Machine Learning Algorithms

14 October 2016

1 Introduction

The first efforts (the formats, texts, references etc.) for the task, including reviews on PCA, LDA and Cloud Computing.

The following materials will contribute to different chapters.

2 PCA

The sheer size of the network traffic data in the modern age is not only a challenge for computer hardware but also a main bottleneck for the performance of many machine learning algorithms. Principal Component Analysis (PCA) is a simple yet popular and useful linear transformation technique that is used in numerous applications, such as stock market predictions, the analysis of gene expression data, and many more. In this thesis we will use it to analyse the huge amount of network traffic data for network security purpose.

The main goal of a PCA analysis is to identify patterns in data; PCA aims to detect the correlation between variables. If a strong correlation between variables exists, the attempt to reduce the dimensionality only makes sense. In a nutshell, this is what PCA is all about: Finding the directions of maximum variance in high-dimensional data and project it onto a smaller dimensional subspace while retaining most of the information.

The Principal Component Analysis (PCA), which is the core of the Eigen-based method, finds a linear combination of features that maximizes the total variance in data. While this is clearly a powerful way to represent data, it doesnt consider any classes and so a lot of discriminative information may be lost when throwing components away.

- Sebastian Raschka
- Face Recognition with OpenCV
- Principal Component Analysis (PCA)
- Careful, PCA does SVD on the covariance matrix, LSA does it on the data matrix directly.

2.1 PCA and Dimensionality Reduction

Often, the desired goal is to reduce the dimensions of a **d**-dimensional dataset by projecting it onto a (**k**)-dimensional subspace (where $\mathbf{k} < \mathbf{d}$) in order to increase the computational efficiency while retaining most of the information.

In Chapter 4, we will compute eigenvectors (the principal components) of the network traffic dataset and collect them in a projection matrix. Each of those eigenvectors is associated with an eigenvalue which can be interpreted as the length or magnitude of the corresponding eigenvector. If some eigenvalues have a significantly larger magnitude than others that the reduction of the dataset via PCA onto a smaller dimensional subspace by dropping the less informative eigenpairs is reasonable.

2.2 A Summary of the PCA Approach

- Standardize the data.
- Obtain the Eigenvectors and Eigenvalues from the covariance matrix or correlation matrix, or perform Singular Vector Decomposition.
- Sort eigenvalues in descending order and choose the **k** eigenvectors that correspond to the kk largest eigenvalues where **k** is the number of dimensions of the new feature subspace ($\mathbf{k} \leqslant \mathbf{d}$).
- Construct the projection matrix **W** from the selected kk eigenvectors.
- Transform the original dataset **X** via **W** to obtain a **k**-dimensional feature subspace **Y**.

3 LDA

Linear Discriminant Analysis (LDA) is most commonly used as dimensionality reduction technique in the pre-processing step for pattern-classification and machine learning applications. The goal is to project a dataset onto a lower-dimensional space with good class-separability in order avoid overfitting (" curse of dimensionality") and also reduce computational costs.

Ronald A. Fisher first formulated the Linear Discriminant in 1936 in his classic work [1], and it also has some practical uses as classifier. The original Linear discriminant was described for a 2-class problem, and it was then later generalized as "multi-class Linear Discriminant Analysis" or "Multiple Discriminant Analysis" by C. R. Rao in 1948 [2].

In a nutshell, often the goal of an LDA is to project a feature space (a dataset n-dimensional samples) onto a smaller subspace **k** (where $k \leqslant n-1$) while maintaining the class-discriminatory information. In general, dimensionality reduction does not only help reducing computational costs for a given classification task, but it can also be helpful to avoid overfitting by minimizing the error in parameter estimation ("curse of dimensionality").

3.1 A Summary of the LDA Method

- Compute the **d**-dimensional mean vectors for the different classes from the dataset.

- Compute the scatter matrices (in-between-class and within-class scatter matrix).
- Compute the eigenvectors ($\mathbf{e}_1$, $\mathbf{e}_2$, ..., $\mathbf{e}_d$) and corresponding eigenvalues (λ_1, λ_2, ..., λ_d) for the scatter matrices.
- Sort the eigenvectors by decreasing eigenvalues and choose k eigenvectors with the largest eigenvalues to form a $d \times k$ dimensional matrix $\mathbf{W}$ (where every column represents an eigenvector).
- Use this $d \times k$ eigenvector matrix to transform the samples onto the new subspace. This can be summarized by the matrix multiplication: $\mathbf{Y}=\mathbf{X} \times \mathbf{W}$ (where $\mathbf{X}$ is a $n \times d$-dimensional matrix representing the n samples, and $\mathbf{y}$ are the transformed $n \times d$-dimensional samples in the new subspace).

4 PCA vs. LDA

Both Linear Discriminant Analysis (LDA) and PCA are linear transformation methods. PCA yields the directions (principal components) that maximize the variance of the data, whereas LDA also aims to find the directions that maximize the separation (or discrimination) between different classes, which can be useful in pattern classification problem (PCA ignores class labels). In other words, PCA projects the entire dataset onto a different feature (sub)space, and LDA tries to determine a suitable feature (sub)space in order to distinguish between patterns that belong to different classes.

The prime difference between LDA and PCA is that PCA does more of feature classification and LDA does data classification. In PCA, the shape and location of the original data sets changes when transformed to a different space whereas LDA doesn?t change the location but only tries to provide more class separability and draw a decision region between the given classes.This method also helps to better understand the distribution of the feature data.

5 Cloud Computing

5.1 A Survey of Security and Privacy Challenges in Cloud Computing

Cloud computing is defined as a service model that enables convenient, on-demand network access to a large shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction [3].

This innovative information system architecture, which is fundamentally changing the way that computing, storage and networking resources are allocated and managed, brings numerous advantages to users, including but not limited to reduced capital costs, easy access to information, improved flexibility, automatic service integration, and quick deployment [4]

5.2 Challenges

The paper identifies several specific security challenges in cloud computing which require the development of advanced security technology.

Loss of Control refers to the situation that cloud users control over their data is diminished when they move the data from their own local servers to remote cloud servers. A great number of concerns about data protection are raised

Lack of Transparency indicates the conflict interests between the Cloud Service Provider (CSP) and Cloud Service Users (CSUs).

Virtualization Related Issues includes New Access Context, Attacks against Hypervisor etc.

Multi-Tenancy Related Issues Multi-tenancy is defined as the practice of placing multiple tenants on the same physical hardware to reduce costs to the user by leveraging economies of scale[5]. It indicates sharing of computational resources, storage, services and applications with other tenants, hosted by the same physical or logical platform at the providers premises.

Managerial Issues Most cloud-specific security and privacy challenges have their own managerial aspect, including Loss of control, the lack of transparency challenge as well as the malicious insider challenge. The fact that managerial challenges are overarching and add to the other challenges is what makes it one of the toughest challenges to deal with.

5.3 Existing Solutions

Diverse defense studies have been launched to secure the cloud computing environment. The state-of-the-art researches that aims to address the security issues in cloud computing are summaries in the following section.

Encryption Algorithms At the current stage, encryption is still the major solution for addressing data confidentiality issues in cloud computing

Access Control Access control, consisting of authentication, authorization, and accounting, is a way of ensuring that the access is provided only to the authorized users, hence the data is stored in a secure manner

Third Party Auditing CSUs and CSPs are not involved in the auditing process except for providing data and information for the independent auditors. TPA can be used to relieve the concerns on data integrity, confidentiality, availability, and privacy. TPA can examine at least two aspects of data integrity: while data is in transit and while it is stationary.

Isolation Current studies handle isolation from several aspects. 1) Hypervisors or virtual machine monitor (VMM), a piece of computer software, firmware or hardware that creates and runs virtual machines, can be utilized to facilitate isolation. 2) Some software-level resource management mechanisms are proposed to perform isolation for cache, disk, memory

bandwidth, and network. 3) Hardware-level solutions are proposed to allocate memory bandwidth and processor caches in a better way. 4) Strict mechanisms to separate customer data are required by cloud users . 5) Security models are established to ensure isolation.

Soft Trust Solutions Trust has been identified as one promising approach to address security and privacy issues in cloud computing. Specifically, soft trust is defined as the relationship between two parties for a specific action or property. Diverse trust models have been proposed to evaluate the trustworthiness of a CSP.

Hard Trust Solutions In the cloud computing model, customer views are limited to a virtual infrastructure typically built on top of non-trusted physical hardware or operating environments. Hardware-based security solutions are envisioned as a natural trend that a CSP will be likely to follow in coming years to resolve different data privacy and integrity issues

Governance Governance refers to a comprehensive set of activities associated with planning and implementing controls. In the context of cloud security, there are some initial signs of a cloud-specific security governance framework emerging.

References

[1] Fisher, Ronald A. "The use of multiple measurements in taxonomic problems." Annals of eugenics 7.2 (1936): 179-188.

[2] Rao C R. The utilization of multiple measurements in problems of biological classification[J]. Journal of the Royal Statistical Society. Series B (Methodological), 1948, 10(2): 159-203.

[3] P. Mell and T. Grance, The NIST definition of cloud computing, 2011; http://csrc.nist.gov/publications/nistpubs/800- 145/SP800-145.pdf.

[4] P. Viswanathan, Cloud computing Is it really all that beneficial? http://mobiledevices.about.com/od/additionalresources/

[5] W. J. Brown, V. Anderson, and Q. Tan, Multitenancy-security risks and countermeasures, in Proceedings of 2012 15th International Conference on Network-Based Information Systems (NBiS), Melbourne, Australia, 2012, pp. 7-13.

[6] A. Behl and K. Behl, An analysis of cloud computing security issues, in Proceedings of 2012 World Congress on Information and Communication Technologies (WICT), Trivandrum, India, 2012, pp. 109-114.