CS 584 – Machine Learning

Topic: Dimensionality Reduction

Mustafa Bilgic

http://www.cs.iit.edu/~mbilgic

https://twitter.com/bilgicm

Motivation

- Remove useless features
- Learn a low dimensional representation
- Visualization

Approaches

- Feature selection
- Feature extraction

Feature Selection

- Select a subset of the features
- Several approaches
  - Univariate feature selection
    - Mutual information, Chi2, …
  - Recursive feature elimination
    - Using an external estimator, recursively remove least useful features
  - Model-based feature selection
    - L1 regularization, decision trees, …
  - Sequential feature selection
    - Can use any model and performance metric

Feature Extraction

- Feature extraction, projection, latent representation learning, manifold learning, …
- Given an input, project it into another dimension (often lower)
- Unsupervised
  - Principal component analysis, isomap, t-SNE, deep learning, …
- Supervised
  - Linear discriminant analysis, deep learning, …

We'll Cover

- Principal component analysis
- Autoencoder

Principal Component Analysis

- Given a dataset $\boldsymbol{x}$ with d dimensions, project it into k dimensions ($k < d$) with minimum loss of information
  - $\boldsymbol{z} = \boldsymbol{w}^T \boldsymbol{x}$ — d, n; z: k, n; $w^T: k \times d$
- Principal component analysis (PCA) maximizes the variance in the projected space so that objects are spread out
  - $argmax_{\boldsymbol{w}} Var(\boldsymbol{z})$

PCA Objective

- $\boldsymbol{z} = \boldsymbol{w}^T \boldsymbol{x}$
- $argmax_{\boldsymbol{w}} Var(\boldsymbol{z})$
- Reminder: $Var(aX + b) = a^2 Var(X)$
- $Var(\boldsymbol{z}) = \boldsymbol{w}^T Var(\boldsymbol{x}) \boldsymbol{w}$
- We can trivially maximize variance by multiplying **w** by a large constant; hence, we enforce that $\boldsymbol{w}$ is unit length
  - $\boldsymbol{w}\boldsymbol{w}^T = 1$
- Objective
  - maximize $\boldsymbol{w}^T Var(\boldsymbol{x}) \boldsymbol{w}$ subject to $\boldsymbol{w}\boldsymbol{w}^T = 1$

$Var(w^T X) = w^T \cdot Var(x)$

$\underline{w} \cdot {}^T Var(x) \cdot \underline{w}$

Background – Constrained optimization

$$\begin{array}{lll} \text{Find} & \boldsymbol{\theta} & \\ \text{maximizing} & f(\boldsymbol{\theta}) & \\ \text{subject to} & & \\ & c_1(\boldsymbol{\theta}) & = & 0 \\ & & \ldots & \\ & c_m(\boldsymbol{\theta}) & = & 0 \end{array}$$

w

$\underline{w^T Va(x)} w$

$ww^T - 1 = 0$

Form the Lagrangian:

$$F(\boldsymbol{\theta}, \boldsymbol{\lambda}) = f(\boldsymbol{\theta}) - \sum_{j=1}^{m} \lambda_j c_j(\boldsymbol{\theta})$$

EXAMPLE

- Maximize xy subject to $x + y = 10$

$x + y - 10 = 0$

$$x \cdot y - \lambda(x + y - 10)$$

$\frac{\partial}{\partial x}$ $\quad y - \lambda = 0 \Rightarrow y = \lambda = 5$

$\frac{\partial}{\partial y}$ $\quad x - \lambda = 0 \Rightarrow x = \lambda = 5$

$\frac{\partial}{\partial \lambda}$ $\quad x + y - 10 = 0$

$\lambda + \lambda - 10 = 0 \Rightarrow \lambda = 5$

PCA WITH ONE DIMENSION

- maximize $Var(z_1) = w_1^T Var(\boldsymbol{x}) w_1$ subject to $w_1^T w_1 = 1$
- Let the variance of $\boldsymbol{x}$ be the covariance matrix Σ

  $d \times d$ $\qquad$ $d \times M$

- Objective
  - Maximize $w_1^T \Sigma w_1$ subject to $w_1^T w_1 = 1$

  $w_1^T w_1 - 1 = 0$

- Lagrange
  - maximize $w_1^T \Sigma w_1 - \lambda(w_1^T w_1 - 1)$
- Take derivative with respect to w_1, set to zero
  - $\frac{\partial L}{\partial w_1}$ $\quad 2\Sigma w_1 - 2\lambda w_1 = 0$
  - $\Sigma w_1 = \lambda w_1$
- w_1 is an eigenvector and λ is an eigenvalue of Σ
- Because we maximize variance
  - maximize $w_1^T \Sigma w_1 = w_1^T \lambda w_1 = \lambda w_1^T w_1 = \lambda$
  - w_1 is the eigenvector that corresponds to the largest eigenvalue of Σ

PCA WITH k DIMENSIONS

- We want the second vector w_2 to be orthogonal to w_1 so that the z_2 is uncorrelated with z_1

$w_2 \cdot w_1^T = 0$

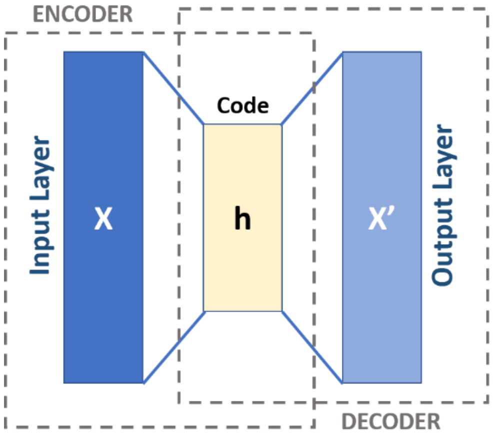
- Skipping the derivation details
  - w_1 is the eigenvector for the largest eigenvalue λ_1
  - w_2 is the eigenvector for the second largest eigenvalue λ_2
  - …
  - w_k is the eigenvector for the k^{th} largest eigenvalue λ_k
- In the end, the transformation is typically centered around zero (in the new dimensions)
  - $\boldsymbol{z} = \mathbf{w}^{\mathrm{T}}(\mathbf{x} - \mathbf{m})$ where $\boldsymbol{m}$ is the mean of $\boldsymbol{x}$

PCA Example

- See OneNote and Jupyter notebook

AutoEncoder

- A neural network architecture where the input and output are the same
  - $\boldsymbol{x} \rightarrow \boldsymbol{h} \rightarrow \boldsymbol{x}$
  - Input $\boldsymbol{x}$ is encoded into $\boldsymbol{h}$, where $\boldsymbol{h}$ is typically lower dimensional than $\boldsymbol{x}$ and $\boldsymbol{h}$ is decoded back to $\boldsymbol{x}$, with some error



https://upload.wikimedia.org/wikipedia/commons/3/37/Autoencoder_schema.png

Other DL Methods

- Transformers
- Contrastive learning
- …
- International conference on learning representations (ICLR)
  - https://iclr.cc/