Gradient and Linear Algebra

Gradient of a least-squares loss in a linear model

Consider the linear model $$y = X \cdot \theta$$

where theta is a parameter vector of length D, X is an n by D input feature matrix and y are the corresponding observations of length n.

Optimizing such a model can be considered as solving

$$\min_{\theta \in \mathbb{R}^{\mathbb{D}}} \left(||y - X \cdot \theta||^2 \right)$$

Gradient of a least-squares loss in a linear model

$$\min_{\theta \in \mathbb{R}^{\mathbb{D}}} \left(||y - X \cdot \theta||^2 \right)$$

This can be solved by computing the gradient of $L = ||e||^2, \; e = y - X \cdot \theta$

$$\frac{\partial L}{\partial e} = 2e^T \qquad\qquad \frac{\partial e}{\partial \theta} = -X$$

$$\frac{\partial L}{\partial \theta} = \frac{\partial L}{\partial e} \frac{\partial e}{\partial \theta} = -2\left(y^T - \theta^T X^T\right) X$$

* Solving derivative equals 0 is sufficient to minimize the loss, since the Hassian of L equals $X^T X$ is PSD.

Least squared loss again

Linear regression with regularization.

$$\min_{a,b} ||y - X \cdot \theta||^2 + \rho||\theta||^2$$

- Calculate the derivative of this objective w.r.t. Theta.
- Is there always a closed-form analytical solution? What is it?

Rank and null space

The rank of a matrix A is the dimension of the vector space generated (or spanned) by its columns

$$A = \begin{bmatrix} 2 & -5 & 3 \\ 0 & 7 & -2 \\ -1 & 4 & 1 \end{bmatrix}$$

The null space of an m × n matrix A is the set of all solutions to the homogeneous equation Ax = 0.

Eigenvalues and Eigenvectors

Theorem 12.8. If matrices A and B are **similar**, i.e., if there is an invertible matrix P such that $A = P^{-1}BP$, then they have the same eigenvalues.

Proof: For any eigenvalue and eigenvector pair (λ, x) , we know

$Ax = \lambda x = P^{-1}BPx$, thus $\lambda Px = BPx$. Therefore (λ, Px)

is a pair of eigenvalue and eigenvector of B.

Eigenvalues and Eigenvectors

Definition: A matrix A is diagonalizable if A is similar to a diagonal matrix.

Theorem 12.9. A is diagonalizable if and only if A has n linearly independent eigenvectors.

Sketch:

1. If A is diagonalizable, then there is an invertible matrix P and a diagonal matrix D, such that $D = P^{-1}AP$, thus PD=AP.
2. Consider P as $[p_1, p_2, \ldots p_n]$, and D has elements $\lambda_1, \ldots \lambda_n$ on the diagonal, then $Ap_i = \lambda_i p_i$.
3. P is invertible, so its columns are linearly independent.
4. Assume A has n linearly independent eigenvectors, and reverse the steps above.

Eigenvalues and Eigenvectors

Theorem 12.10. Let A be a real symmetric matrix, then all its eigenvalues and eigenvectors are real. Besides, A is orthogonally diagonalizable and

$$A = VDV^T = \sum_i \lambda_i v_i v_i^T$$

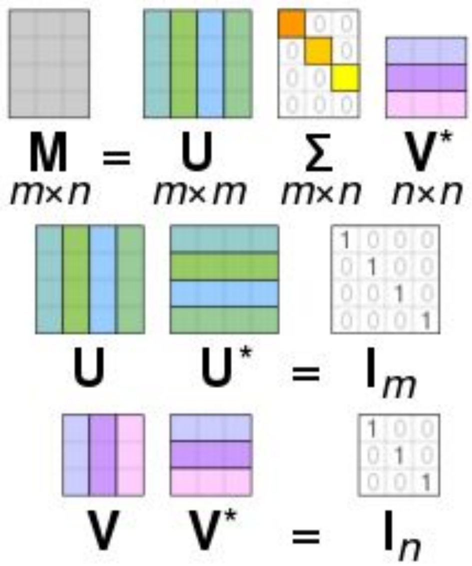
Where V is a matrix with eigenvectors as columns and D is a diagonal matrix with corresponding eigenvalues.

SVD

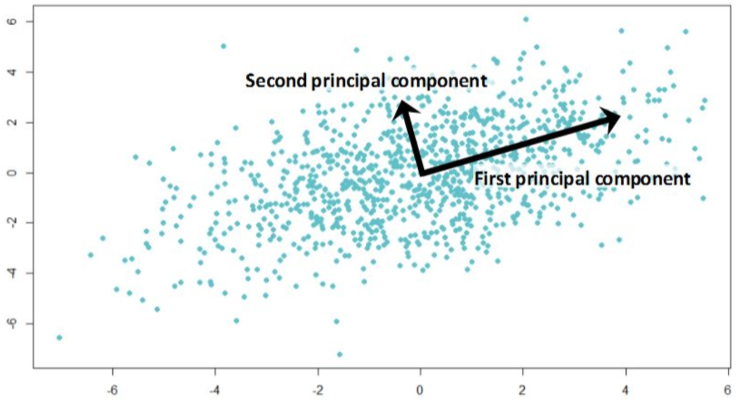
Singular value decomposition for matrix M:

$$M = U\Sigma V^T$$

- U and V are unitary matrices. This means rows/columns of U are orthonormal.
- $\sum$ is a rectangular diagonal matrix with singular values on the diagonal.



PCA



Consider a n by p matrix X with column-wise zero empirical mean.

- Each column can be considered as a feature
- Each row is a data sample

First round:

We want to find a unit vector $w_{(1)}$, such that the projections of data samples on this vector have the largest variance.

$$w_{(1)} = \arg\max_{||w||=1} \left\{ \sum_{i=1}^{n} (x_i \cdot w)^2 \right\} = \arg\max_{||w||=1} \left\{ ||Xw||^2 \right\} = \arg\max_{||w||=1} \left\{ w^T X^T X w \right\}$$

PCA

$$w_{(1)} = \arg\max_{||w||=1} \{w^T X^T X w\} = \arg\max_{||w||=1} \left\{ \frac{w^T X^T X w}{w^T w} \right\}$$

k-th round

We can form a new data matrix by subtracting all previous principal components from X.

$$X_k = X - \sum_{s=1}^{k-1} X w_{(s)} w_{(s)}^T$$

Then repeat the process of finding the unit vector that leads to the max variance of projections.

Details not required in the lab: This is called Rayleigh quotient. Since $X^T X$ is a positive semidefinite matrix, the maximum value of it is the largest eigenvalue of the matrix when w is the corresponding eigenvector.

Project matrix (each data sample) to s-th principal component.

PCA

A matrix W can be formed as $\left[w_{(1)}|\ldots|w_{(l)}\right]$, where $l \leq p$. And the final transformed data is $T = XW$.

Connection to SVD

By SVD, we know that $$X^T X = V\Sigma^T U^T U \Sigma V^T = V\Sigma^T \Sigma V^T$$

$$= V\hat{\Sigma}^2 V^T$$, where $\hat{\Sigma}^2$ is a diagonal matrix.

This is the format of eigen-decomposition, which implies the right singular vectors V of X are also the eigenvectors of $X^T X$, i.e., V=W, exactly the solution we need for PCA.

This is it