Sparse Coding and Dictionary Learning

Pardis Noorzad

Department of Computer Engineering and IT
Amirkabir University of Technology

Mehr 1390

Outline

The Setting

- We have the linear inverse problem

$$\mathbf{b} = \mathbf{A}\mathbf{x} \tag{1}$$

where $\mathbf{b} \in \mathbb{R}^m, \mathbf{x} \in \mathbb{R}^p, \mathbf{A} \in \mathbb{R}^{m \times p}$; $\mathbf{x}$ is unknown.

- Many problems in ML are linear inverse problems, for e.g.,
  - regression and classification: $\mathbf{y} = \mathbf{X}\mathbf{a}$, $\mathbf{a}$ is unknown;
  - sparse coding: $\mathbf{x} = \mathbf{D}\mathbf{a}$, $\mathbf{a}$ is unknown;
  - dictionary learning: $\mathbf{x} = \mathbf{D}\mathbf{a}$, both $\mathbf{a}$ and $\mathbf{D}$ are unknown.

Solution

Take I

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} \tag{2}$$

- What's the problem here?
- $\mathbf{A}$ is almost never **invertible** in our problems:
  - needs to be square
  - needs to have full column rank

Ill-posedness

- **Case I**
  - If $m = p$ or $m > p$, we say that the system of equations is **overdetermined**.
  - In this case, the solution to (1) does **not exist**.
- **Case II**
  - If $m < p$, the system is **underdetermined**,
  - and there exists **infinitely many** solutions.

Solution

Case I, Take II

- Instead of the equations, $\mathbf{b} = \mathbf{A}\mathbf{x}$, only minimize the residual,
$$\min_{\mathbf{x}} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 \tag{3}$$
- where (3) yields an approximate solution to (1), i.e.,
$$\mathbf{x} = (\mathbf{A}^\mathsf{T}\mathbf{A})^{-1}\mathbf{A}^\mathsf{T}\mathbf{b}$$
- The solution exists if $\mathbf{A}^\mathsf{T}\mathbf{A}$ is invertible, i.e.,
  - $\mathbf{A}$ must have full column rank
  - o.w., (3) is no better than (1), which is the case for **Case II**.

Schemes

- Regularize to incorporate a priori assumptions about the **size** and **smoothness** of the solution.
  - for e.g. by using the ℓ_2 norm as the measure of size
- Regularization is done using one of the following schemes:

$$\min \frac{1}{2}\|\mathbf{b}-\mathbf{A}\mathbf{x}\|_2^2 \ \text{ s.t. } \ \|\mathbf{x}\|_1 \leq T \tag{4}$$

$$\min \|\mathbf{x}\|_1 \ \text{ s.t. } \ \|\mathbf{b}-\mathbf{A}\mathbf{x}\|_2^2 \leq \epsilon \tag{5}$$

$$\min \frac{1}{2}\|\mathbf{b}-\mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1 \ \text{ (Lagrangian form)} \tag{6}$$

- Note that the schemes are equivalent in theory but not in practice, since relations between T, ϵ, and λ are unknown.

Solution

Take III

- Regularize, i.e.,

$$\min \frac{1}{2}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_2^2 \quad (7)$$

- now (7) has the **unique** solution,

$$\mathbf{x}^* = (\mathbf{A}^\mathsf{T}\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^\mathsf{T}\mathbf{b}.$$

- Note that $\mathbf{A}^\mathsf{T}\mathbf{A} + \lambda\mathbf{I}$ is nonsingular even when $\mathbf{A}^\mathsf{T}\mathbf{A}$ is singular.

When $m \ll p$

- Standard procedure is to constrain with **sparsity**.
- To measure sparsity, we introduce the ℓ_0 quasi-norm,

$$\|\mathbf{x}\|_0 = \#\{i : x_i \neq 0\}. \tag{8}$$

- The problem becomes,

$$\min \|\mathbf{x}\|_0 \;\; \text{s.t.} \;\; \mathbf{b} = \mathbf{A}\mathbf{x}. \tag{9}$$

- Because of the **combinatorial** aspect of the ℓ_0 norm, the problem (9) is intractable.

Solution

Take I: Convex Relaxation

- **Basis pursuit** (Chen et al., 1995)

$$\min \|\mathbf{x}\|_1 \ \text{ s.t. } \ \mathbf{b} = \mathbf{A}\mathbf{x}. \tag{10}$$

- (10) is a linear program for which a tractable algorithm exists, in this case:
  - primal-dual interior point method
  - solves the **approximate** problem, **exactly**
- To allow for some noise, Chen et al. proposed **basis pursuit de-noising**

$$\frac{1}{2}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1. \tag{11}$$

Solution

Take II: Greed

- Greedy algorithms like **matching pursuit** (Mallat and Zhang, 1993) solve the following problem **approximately**.

$$\min \frac{1}{2}\|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 \text{ s.t. } \|\mathbf{x}\|_0 \leq s. \tag{12}$$

- where s is the desired sparsity of the solution.
- (12) can't be solved exactly since it is NP-hard.
- However, greedy methods like MP, OMP, LARS, etc., can result in good local optima.

Power family of penalties

ℓ_p norms raised to the pth power

$$\|\mathbf{x}\|_p^p = \left(\sum_i |x_i|^p \right) \tag{13}$$

- For $1 \leq p < \infty$, (13) is convex.
- $0 < p \leq 1$, is the range of p useful for measuring sparsity.

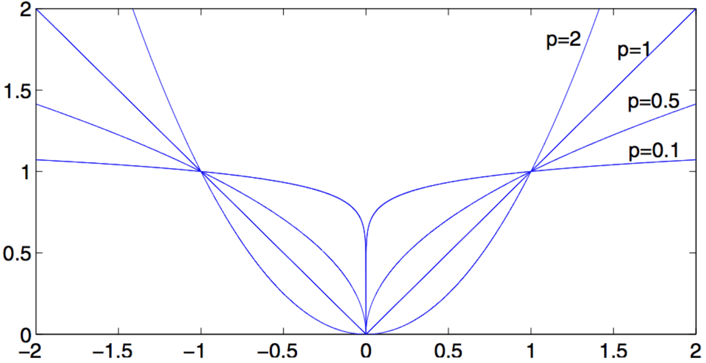

Figure: As p goes to 0, $|x|^p$ becomes the indicator function and $|\mathbf{x}|^p$ becomes a count of the nonzeros in $\mathbf{x}$ (Bruckstein et al., 2009).

Sparse representation

$$\mathbf{x} = \mathbf{D}\mathbf{a}$$

- The dictionary should be redundant:
  dimensionality of input $\ll$ number of columns of the dictionary
- Use the algorithms that we talked about, e.g., OMP or LARS.

Unsupervised feature learning

Application to image classification

$$\mathbf{x} = \mathbf{D}\mathbf{a}$$

- An example is the recent work by Coates and Ng (2011).
  - where $\mathbf{x}$ is the input vector
  - could be a vectorized image patch, or a SIFT descriptor
  - $\mathbf{a}$ is the **higher-dimensional sparse representation** of $\mathbf{x}$
  - $\mathbf{D}$ is usually learned—we'll talk about it later

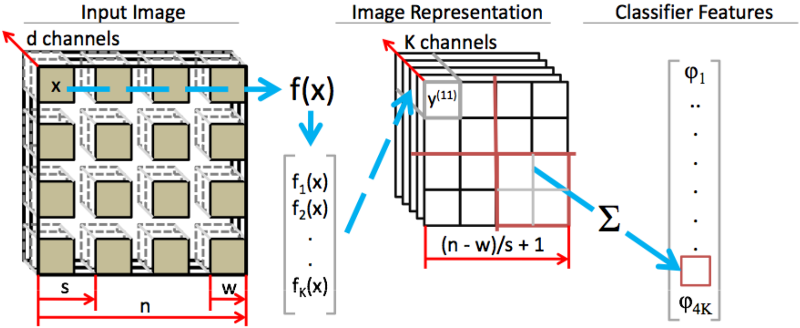

Figure: Image classification (Coates et al., 2011).

Multiclass classification

(Wright et al., 2009)

$$\mathbf{x} = \mathbf{D}\mathbf{a}$$

- $\mathbf{x}$ is a test sample
- $\mathbf{D} = [\mathbf{x}^1|\mathbf{x}^2|\ldots|\mathbf{x}^p]$ contains training samples as its columns
- $\delta_i(\mathbf{a})$ gives a new vector whose nonzero entries are those in $\mathbf{a}$ associated with class i

$$i^* = \arg\min \|\mathbf{x} - \mathbf{D}\delta_i(\mathbf{a})\|_2^2.$$

Dictionary learning as matrix factorization

$$\min_{\substack{\mathbf{D}\in\mathcal{D}\\ \mathbf{A}\in\mathcal{A}}} \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{2}\|\mathbf{x}^i-\mathbf{D}\mathbf{a}^i\|_2^2+\lambda\Omega(\mathbf{a}^i)\right]=$$

$$\min_{\substack{\mathbf{D}\in\mathcal{D}\\ \mathbf{A}\in\mathcal{A}}} \left[\frac{1}{2}\|\mathbf{X}-\mathbf{D}\mathbf{A}\|_F^2+\lambda\Omega^{'}(\mathbf{A})\right]$$

- $\Omega(.)$ is a "sparsity-inducing" norm
- $\Omega^{'}(\mathbf{A})=\frac{1}{n}\sum_{i=1}^{n}\Omega(\mathbf{a}^i)$
- $\mathbf{X}=[\mathbf{x}^1,\ldots,\mathbf{x}^n]\in\mathbb{R}^{m\times n}$: samples
- $\mathbf{A}=[\mathbf{a}^1,\ldots,\mathbf{a}^n]\in\mathbb{R}^{p\times n}$: sparse codes for each sample
- $\|\mathbf{X}\|_F=\left(\sum_{i=1}^{m}\sum_{i=1}^{n}x_{ij}^2\right)^{\frac{1}{2}}$: Frobenius norm

Classical matrix factorization

PCA

$$\min_{\substack{\mathbf{D}\in\mathbb{R}^{m\times p}\\ \mathbf{A}\in\mathbb{R}^{p\times n}}} \frac{1}{2}\|\mathbf{X}-\mathbf{DA}\|_F^2 \ \text{ s.t. } \ \mathbf{D}^\mathsf{T}\mathbf{D}=\mathbf{I}_m \ \text{ and } \ \mathbf{A}\mathbf{A}^\mathsf{T} \ \text{ is diagonal}$$

***k*-means**

$$\min_{\substack{\mathbf{D}\in\mathbb{R}^{m\times k}\\ \mathbf{A}\in\{0,1\}^{k\times n}}} \frac{1}{2}\|\mathbf{X}-\mathbf{DA}\|_F^2 \ \text{ s.t. } \ \sum_{j=1}^{k}\mathbf{a}_j^i = 1, \ \text{ for all } \ i\in\{1,\ldots,p\}$$

Algorithms

MF with ℓ_1 regularization

$$\min_{\substack{\mathbf{D}\in\mathcal{D}\\ \mathbf{A}\in\mathbb{R}^{m\times p}}} \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{2}\|\mathbf{x}^i-\mathbf{D}\mathbf{a}^i\|_2^2+\lambda\|\mathbf{a}^i\|_1\right]$$

- Optimization is not jointly convex in $(\mathbf{D},\mathbf{A})$
- BUT, is convex w.r.t. each when the other is fixed
- use LARS and gradient descent interchangeably, i.e., separate **sparse coding** and **dictionary learning** steps

De-noising

$$\min_{\mathbf{x}\in\mathbf{R}^n} \frac{1}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 + \psi(\mathbf{x})$$

- data fitting term + regularization term such that estimate respect image model

$$\min_{\substack{\mathbf{D}\in\mathcal{D}\\ \mathbf{A}\in\mathbb{R}^{m\times p}}} \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{2}\|\mathbf{x}^i - \mathbf{D}\mathbf{a}^i\|_2^2 + \lambda\|\mathbf{a}^i\|_1\right]$$

$$\mathbf{x} = \frac{1}{m}\sum_{i=1}^{n}\mathbf{R}^i\mathbf{D}\mathbf{a}^i$$

Inpainting

$$\min_{\substack{\mathbf{D}\in\mathcal{D}\\ \mathbf{A}\in\mathbb{R}^{m\times p}}} \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{2}\|(\mathbf{x}^i - \mathbf{D}\mathbf{a}^i)\|_2^2 + \lambda\|\mathbf{a}^i\|_1\right]$$

- can only handle holes that are smaller than the patch size

Figure: Damaged image (Mairal, 2010).

Figure: Restored image (Mairal, 2010).

References I

Alfred M. Bruckstein, David L. Donoho, and Michael Elad. From sparse solutions of systems of equations to sparse modeling of signals and images. *SIAM Review*, 51(1):34–81, 2009.

Scott S. Chen, David L. Donoho, and Michael A. Saunders. Atomic decomposition by basis pursuit. Technical Report 479, Department of Statistics, Stanford University, May 1995.

Adam Coates and Andrew Y. Ng. The importance of encoding versus training with sparse coding and vector quantization. In *Twenty-Eighth International Conference on Machine Learning*, 2011.

Adam Coates, Honglak Lee, and Andrew Y. Ng. Self-taught learning: Transfer learning from unlabeled data. In *Int. Conf. on AI and Stats.*, 2011.

Julien Mairal. *Sparse Coding for Machine Learning, Image Processing and Computer Vision*. PhD thesis, Ecole Normale Superieure de Cachan, 2010.

References II

Stéphane Mallat and Zhifeng Zhang. Matching pursuit with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41:3397–3415, 1993.

Hosein Mohimani, Massoud Babaie-Zadeh, and Christian Jutten. A fast approach for overcomplete sparse decomposition based on smoothed ℓ_0 norm. *Transactions on Signal Processing*, 57:289–301, January 2009. ISSN 1053-587X.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society (Series B)*, 58:267–288, 1996.

John Wright, Allen Y. Yang, Arvind Ganesh, S. Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:210–227, February 2009. ISSN 0162-8828.