Stat 991: Multivariate Analysis, Dimensionality Reduction, and Spectral Methods **Lecture: 2**

The SVD and applications

Instructor: Sham Kakade

1 Applications

1.1 Latent Semantic Analysis (LSA) or LSI (LSIndexing)

Let look at an application to information retrieval.

Say we represent a document by a vector d and a query by a vector q, then one score of a match is the cosine score:

$$\text{similarity} = \frac{d \cdot q}{\|d\| \|q\|}$$

The naive approach is to just use a bag of words to represent these vectors — so the length of the vector is the number of words (in the language or corpus) and the entry in the k-th position denote the number of times that word appears. Using just bag of word counts, two difficulties with this approach are synonymy and polysemy.

LSA is a simple way to address this, using a vector space method. Here, let X be the term/document matrix. Let:

$$X = UDV^\top$$

be the SVD of X. We can work with the k-rank approximation to X:

$$X_k = U_k D_k V_k^\top$$

So we represent each document and (new) query as a k-vector. The document j is just represented by V_j. A vector query q is now represented as:

$$x_{term}(q) = D_k^{-1} U_k^\top q x_{document}(d) = D_k^{-1} V_k^\top d$$

Now for recall we can just use the cosine score for retrieval.

1.2 EigenFaces

In vision, a common (and simple) way to represent centered faces is by projecting onto their top singular vectors.

1.3 PageRank

See Wikipedia. Essentially, the PageRank algorithm computes the stationary distribution of the web graph, along with some 'transition' noise (which ensure the induced random walk is ergodic). Also, for IR applications, there is a notion of 'anchor text', where it is also important to consider the text that co-occurs with a link to another page (e.g. the website for IBM does not even have the word computer on the front page, but the anchor text will often have more information about IBM then IBM's homepage itself).

2 PCA

Given a finite sample $X_1, \dots X_n$, we have the empirical covariance matrix:

$$\hat{K} = \frac{1}{n} \sum_{i=1}^{n} X_i X_i^\top$$

PCA is just the KL transform (discussed later) of the empirical Kernel matrix.

Alternative viewpoint:

$$w_1 = \arg \max_{w:\|w\|=1} \hat{\sigma}^2(w \cdot X) = \arg \max_{w:\|w\|=1} \frac{1}{n} \sum_{i=1}^{n} (w \cdot X_i)^2$$

and z_1 is the value. Next,

$$X_i \leftarrow X_i - \sum_j (w_1 \cdot X_i) w_1$$

and repeat to find e_2 and z_2 and so on.

3 Related Ideas for Functions

One can view functions as essentially a "generalized" vector — namely, once can view functions as living in linear spaces, with associated norms and inner products. For example for functions $f(x)$ and $g(x)$, we can define the norm $\|f\| = \int f(x) d\mu(x)$ and the inner product $f \cdot g = \int f(x) g(x) d\mu(x)$. Similarly, many of the decomposition methods apply.

4 Mercer's Theorem

Theorem 4.1. *Suppose K is a continuous symmetric non-negative definite kernel. Then there is an orthonormal basis $\{e_i\}$ on $L_2[0,1]$ consisting of eigenfunctions of T_K such that the corresponding sequence of eigenvalues $\{\lambda_i\}$ is nonnegative. The eigenfunctions corresponding to non-zero eigenvalues are continuous on $[0,1]$ and K has the representation:*

$$K(s,t) = \sum_{i=1}^{\infty} \lambda_j e_j(s) e_j(t)$$

where the convergence is absolute and uniform.

In finite dimensions,

Theorem 4.2. *Suppose K is a square symmetric matrix. Then there exists a decomposition:*

$$S = UDU^\top$$

where D is diagonal and U is orthogonal. The diagonal entries of D are the eigenvalues and corresponding columns of U are the eigenvalues. If K is non-negative definite then all the eigenvalues are positive.

5 Karhunen-Loeve theorem

Consider a centered stochastic process $[X]_t$, for $t \in [0,1]$. Centered means that $\mathbb{E}[X]_t = 0$. In the discrete case we have a random vector $X \in \mathbb{R}^d$ where $[X]_t$ is the $t - thcomponent$.

The autocovariance function is:

$$K(t,s) = Cov(X_t, X_s) =< X_t | X_s >= \mathbb{E}[X_t X_s]$$

which can be viewed as a kernel.

The corresponding integral operator is:

$$T_K \Phi(t) = \int_0^1 K(t,s)\Phi(t)ds$$

which has eigenvectors and eigenvalues.

Theorem 5.1. *(KL) Consider the centered stochastic process X_t for $t \in [0,1]$ with covariance function $K(t,s)$. Suppose this covariance function is continuous in t, s. By Mercer's theorem, the corresponding integral operator on T_K has an orthonormal basis of eigenvectors, $\{e_i(t)\}$. Define:*

$$Z_i = \int_0^1 X_t e_i(t)dt$$

Then Z_i are centered orthogonal random variables and:

$$X_t = \sum_i^\infty e_i(t) Z_i$$

(where convergence is in the mean and uniform in t). Also,

$$Var(Z_i) = \mathbb{E}(Z_i^2) = \lambda_i$$

where λ_i is the eigenvalue corresponding to e_i.

6 The Pseudo Inverse

For $X \in \mathbb{R}^{n\times d}$ and $Y \in \mathbb{R}^n$, suppose that the equation:

$$X\beta = Y$$

has a unique solution and that X is invertible, then:

$$\beta = X^{-1}Y$$

In regression, there is typically noise, and we find a β which minimizes:

$$\|X\beta - Y\|$$

Clearly, if there is no noise, then a solution is given by $\beta = X^{-1}Y$, assuming no degeneracies. In general though, the least squares solution is given by:

$$\beta = (X^\top X)^{-1} XY \tag{1}$$

which one can argue is a less intuitive (and elegant) expression than when there is an exact solution. Furthermore, Equation 1 above only holds if X is of rank d (else $(X^\top X)^{-1}$ would not be invertible).

Now let us define the Moore-Penrose pseudo-inverse. While there are a variety of more elegant definitions of the pseudo-inverse, in terms of certain desirable properties, we take the more brute force definition.

First, let us define the 'thin' SVD.

Definition 6.1. *We say $X = U\Sigma V^\top$ is the "thin" SVD of $X \in \mathbb{R}^{n\times d}$ if: $U^{n\times r}$ and $V^{d\times r}$ have orthonormal columns (e.g. where r is the number of columns) and $\sigma \in \mathbb{R}^{r\times r}$ is diagonal, with all it's diagonal entries being non-zero.*

Now we define the pseudo-inverse as follows:

Definition 6.2. *Let $X = U\Sigma V^\top$ be the thin SVD of X. The Moore-Penrose pseudo-inverse of X, denoted by X^+, is defined as:*

$$X^+ = V\Sigma^{-1}U^\top$$

Let us make some observations:

1. First, if X is invertible (so X is square) then $X^+ = X^{-1}$.
2. Suppose that X isn't square and that $Xw = Y$ has a (unique) solution, then $w = X^+Y$.
3. Now suppose that $Xw = Y$ has (at least one) solution. Then one solution is given by $w = X^+Y$. This solution is the minimum norm solution w.
4. (geometric interpretation) The matrix X^+ maps any point in the range of X to the minimum norm point in the domain.

With the pseudo-inverse, we have the much more elegant least squares estimator:

Lemma 6.3. *The least squares estimator is:*

$$\beta = X^+Y$$

(Note that the above is alway a minimizer, while the solution provided in Equation 1 only holds if $X^\top X$ is invertible, in which case the minimizer is unique).

7 References

Material used was Wikipedia and Santosh Vempala's lecture notes.