} - \mathbf{b}\right)$

11.2 Alternating Direction Method of Multipliers (ADMM)

$\min_{\mathbf{x}_1,\mathbf{x}_2} f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2)$ s. t. $\mathbf{A}_1\mathbf{x}_1 + \mathbf{A}_2\mathbf{x}_2 = \mathbf{b}$, f_1, f_2 convex • Augmented Lagrangian: $L_p(\mathbf{x}_1,\mathbf{x}_2,\nu) = f_1(\mathbf{x}_1) + f_2(\mathbf{x}_2) + \nu^\top(\mathbf{A}_1\mathbf{x}_1 + \mathbf{A}_2\mathbf{x}_2 - \mathbf{b}) + \frac{p}{2}\|\mathbf{A}_1\mathbf{x}_1 + \mathbf{A}_2\mathbf{x}_2 - \mathbf{b}\|_2^2$
• ADMM: $\mathbf{x}_1^{(t+1)} := \arg\min_{\mathbf{x}_1} L_p(\mathbf{x}_1,\mathbf{x}_2^{(t)},\nu^{(t)})$, $\mathbf{x}_2^{(t+1)} := \arg\min_{\mathbf{x}_2} L_p(\mathbf{x}_1^{(t+1)},\mathbf{x}_2,\nu^{(t)})$, $\nu^{(t+1)} := \nu^{(t)} + p(\mathbf{A}_1\mathbf{x}_1^{(t+1)} + \mathbf{A}_2\mathbf{x}_2^{(t+1)} - \mathbf{b})$ • ADMM for RPCA: $f_1(\mathbf{L}) = \|\mathbf{L}\|_\star$, $f_2(\mathbf{S}) = \lambda\|\mathbf{S}\|_1$, $\mathbf{A}_1\mathbf{x}_1 + \mathbf{A}_2\mathbf{x}_2 = \mathbf{b}$ becomes $\mathbf{L}+\mathbf{S} = \mathbf{X}$, therefore $L_p(\mathbf{L},\mathbf{S},\nu) = \|\mathbf{L}\|_* + \nu\|\mathbf{S}\|_1 + \langle \nu, \operatorname{vec}(\mathbf{L}+\mathbf{S}-\mathbf{X})\rangle + \frac{p}{2}\|\mathbf{L}+\mathbf{S}-\mathbf{X}\|_F^2$