Algorithms for massive datasets

Francesco Tomaselli

September 4, 2021

Contents

1 Link analysis

1.1 Power method

Computing the pagerank index given a connection matrix consists of finding a probability distribution over websites in the web graph.

We can initialize a vector with n components, each with value $\frac{1}{n}$, where n is the number of websites and multiply it to the connection matrix many times. We will converge to an eigenvector of the matrix, with eigenvalue 1, that represent the final probability distribution over websites.

Result is an eigenvector We start with a matrix $A = [a_{i,j}]_{n \times n}$, given an eigenpair $(\lambda, \vec{x})$, it holds that $A\vec{x} = \lambda\vec{x}$.

Also, for A dimensionality, we have n eigenpairs, and we can sort them by non-increasing values of eigenvalue

$$(\lambda_1, \vec{x}_1), \ldots, (\lambda_n, \vec{x}_n)$$

Eigenvalues are orthogonal and they constitute a basis for the space in which vectors such as $\vec{x}$ lives. If we now consider a starting vector v_0 as the starting probability distribution, we can rewrite it as follows.

$$\vec{v}_0 = \alpha_1\vec{x}_1 + \cdots + \alpha_n\vec{x}_n$$

Thus the matrix vector multiplication to find the final distribution becomes:

$$\begin{aligned}
\vec{v}_1 = A\vec{v}_0 &= A(\alpha_1\vec{x}_1 + \cdots + \alpha_n\vec{x}_n) && \text{For the above equation} \\
&= \alpha_1 A\vec{x}_1 + \cdots + \alpha_n A\vec{x}_n && \text{Distribute multiplication} \\
&= \alpha_1\lambda_1\vec{x}_1 + \cdots + \alpha_n\lambda_n\vec{x}_n && \text{They are eigenvectors}
\end{aligned}$$

Computing $\vec{v}_2$ now becomes:

$$\begin{aligned}
\vec{v}_2 = A\vec{v}_1 &= A(\alpha_1\lambda_1\vec{x}_1 + \cdots + \alpha_n\lambda_n\vec{x}_n) && \text{Using the previous result} \\
&= \alpha_1\lambda_1^2\vec{x}_1 + \cdots + \alpha_n\lambda_n^2\vec{x}_n && \text{We distribute A, as before}
\end{aligned}$$

At the generic $\vec{v}_k$ we have:

$$\begin{aligned}
\vec{v}_k &= \alpha_1\lambda_1^k\vec{x}_1 + \cdots + \alpha_n\lambda_n^k\vec{x}_n \\
&= \lambda_1^k\left(\alpha_1\vec{x}_1 + \alpha_2\left(\frac{\lambda_2}{\lambda_1}\right)^k\vec{x}_2 \ldots\right) \\
&\approx \lambda_1^k\alpha_1\vec{x}_1 && \text{As k grows, lambdas are sorted}
\end{aligned}$$

We have a practical problem, as if λ_1 is smaller than 1 this method diverges. It holds that the maximum eigenvalue of a row-stochastic matrix is equals to 1.

Determinant of the transpose To prove it, we start to prove that $detA^T = detA$:

$$detA = \sum_i a_{ij}c_{ij} = \sum_j a_{ij}c_{ij}$$
$$detA^T = \sum_i a^T_{ij}c^T_{ij} = \sum_i a_{ji}c_{ji} = \sum_j a_{ij}c_{ij} = detA$$

It also holds that A and A^T have the same eigenvalues:

$$det(A - \lambda I) = 0 \longleftrightarrow det(A - \lambda I)^T = 0$$
$$\longleftrightarrow det(A^T - \lambda I) = 0$$

One as eigenvalue We can now show that 1 is always an eigenvalue of a row-stochastic matrix:

$$A \cdot \vec{1} = \left[\sum_j a_{ij} \cdot 1\right]_n = \vec{1} \quad \text{As the components in } A \text{ row-wise are probabilities}$$

We can multiply the unit vector by one showing that it is a valid eigenvalue. Similarly, this results holds even if A is column-stochastic.

Power of stochastic matrix We now prove that if A is row-stochastic, A^k is as well.

$$k = 1: \qquad \text{trivial}$$

$$A^k \; r.s. \rightarrow A^{k+1} \; r.s.: \qquad a^{k+1}_{ij} = \sum_s a^k_{is} \cdot a_{sj}$$
$$\sum_j a^{k+1}_{ij} = \sum_j \sum_s a^k_{is} \cdot a_{sj}$$
$$= \sum_s a^k_{is} \sum_j a_{sj}$$
$$= \sum_s a^k_{is} \cdot 1 = 1$$

One is highest eigenvalue We are left to show 1 is the highest eigenvalue of a column-wise stochastic matrix. Suppose it exists an eigenvalue greater than one.

$$A^T = \lambda \vec{v}$$
$$(A^T)^k = \lambda^k \vec{v}$$
$$\sum_j ((a^T)^k)_{ij} v_j = \lambda^k v_i$$

We can overestimate the left end with v_{max}, the maximum vector component, we underestimate tghe right end with G, a value of choice. It exists a k that

makes the equation true as lambda is greater than one. We then divide by v_{max}.

$$\sum_j ((a^T)^k)_{ij} v_{max} > G$$

$$1 = \sum_j ((a^T)^k)_{ij} > \frac{G}{v_{max}}$$

A is column wise stochastic, the transpose is row stochastic, thus if we raise to the k we get a sum of one. This means it is absurd to assume we have an eigenvalue greater than one, as G is arbitrary big.