CS 189 Fall 2017 | Introduction to Machine Learning | DIS1

1 Unitary invariance

Prove that the regular Euclidean norm (also called the 2-norm) is unitary invariant; in other words, the 2-norm of a vector is the same, regardless of how you apply a rigid transformation to the vector (i.e., rotate or reflect). Note that rigid transformation of a vector $\vec{v} \in \mathbb{R}^d$ means multiplying by an orthogonal $U \in \mathbb{R}^{d\times d}$.

Solution:

Recall that an orthogonal matrix U is one whose columns are orthonormal — i.e. each has norm 1 and their Euclidean inner products with each other are zero. This then means that $U^T U = I$.

Take a rotated or reflected version of v to then be $v_2 = Uv$ for an orthogonal matrix U.

$$\|v_2\|_2^2 = \|Uv\|_2^2 = (Uv)^T(Uv) = v^T U^T U v = v^T v = \|v\|_2^2$$

Take the square root of both sides; this is valid since the L2-norm is always non-negative.

$$\|v_2\|_2 = \|v\|_2$$

2 Eigenvalues

(a) Let A be an invertible matrix. Show that if $\vec{v}$ is an eigenvector of A with eigenvalue λ, then it is also an eigenvector of A^{-1} with eigenvalue λ^{-1}.

Solution: By definition, this means $A\vec{v} = \lambda\vec{v}$. Then

$$\vec{v} = A^{-1}A\vec{v} = A^{-1}(\lambda\vec{v}) = \lambda A^{-1}\vec{v}$$

We know $\lambda \neq 0$ since A is invertible, so division by λ is valid, giving $\lambda^{-1}\vec{v} = A^{-1}\vec{v}$, which proves the result.

(b) A square and symmetric matrix A is said to be positive semidefinite (PSD) ($A \succeq 0$) if $\forall \vec{v} \neq 0, \vec{v}^T A\vec{v} \geq 0$. Show that A is PSD if and only if all of its eigenvalues are nonnegative.

Hint: Use the eigendecomposition of the matrix A.

Solution: *Start with the reverse direction. We wish to prove: if eigenvalues are nonnegative, A is PSD.*

The spectral theorem of A allows us to decompose a symmetric matrix A into $U\Lambda U^T$, where Λ is diagonal with eigenvalues λ_i as its non-zero entries, U is orthonormal. Define $z = U^T v$; since U is orthonormal, there exists a one-to-one mapping between all z, v.

$$\vec{v}^T A\vec{v} = \vec{v}^T (U\Lambda U^T)\vec{v} = z^T \Lambda z = \Sigma_{i=1}^n \lambda_i z_i^2$$

We assume $\lambda_i \geq 0$, so $\forall \vec{v}, \vec{v}^T A\vec{v} = \sum_{i=1}^n \lambda_i z_i^2 \geq 0$, which is the definition of PSD.

Take the forward direction. We wish to prove: if A is PSD, the eigenvalues are nonnegative.

Since A is PSD, we know $\forall \vec{x}, \vec{x}^T A\vec{x} \geq 0$. So for all i, take the ith eigenvector u_i for A. Then,

$$u_i^T A u_i = u_i^T (\lambda_i u_i) = \lambda_i u_i^T u_i = \lambda_i \|u_i\|_2^2 \geq 0$$

Since $\lambda_i \|u_i\|_2^2 \geq 0$ and $\|u_i\|_2^2 \geq 0$, we must have that $\lambda_i \geq 0$

3 Least Squares (using vector calculus)

(a) In ordinary least-squares linear regression, there is typically no $\vec{x}$ such that $A\vec{x} = \vec{y}$ (these are typically overdetermined systems — too many equations given the number of unknowns). Hence, we need to find an approximate solution to this problem. The residual vector will be $\vec{r} = A\vec{x} - \vec{y}$ and we want to make it as small as possible. The most common case is to measure the residual error with the standard Euclidean 2-norm. So the problem becomes:

$$\min_{\vec{x}} \|A\vec{x} - \vec{y}\|_2^2$$

Where $A \in \mathbb{R}^{m\times n}, \vec{x} \in \mathbb{R}^n, \vec{y} \in \mathbb{R}^m$. Derive using vector calculus an expression for an optimal estimate for $\vec{x}$ for this problem assuming A is full rank.

Solution: Take the gradient first, and set to 0. We'll elaborate on how the gradient is taken below.

$$\begin{aligned}
\nabla \|A\vec{x} - \vec{y}\|_2^2 &= 0 \\
2A^T (A\vec{x} - \vec{y}) &= 0 \\
2A^T A\vec{x} - 2A^T y &= 0 \\
2A^T A\vec{x} &= 2A^T y \\
A^T A\vec{x} &= A^T y \\
\vec{x} &= (A^T A)^{-1} A^T y
\end{aligned}$$

To take the gradient rigorously, we expand the L2-norm. First, note the following:

$$\frac{\partial \vec{x}^T B\vec{x}}{\partial x} = (B + B^T)\vec{x}$$

$$\frac{\partial x^T b}{\partial x} = b$$

We start by expanding the L2-norm:

$$\begin{aligned}
&\nabla(A\vec{x}-\vec{y})^T(A\vec{x}-\vec{y}) \\
&= \nabla((A\vec{x})^T(A\vec{x}) - (A\vec{x})^T(\vec{y}) - \vec{y}^T(A\vec{x}) + \vec{y}^T\vec{y}) && \text{Combine middle terms, identical scalars.} \\
&= \nabla(\vec{x}^T A^T A\vec{x} - 2\vec{x}^T A^T\vec{y} + \vec{y}^T y) && \text{Apply two derivative rules above} \\
&= (A^T A + A^T A)\vec{x} - 2A^T\vec{y} \\
&= 2A^T(A\vec{x}-\vec{y})
\end{aligned}$$

(b) What should we do if A is not full rank?

Solution: Basic idea: If $A \in \mathbb{R}^{m\times n}$ is not full rank, there is no unique answer. One possibility is to use the solution that minimizes the norm of $\vec{x}$. This solution is known as the pseudo-inverse $A^\dagger$. More intuitively, $A^\dagger$ behaves most similarly to the inverse: it is the matrix that, when multiplied by A, minimizes distance to the identity. $A^\dagger = \text{argmax}_{X\in\mathbb{R}^{n\times m}} \|AX - I_m\|_F$.