Statistical Machine Learning for Data Science

Zaid Harchaoui

DATA 558

Week 4

Lecture 4: Outline

- Overview of supervised learning
- Principal component analysis

Supervised learning

General objective

Let $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \{1, \ldots, k\}$ be labelled training examples

$$\min_{B \in \mathbb{R}^{d \times k}} \quad \lambda \, \Omega(B) + \frac{1}{n} \sum_{i=1}^{n} L\big(y_i, B^T x_i\big)$$

Large-scale setting

$$n \gg 1, \quad d \gg 1, \quad k \gg 1$$

Gradient descent with adaptive step-size

- **Initialize**: $B_0 = 0$.
- **Iterate**:

Find η_t with backtracking rule.

$$\begin{aligned} B_{t+1} &= B_t - \eta_t \nabla_B F(B) \\ &= B_t - \eta_t \nabla_B \left\{ \frac{1}{n} \sum_{i=1}^{n} L(B; x_i, y_i) \right\} \end{aligned}$$

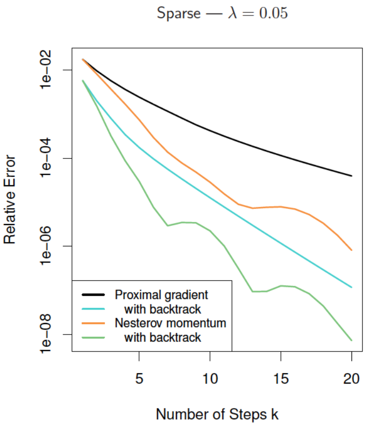
Fast Gradient Method

- **Initialize**: $B = 0$ and $\theta_0 = 0$.
- **Iterate**:

Find η_t with backtracking rule.

$$B_{t+1} = \theta_t - \eta_t \nabla_\theta F(\theta)$$
$$\theta_{t+1} = B_{t+1} + \frac{t}{t+3}(B_{t+1} - B_t)$$

Accelerated Gradient Method



Performance of the gradient descent versus accelerated gradient on a regression problem.

Large-scale supervised learning

General form

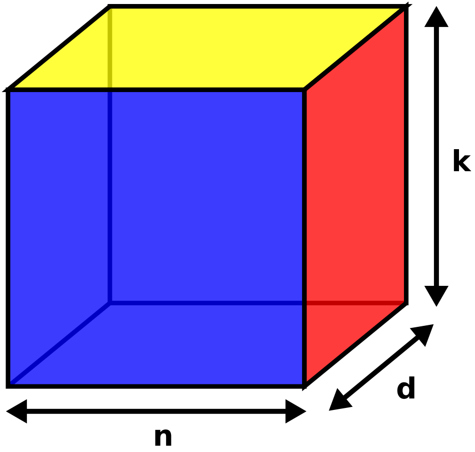
Let $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \{1, \ldots, k\}$ be labelled training examples

$$\min_{B \in \mathbb{R}^{d \times k}} \quad \lambda \, \Omega(B) + \frac{1}{n} \sum_{i=1}^{n} L\big(y_i, B^T x_i\big)$$

Problem: minimizing such objectives in the **large-scale** setting

$$n \gg 1, \quad d \gg 1, \quad k \gg 1$$

Machine learning cuboid



An example: ImageNet dataset

ImageNet dataset

- Large number of images/examples: $n = 17,000,000$
- Large number of pixels/image: $d = 200,030$
- Large number of categories: $k = 10,000$

Zoom on the ImageNet Dataset

Hierarchy of classes:

Deng, Dong, Socher, Li, Li and Fei-Fei, “Imagenet: a large-scale hierarchical image database”, CVPR’09.

Fine-grained subsets: generally more practical problems

→ Fungus: 134 classes, 90K images

Zoom on the ImageNet Dataset

Hierarchy of classes:

Deng, Dong, Socher, Li, Li and Fei-Fei, “Imagenet: a large-scale hierarchical image database”, CVPR’09.

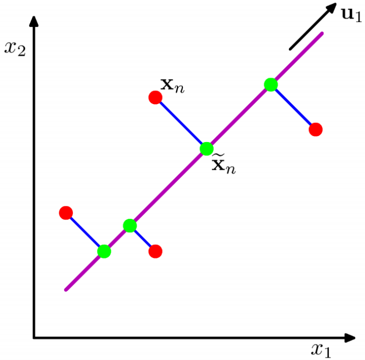
Fine-grained subsets: generally more practical problems

→ Vehicle: 262 classes, 226K images

Dimension Reduction: Principal Component Analysis

Goal

- Project data onto a space with dimensional $M < D$
- Maximize the variance of the projected data



Principal Component Analysis

Goal:

- Maximum variance criterion corresponds to a Rayleigh quotient
- PCA boils down to an eigenvalue problem on the *centered* covariance matrix $\hat{\Sigma}$ of the dataset, that is the principal components $w_1, \ldots, w_d$ are the eigenvectors of $\hat{\Sigma}$ (assuming $n > d$)
- Computational complexity: $O(ndc)$ in time with a *Singular Value Decomposition* (SVD; see `eigs` in Matlab/Octave), with n the number of points, d the dimension, c the number of principal components retained; stochastic approximation version for nonstationary/large-scale datasets.

Principal Component Analysis

$$\text{Empirical mean} \quad \bar{x} = \frac{1}{N} \sum_{j=1}^{d} x_j$$

$$\text{Empirical covariance} \quad \hat{\Sigma} = \frac{1}{N} \sum_{j=1}^{d} (x_j - \bar{x})(x_j - \bar{x})^T$$

Projection along the direction w

- $\mathrm{Proj}_w(x_j) = w^T x_j$, for all $j = 1, \ldots, N$
- $\mathrm{Proj}_w(\bar{x}) = w^T \bar{x}$

Principal Component Analysis

Projection along the direction w

- $\mathrm{Proj}_w(x_j) = w^T x_j$, for all $j = 1, \dots, N$
- $\mathrm{Proj}_w(\bar{x}) = w^T \bar{x}$

Variance of $\mathrm{Proj}_w(x_j)$

$$\frac{1}{N}\sum_{j=1}^{N}(w^T x_j - w^T \bar{x})^2 = w^T \hat{\Sigma} w \ .$$

First Principal Component

Projection along the direction w

- $\mathrm{Proj}_w(x_j) = w^T x_j$, for all $j = 1, \ldots, N$
- $\mathrm{Proj}_w(\bar{x}) = w^T \bar{x}$

Variance of $\mathrm{Proj}_w(x_j)$

$$\frac{1}{N}\sum_{j=1}^{N}(w^T x_j - w^T \bar{x})^2 = w^T \hat{\Sigma} w \ .$$

How to compute the top pair of eigenvalue and eigenvector

For a matrix A, the Power Iteration algorithm returns the top pair of eigenvalue λ and eigenvector v of the matrix A.

Algorithm 1 Power Iteration Algorithm

initialization v_0 random vector, and large number N .
repeat for $k = 1, 2, 3, \cdots, N$

- Perform update $z_k = Av_{k-1}$,
- Perform update $v_k = \frac{z_k}{\|z_k\|_2}, \lambda_k = v_k^T A v_k$.

until the stopping criterion is satisfied.

Variance along a direction and Rayleigh quotients

PCA seeks for directions $w_1, \ldots, w_c$ such that

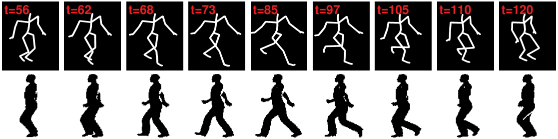
$$
\begin{aligned}
w_j &= \operatorname{argmax}_{w \in \mathbb{R}^d; w_j \perp \{w_1, \ldots, w_{j-1}\}} \operatorname{Var} \frac{(w, x)}{(w, w)} \\
&= \operatorname{argmax}_{w \in \mathbb{R}^d; w_j \perp \{w_1, \ldots, w_{j-1}\}} \frac{1}{m} \sum_{i=1}^{m} \frac{(w, x_i)^2}{(w, w)} \\
&= \operatorname{argmax}_{w \in \mathbb{R}^d; w_j \perp \{w_1, \ldots, w_{j-1}\}} \underbrace{\frac{(w, \hat{\Sigma} w)}{(w, w)}}_{\text{Rayleigh quotient}} .
\end{aligned}
$$

Principal components $w_1, \ldots, w_c$ are the first c eigenvectors of $\hat{\Sigma}$.

Low-dimensional representation with PCA

- Walking sequence of length 400 (containing about 3 walking cycles) obtained from the CMU Mocap database
- Data: silhouette images taken at a side view

Human body pose representation (Kim & Pavlovic, 2008). Selected skeleton and silhouette images for a half walking cycle.



Low-dimensional representation with PCA

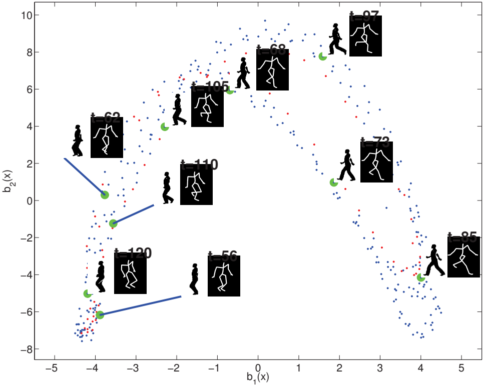


Figure: Central subspaces for silhouette images from walking motion

Low-dimensional representation with PCA

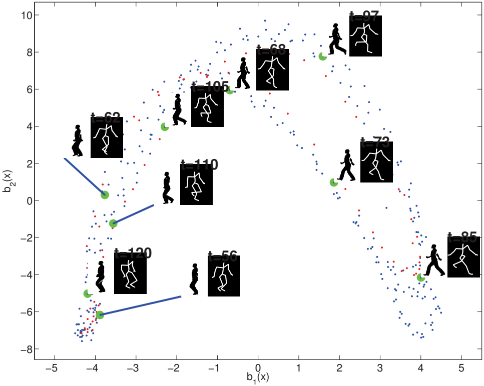


Figure: Central subspaces for silhouette images from walking motion

Super-resolution with PCA (Kim et al., 2005)

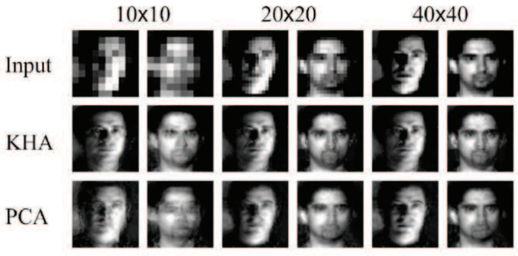


Figure: Super-resolution from low-resolution images of faces

Applications

- Image denoising (digits, faces, etc.)
- Visualization of bioinformatics data (strings, proteins, etc.)
- Dimension-reduction of high-dimensional features (appearance, interest points, etc.)