MACHINE LEARNING NOTES

Alfred O. Hero

September 9, 2019

Contents

1 Matrices and Linear Algebra Background

Keywords: vectors and matrices, matrix operations, matrix inverse identities.

Machine learning starts with data which is often multi-indexed over different index sets, e.g., collection time, collection place, individuals in a population, and attributes of these individuals. Thus they are best described as multi-indexed data arrays. When we collapse all but one of the index sets into a single index set, e.g., by lexigographic ordering, the multiindexed array becomes a matrix. The representation and manipulation of data therefore reposes on the the properties of matrices, for which linear algebra is the domain of discourse. Here some elementary concepts in linear algebra are reviewed.

Vector valued and matrix valued quantities, e.g. $\mathbf{x}$ and $\mathbf{A}$, are denoted by bold font symbols. Matrices are always in upper case but sometimes vectors may also be in upper case, e.g., when we wish to differentiate random variables (upper case) from their realizations (lower case). The elements of an $m \times n$ matrix $\mathbf{A}$ are denoted generically by a_{ij}, where i is the row index and j is the column index. We will try to avoid the longhand notation for this matrix ($m < n$ here)

$$\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & \cdots & a_{1n} \\ \vdots & \ddots & \ddots & \vdots \\ a_{m1} & \cdots & \cdots & a_{mn} \end{bmatrix}.$$

The following shorthand notation will be preferred

$$\mathbf{A} = (a_{ij})_{i,j=1}^{m,n}.$$

1.1 Row and column vectors in $\mathbb{R}^n$

A n-element vector $\mathbf{x} \in \mathbb{R}^n$ is an ordered list of n scalar elements x_i, generally assumed to be real valued, denoted as $x_i \in \mathbb{R}$:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix},$$

which resides in $\mathbb{R}^n$. This is to be distinguished from the unordered set of values

$$\{x_1, \ldots, x_n\}.$$

Convention: unless otherwise specified, $\mathbf{x}$ is always a column vector. Its transpose is the row vector

$$\mathbf{x}^T = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}$$

Some common vectors we will see are the vector of all ones

$$\mathbf{1} = [1, \ldots, 1]^T$$

the vector of all zeros,

$$\mathbf{0} = [0, \ldots, 0]^T$$

and the j-th elementary vector, which is the j-th column of the identity matrix

$$\mathbf{e}_j = [0, \ldots, 0, \underbrace{1}_{j-th}, 0, \ldots 0]^T$$

1.2 Inner and outer products

For two vectors $\mathbf{x}$ and $\mathbf{y}$ with the same number n of entries, their inner product (sometimes called the dot product) is the scalar

$$\mathbf{x}^T\mathbf{y} = \sum_{i=1}^n x_i y_i$$

The inner product is sometimes denoted with bracket notation ¡$\mathbf{x}, \mathbf{y}$¿ and sometimes as dot notation $\mathbf{x} \cdot \mathbf{y}$.

For two vectors $\mathbf{x}$ and $\mathbf{y}$ of possibly different dimensions n, m their "outer product" is the $n \times m$ matrix

$$\begin{aligned}
\mathbf{x}\mathbf{y}^T &= (x_i y_j)_{i,j=1}^{n,m} \\
&= [\mathbf{x}y_1, \ldots, \mathbf{x}y_m] \\
&= \begin{bmatrix} x_1 y_1 & \cdots & x_1 y_m \\ \vdots & \ddots & \vdots \\ x_n y_1 & \cdots & x_n y_m \end{bmatrix}
\end{aligned}$$

1.3 Vector norms: measuring length

The norm of a vector $\mathbf{x} \in \mathbb{R}^n$ is a measure of its length" or magnitude. All valid vector norms satisfy the following properties for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

- $\|\mathbf{x}\| \geq 0$ (non-negativity)
- $\|c\mathbf{x}\| = c\|\mathbf{x}\|$ (scaling property)
- $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (triangle inequality)

The following are examples of vector-norms that will be useful to us.

- The Euclidean norm, also known as the vector-l_2 norm, is simply the square root of the inner product of $\mathbf{x}$ with itself

  $$\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^T\mathbf{x}} = \sqrt{\sum_{i=1}^n x_i^2}$$

  The set $\{\mathbf{x} : \|\mathbf{x}\|_2 \leq r\}$ is convex. It is shaped as a sphere centerered at $\mathbf{0}$ of radius r.

- The vector-l_1 norm is the sum of the magnitudes of the elements of $\mathbf{x}$

  $$\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$$

  The set $\{\mathbf{x} : \|\mathbf{x}\|_1 \leq r\}$ is convex. It is shaped as a diamond centered at $\mathbf{0}$ with vertices located on the coordinate axes of $\mathbb{R}^n$ at coordinate location r.

- The vector-l_∞ norm is the maximum magnitude element of $\mathbf{x}$

  $$\|\mathbf{x}\|_\infty = \max_i\{|x_1|, \ldots, |x_n|\}$$

  The set $\{\mathbf{x} : \|\mathbf{x}\|_\infty \leq r\}$ is convex. It is shaped as a cube centered at $\mathbf{0}$ whose faces are each orthogonal to one of the axes of $\mathbb{R}^n$ and intersect at location r.

- The above are all special cases of the vector-l_p norm, defined for any $p > 0$

  $$\|\mathbf{x}\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$$

  The set $\{\mathbf{x} : \|\mathbf{x}\|_p \leq r\}$ is convex when $p \geq 1$ but is not convex when $p < 1$.

A useful relation is Holder's inequality, which holds when $\frac{1}{p} + \frac{1}{q} = 1$

$$|\mathbf{x}\mathbf{y}| \leq \|\mathbf{x}\|_p \ \|\mathbf{y}\|_q$$

This gives rise to the Cauchy-Schwarts inequality when specialized to $p = q = 2$

$$|\mathbf{x}\mathbf{y}| \leq \|\mathbf{x}\|_2 \ \|\mathbf{y}\|_2$$

and to the dual norm inequality when $p = 1$, $q = \infty$

$$|\mathbf{x}\mathbf{y}| \leq \|\mathbf{x}\|_1 \ \|\mathbf{y}\|_\infty.$$

1.4 Inner products: orthogonality and co-linearity

The inner product measures the amount of linear dependence between two (non-zero) vectors as it can be represented in terms of an angle θ:

$$\mathbf{x}^T\mathbf{y} = \|\mathbf{x}\|_2\|\mathbf{y}\|_2\cos(\theta).$$

When $\theta = \pm\pi/2$, i.e., $\mathbf{x}^T\mathbf{y} = 0$, $\mathbf{x}$ and $\mathbf{y}$ are said to be orthogonal, denoted $\mathbf{x} \perp \mathbf{y}$. When $\theta = 0$ the vectors are co-linear, i.e., $\mathbf{y} = a\mathbf{x}$ for some $a \neq 0$

If $\mathbf{x} \perp \mathbf{y}$ and $\|\mathbf{x}\| = \|\mathbf{y}\| = 1$, then $\mathbf{x}$ and $\mathbf{y}$ are said to be orthonormal vectors.

1.5 Matrix vector multiplication

Let $\mathbf{A}$ be an $m \times n$ matrix and let $\mathbf{x}$ be any n-element vector.

The product $\mathbf{Ax}$ is a (column) vector composed of linear combinations of the n columns $\mathbf{a}_i$ of $\mathbf{A}$

$$\mathbf{Ax} = \sum_{i=1}^n x_i \ \mathbf{a}_i$$

For $\mathbf{y}$ an m-element vector the product $\mathbf{y}^T\mathbf{A}$ is a (row) vector composed of linear combinations of the m rows, denoted here as $\mathbf{a}^{(i)}$, of $\mathbf{A}$

$$\mathbf{y}^T\mathbf{A} = \sum_{i=1}^m y_i \ \mathbf{a}_{(i)}.$$

1.6 Linear span and linear subspaces

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a set of (column) vectors in $\mathbb{R}^p$ and construct the $p \times n$ matrix

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n].$$

Let $\mathbf{a} = [a_1, \ldots, a_n]^T$ be a vector of coefficients. Then $\mathbf{y} = \sum_{i=1}^n a_i \mathbf{x}_i = \mathbf{X}\mathbf{a}$ is another p dimensional vector that is a linear combination of the columns of $\mathbf{X}$. The linear span of the vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$, equivalently, the column space or range of $\mathbf{X}$, is defined as the linear subspace of $\mathbb{R}^p$ that contains all such linear combinations:

$$\text{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_n\} = \{\mathbf{y} : \mathbf{y} = \mathbf{X}\mathbf{a},\ \mathbf{a} \in \mathbb{R}^n\}.$$

This subspace is denoted as colspan($\mathbf{X}$) or, equivalently, colspace($\mathbf{X}$). In other words, when we allow $\mathbf{a}$ to sweep over its entire domain $\mathbb{R}^n$, $\mathbf{y}$ sweeps over the linear subspace of $\mathbb{R}^p$ spanned by $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Likewise the of $\mathbf{X}$ is the of all linear combinations of the (transposed) rows of $\mathbf{X}$, which is denoted rowspan($\mathbf{X}$) and is equal to colspan($\mathbf{X}^T$), a lineear subspace of $\mathbb{R}^n$. We will sometimes denote the (transposed) rows of $\mathbf{X}$ as the vectors $\mathbf{x}^{(i)} \in \mathbb{R}^n$ for which $\mathbf{X}$ has the representation

$$\mathbf{X} = \left[\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(p)}\right]^T$$

1.7 Linear independent sets, bases and dimension

The set of p vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$, $\mathbf{x}_i \in \mathbb{R}^n$, are linearly independent when no non-trivial linear combination of them equals zero, i.e., $\sum_{i=1}^p a_i \mathbf{x}_i = \mathbf{0}$ implies that $a_i = 0$, for all i. More compactly

$$\sum_{i=1}^p a_i \mathbf{x}_i = \mathbf{0} \Rightarrow a_i = 0,\ \forall i.$$

If there exist $\{a_i\}_{i=1}^p$'s not identically zero for which the linear combination is equal to zero, then the set $\{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ is said to be linearly dependent. In the case where the set $\{\mathbf{x}_i\}_{i=1}^p$ is linear independent the set is called a basis for the subspace span$\{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$. The dimension of a subspace is the cardinality of any basis spanning the subspace, equal to p for the linearly independent set span$\{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$. We have the following facts:

- The basis for a linear space is not unique. For example, replacing $\mathbf{x}_1, \mathbf{x}_2$ with $\mathbf{x}_1 + \mathbf{x}_2, \mathbf{x}_1 - \mathbf{x}_2$ (a rotation) gives another basis that spans the same linear space.

- If for $\mathbf{x}_i \in \mathbb{R}^n$ the vectors $\{\mathbf{x}_i\}_{i=1}^p$ are orthogonal then they are also linear independent.

- If $\mathbf{x}_i \in \mathbb{R}^n$ and $p > n$ the set $\{\mathbf{x}_i\}_{i=1}^p$ is necessarily linear dependent. In particular, there exists a set of coefficients a_i such that for some $k \in \{1, \ldots, p\}$, $\mathbf{X}_k$ can be perfectly reconstructed as a linear combination of the remaining vectors:

  $$\mathbf{x}_k = \sum_{i \neq k}^p a_i \mathbf{x}_i.$$

  In this sense linear dependent sets of vectors have perfectly predictable components (members of the set can be found by linear combination of the other members).

1.8 Matrix rank and dimension of column space

The column rank of a matrix $\mathbf{A}$ is equal to the dimension of the column space of $\mathbf{A}$. The row rank of $\mathbf{A}$ is equal to the column rank of $\mathbf{A}^T$. If a $p \times n$ matrix $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_n]$ has $\text{rank}(\mathbf{A}) = \min(p, n)$ then it is said to be full rank. If $\mathbf{A}$ is square ($p = n$) and full rank, i.e., $\text{rank}(\mathbf{A}) = n$, then it is non-singular. Non-singular matrices have inverses.

1.9 Matrix inversion

If $\mathbf{A}$ is non-singular square matrix then it has an inverse $\mathbf{A}^{-1}$ that satisfies the relation $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$. In the special case of a 2×2 matrix the matrix inverse is given by (Cramèr's formula)

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \quad \text{if} \quad ad \neq bc$$

Sometimes when a matrix has special structure its inverse has a simple form. The books by Graybill [5] and Golub and VanLoan [4] give many interesting and useful examples. Some results which we will need in this text are: the *Sherman-Morrison-Woodbury identity* (SMW)

$$[\mathbf{A} + \mathbf{U}\mathbf{V}^T]^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}[\mathbf{I} + \mathbf{V}^T\mathbf{A}^{-1}\mathbf{U}]^{-1}\mathbf{V}^T\mathbf{A}^{-1}, \tag{1}$$

where $\mathbf{A}, \mathbf{U}, \mathbf{V}$ are compatible matrices, $[\mathbf{A} + \mathbf{U}\mathbf{V}^T]^{-1}$ and $\mathbf{A}^{-1}$ exist; and the *partitioned matrix inverse identity*

$$\begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} [\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}]^{-1} & -\mathbf{A}_{11}^{-1}\mathbf{A}_{12}[\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}]^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}[\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}]^{-1} & [\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}]^{-1} \end{bmatrix}, \tag{2}$$

assuming that all the indicated inverses exist.

A related identity to the SMW is the *Push-through identity*

$$[\mathbf{A} + \mathbf{U}\mathbf{V}^T]^{-1}\mathbf{U} = \mathbf{A}^{-1}\mathbf{U}[\mathbf{I} + \mathbf{V}^T\mathbf{A}^{-1}\mathbf{U}]^{-1}, \tag{3}$$

1.10 Projection matrices

Let $p \geq n$ and assume that the matrix $\mathbf{A} \in \mathbb{R}^{p \times n}$ has full rank, i.e. it has linearly independent columns. The projection of a vector $\mathbf{x} \in \mathbb{R}^p$ onto the colspace of $\mathbf{A}$ is defined as a vector, denoted $\hat{\mathbf{x}}$, that satisfies two properties:

1. $\hat{\mathbf{x}} \in$colspace($\mathbf{A}$).
2. $\hat{\mathbf{x}}$ is the best approximation to $\mathbf{x}$ among all possible vectors in the colspan($\mathbf{A}$), in the sense of minimum norm of the residual approximation error

$$\hat{\mathbf{x}} = \text{argmin}_{\mathbf{v} \in \text{colspace}(\mathbf{A})} \|\mathbf{x} - \mathbf{v}\|.$$

The solution $\hat{\mathbf{x}}$ to the above residual norm minimization can be obtained by orthogonally projecting $\mathbf{x}$ into the column space of $\mathbf{A}$. Specifically, $\hat{\mathbf{x}} = \mathbf{\Pi}\mathbf{x}$ where $\mathbf{\Pi}$ is the $p \times p$ projection matrix:

$$\mathbf{\Pi} = \mathbf{A}[\mathbf{A}^T\mathbf{A}]^{-1}\mathbf{A}^T. \tag{4}$$

Note that the inverse in (4) exists due to the full rank assumption on $\mathbf{A}$.

A special case is the rank 1 projection matrix which projects to a one dimensional subspace. Specifically, let $\mathbf{w} \in \mathbb{R}^p$ and assume that $\|\mathbf{x}\| > 0$. The projection matrix (4) takes the simple form

$$\mathbf{\Pi} = \frac{\mathbf{w}\,\mathbf{w}^T}{\|\mathbf{w}\|^2} \tag{5}$$

This matrix will orthogonally project any vector $\mathbf{x} \in \mathbb{R}^p$ onto the line spanned by $\mathbf{w}$.

The matrix $\mathbf{\Pi}$ is idempotent, a property defined by two properties

- $\mathbf{\Pi}^2 = \mathbf{\Pi}$
- $\mathbf{\Pi}^T = \mathbf{\Pi}$

Any idempotent matrix is a projection and vice-versa. The first property simply says that projecting a vector $\mathbf{x}$ multiple times onto a supspace is futile - it suffices to apply the projection operator $\mathbf{\Pi}$ only once.

Associated with $\mathbf{\Pi}$ is the matrix $\mathbf{I} - \mathbf{\Pi}$, which is also idempotent and therefore a projection matrix. Note that

$$\mathbf{\Pi}(\mathbf{I} - \mathbf{\Pi}) = \mathbf{O}$$

where $\mathbf{O}$ is square matrix of zeros. Hence, when $\mathbf{I} - \mathbf{\Pi}$ is applied to $\mathbf{x}$ it yields a vector $\mathbf{e}$ that is orthogonal to colspace($\mathbf{A}$) since any vector $\mathbf{m}$ in this subspace must satisfy $\mathbf{\Pi m}$, we have $\mathbf{e}^T\mathbf{m} = \mathbf{x}(\mathbf{I} - \mathbf{\Pi})\mathbf{\Pi m} = 0$.

1.10.1 The projection theorem

The projection theorem formalizes the above and asserts the uniqueness of the norm minimizing vector $\hat{\mathbf{x}}$ described above.

Theorem 1 *Let H be a linear space, e.g., $H = \mathbb{R}^p$, and assume that M a linear subspace of H, e.g., colspace($\mathbf{A}$). Corresponding to any vector $\mathbf{x}$ in H, there is a unique vector $\mathbf{m}^*$ in M such that*

$$\|\mathbf{x} - \mathbf{m}^*\| \leq \|\mathbf{x} - \mathbf{m}\|$$

for all $\mathbf{m} \in M$. Furthermore, a necessary and sufficient condition that $\mathbf{m}^* \in M$ be the unique minimizing vector is that $\mathbf{x} - \mathbf{m}^*$ be orthogonal to M

Put into the context of the vector approximation problem developed above, the optimal residual error $\mathbf{e} = \mathbf{x} - \hat{\mathbf{x}}$ is orthogonal to any vector $\mathbf{a}$ in the column space of $\mathbf{A}$. To see that $\mathbf{m}^* = \mathbf{\Pi x}$ is in fact the unique minimizing vector in colspace($\mathbf{A}$), it suffices to show that $\mathbf{e} = \mathbf{x} - \mathbf{\Pi x}$ is orthogonal to any $\mathbf{m}$ in colspace($\mathbf{A}$). Simply observe that

$$\mathbf{e} = (\mathbf{I} - \mathbf{\Pi})\mathbf{x}$$

and, since $\mathbf{m} \in$colspace($\mathbf{A}$) we have $\mathbf{m} = \mathbf{\Pi m}$. Hence,

$$(\mathbf{I} - \mathbf{\Pi})\mathbf{m} = (\mathbf{I} - \mathbf{\Pi})\mathbf{\Pi m} = 0,$$

which is a consequence of the property $\mathbf{\Pi}(\mathbf{I} - \mathbf{\Pi}) = 0$

As pointed out in Wolfram MathWorld ”This theorem can be viewed as a formalization of the result that the closest point on a plane to a point not on the plane can be found by dropping a perpendicular.”

1.11 Orthogonal and unitary matrices

A real square matrix $\mathbf{A}$ is said to be orthogonal if all of its columns are orthonormal, i.e.,

$$\mathbf{A}^T\mathbf{A} = \mathbf{I}. \tag{6}$$

The generalization of orthogonality to complex matrices $\mathbf{A}$ is the property of being unitary,

$$\mathbf{A}^H\mathbf{A} = \mathbf{I}.$$

The relation (6) implies that if $\mathbf{A}$ is an orthogonal matrix it is invertible and has a very simple inverse

$$\mathbf{A}^{-1} = \mathbf{A}^T.$$

An example of an orthogonal matrix in 2D is the matrix $\mathbf{A}$ that rotates a vector $[x_1, x_2]^T$ by angle θ:

$$\mathbf{A} = \left[\begin{array}{cc} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{array}\right].$$

1.12 Gram-Schmidt orthogonalization of linearly independent sets

Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a set of n linearly independent p dimensional column vectors ($n \leq p$) whose linear span is the subspace $\mathcal{H}$. Gram-Schmidt orthogonalization is an algorithm that can be applied to this set of vectors to obtain a set of n orthogonal vectors $\mathbf{y}_1, \ldots, \mathbf{y}_n$ that spans the same subspace. This algorithm can also be used to compute the SVD and is called Arnoldi's method [15]. It has the following steps:

- **Step 1**: select $\mathbf{y}_1$ as an arbitrary starting point in $\mathcal{H}$. For example, choose any coefficient vector $\mathbf{a}_1 = [a_{11}, \ldots, a_{1n}]^T$ and define $\mathbf{y}_1 = \mathbf{X}\mathbf{a}_1$ where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$.
- **Step 2**: construct the other $n-1$ vectors $\mathbf{y}_2, \ldots, \mathbf{y}_n$ by the following recursive procedure:
  `For` $j = 2, \ldots, n$`:` $\mathbf{y}_j = \mathbf{x}_j - \sum_{i=1}^{j} K_i \mathbf{y}_{i-1}$ `where` $K_j = \mathbf{x}_j^T \mathbf{y}_{j-1} / \mathbf{y}_{j-1}^T \mathbf{y}_{j-1}$.

The above Gram-Schmidt procedure can be expressed in compact matrix form [13]

$$\mathbf{Y} = \mathbf{G}\mathbf{X},$$

where $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]$ and $\mathbf{G}$ is called the Gram-Schmidt matrix.

After each step $j = 1, \ldots, n$ the length of $\mathbf{y}_j$ is normalized, i.e., $\mathbf{y}_j$ is replaced by $\tilde{\mathbf{y}}_j = \mathbf{y}_j / \|\mathbf{y}_j\|$. In this way we produce a matrix $\mathbf{Y}$ with orthonormal columns that has identical column span as that of $\mathbf{X}$. Thes process is called Gram-Schmidt orthonormalization. The procedure requires an initial vector $\mathbf{y}_1$ to generate an orthonormal basis for $\mathbf{x}_1, \ldots, \mathbf{x}_p]$. In this way the Gram-Schmidt procedure is said to accomplish *completion of the basis* with respect to an initial vector $\mathbf{y}_1$. Often $\mathbf{y}_1$ selected as the first vector $\mathbf{x}_1$ in the matrix $\mathbf{X}$. The matrix formed from such a basis will have the structure

$$\mathbf{Y} = \left[\begin{array}{c} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_n \end{array}\right]$$

which is an orthogonal matrix since the basis is orthonormal: $\mathbf{Y}^T\mathbf{Y} = \mathbf{I}.$.

1.13 Singular value decomposition (SVD) and PCA

Let $p < n$ and consider a (short and fat) matrix $\mathbf{A} \in \mathbb{R}^{p\times n}$. The singular value decomposition (SVD) is a matrix factorization of the form

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

where

- $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_p] \in \mathbb{R}^{p\times p}$ is an orthogonal matrix of vectors $\mathbf{u}_i$ that constitute an orthonormal basis for $\mathbb{R}^p$.
- $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_n] \in \mathbb{R}^{n\times n}$ is an orthogonal matrix of vectors $\mathbf{v}_i$ that constitute an orthonormal basis for $\mathbb{R}^n$.
- $\mathbf{D} = [\text{diag}(\sigma_i), \mathbf{O}] = \mathbb{R}^{p\times n}$ with $\text{diag}(\sigma_i) \in \mathbb{R}^{p\times p}$ a diagonal matrix of rank ordered positive values $\sigma_1 \geq \ldots \geq \sigma_p$ and $\mathbf{O}$ a $p \times n$ matrix of zeros.

The vectors $\{\mathbf{u}_i\}_{i=1}^n$ and $\{\mathbf{v}_i\}_{i=1}^p$ are called the left and right singular vectors, respectively, of $\mathbf{A}$. The positive scalars $\{\sigma_i\}_{i=1}^p$ are called singular values of $\mathbf{A}$. The number of non-zero singular values is equal to the rank of $\mathbf{A}$.

An alternative form for the SVD uses the diagonal structure of $\mathbf{D}$ to express the $\mathbf{U}\mathbf{D}\mathbf{V}^T$ as a weighted sum of p orthogonal rank one matrices

$$\mathbf{A} = \sum_{i=1}^{p} \sigma_i \mathbf{u}_i \mathbf{v}_i^T. \tag{7}$$

This SVD representation motivates the terminology that the singular vectors $\mathbf{u}_i$ and $\mathbf{v}_i$ are *associated* with the singular value σ_i since the the sense that the remaining singular vectors do not multiply against σ_i.

The SVD has several important properties

- Principal components (PC): these are the top k singular vectors $\mathbf{u}_1, \ldots, \mathbf{u}_k$ in $\mathbf{U}$. They are associated with the k largest singular values $\sigma_1, \ldots, \sigma_k$ as seen in (7).
- PC approximation property: the best rank-k matrix approximation to $\mathbf{A}$ is the matrix $\mathbf{A}^*$ obtained by setting $\sigma_{k+1} \ldots \sigma_d$ to zero in (7)

  $$\mathbf{A}^* = \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^T,$$

  in the sense that among rank k matrices $\mathbf{C} \in \mathbb{R}^{p\times n}$, $\mathbf{A}^*$ achieves the minimum approximation error $\|\mathbf{A} - \mathbf{C}\|$ where $\|\mathbf{E}\|$ is any matrix-norm[1] that depends on the singular values of the matrix $\mathbf{E}$. See subsection on matrix norms below.
- Subspace projection property: Since the PC's $\mathbf{u}_1, \ldots, \mathbf{u}_p$ are all orthonormal, for any specified $k < p$ a vector $\mathbf{x} \in \mathbb{R}^p$ can be simply projected to the principal k-dimensional subspace $\text{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_k\}$ by sequentially accumulating the individual projections of $\mathbf{x}$ onto each dimension. For example, if $k = 2$ the projected $\mathbf{x}$ is

  $$\hat{\mathbf{x}} = (\mathbf{\Pi}_2 + \mathbf{\Pi}_1)\mathbf{x},$$

  where $\mathbf{\Pi}_i = \mathbf{u}_k \mathbf{u}_k^T$ is the $p\times p$ idempotent rank 1 matrix projecting onto $\mathbf{u}_k$ (recall the definition (5). This provides a scalable recursive method for exploring the principal subspaces if $\mathbf{A}$, useful for PCA.

[1] These matrix norms include the spectral norm, frobenius norm or nuclear norm

1.13.1 Principal component analysis (PCA)

Principal component analysis (PCA) uses the PC's to reduce the number of rows of $\mathbf{A}$ and in data science is often used for visualization and exploratory analysis of the information content of a data matrix $\mathbf{A}$ whose columns live in a high extrinsic dimension[2] p. PCA is typically applied when $\mathbf{A}$ is a data matrix whose columns are n instances, or samples, of high dimensional p-element feature vectors $\{\mathbf{x}_i\}_{i=1}^n$ collected from a database or a sensor. PCA works as follows.

For $k < p$ define the $p \times k$ matrix of PC's $\mathbf{U}_{1:k} = [\mathbf{u}_1, \ldots, \mathbf{u}_k]$. This is a reduced rank version of the full $p \times p$ matrix $\mathbf{U}$. Left multiplying $\mathbf{A}$ by $\mathbf{U}_{1:k}^T$ yields the following dimension reduced $k \times n$ matrix[3]:

$$\mathbf{A}_k = \mathbf{D}_k \mathbf{V}^T$$

where $\mathbf{D}_k = [\text{diag}(\sigma_1, \ldots, \sigma_k), \mathbf{O}] \in \mathbb{R}^{k\times n}$ is obtained from $\mathbf{D}$ by truncating its last $p-k$ rows. When k is small, e.g., 1, 2 or 3, $\mathbf{A}_k$ provides a visualizable "best approximation" of the high dimensional matrix $\mathbf{A}$. This can be very usefule when the extrinsic dimension p of $\mathbf{A}$ is large.

1.14 Eigenvalues of a symmetric matrix

If $\mathbf{R}$ is arbitrary $n \times n$ **symmetric matrix**, that is, $\mathbf{R}^T = \mathbf{R}$, then there exist a set of n orthonormal eigenvectors $\mathbf{u}_i$,

$$\mathbf{u}_i^T \mathbf{u}_j = \Delta_{ij} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}$$

and a set of associated eigenvectors λ_i such that:

$$\mathbf{R}\mathbf{u}_i = \lambda_i \mathbf{u}_i, \qquad i = 1, \ldots, n.$$

These eigenvalues and eigenvectors satisfy:

$$\begin{aligned} \mathbf{u}_i^T \mathbf{R} \mathbf{u}_i &= \lambda_i \\ \mathbf{u}_i^T \mathbf{R} \mathbf{u}_j &= 0, \quad i \neq j. \end{aligned}$$

1.15 Symmetric matrix eigendecomposition and diagonalization

Let $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$ be the $n \times n$ orthogonal matrix formed from the eigenvectors of a symmetric matrix $\mathbf{R}$. The matrix $\mathbf{U}$ can be used to diagonalize $\mathbf{R}$

$$\mathbf{U}^T \mathbf{R} \mathbf{U} = \mathbf{u}, \tag{8}$$

In cases of both symmetric $\mathbf{R}$ the matrix $\mathbf{u}$ is diagonal and real valued

$$\mathbf{u} = \text{diag}(\lambda_i) = \begin{bmatrix} \lambda_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \lambda_n \end{bmatrix},$$

where λ_i's are the eigenvalues of $\mathbf{R}$.

The expression (8) implies

$$\mathbf{R} = \mathbf{U}\mathbf{u}\mathbf{U}^T,$$

[2] The extrinsic dimension of the columns of $\mathbf{A} \in \mathbb{R}^{p\times n}$ is always equal to p. This is not to be confused with the span of the columns colspace($\mathbf{A}$) which can be of significantly lower dimension that p and is called the intrinsic dimension.

[3] This representation follows by orthogonality of the columns of $\mathbf{U}$

which is called the *eigendecomposition* (EVD) of $\mathbf{R}$. As $\mathbf{u}$ is diagonal, an equivalent summation form for this eigendecomposition is

$$\mathbf{R} \;=\; \sum_{i=1}^{n} \lambda_i \mathbf{u}_i \mathbf{u}_i^T. \tag{9}$$

1.15.1 EVD and SVD for Gram product matrices

Assume that $\mathbf{R}$ is equal to a Gram product of the form $\mathbf{R} = \mathbf{A}\mathbf{A}^T$, where $\mathbf{A}$ has SVD

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T$$

Then it is easy to verify that the EVD of $\mathbf{R}$ is closely related to the SVD of $\mathbf{A}$ as follows

$$\mathbf{R} = \mathbf{U}\mathbf{D}^2\mathbf{U}^T$$

Thus comparing to the EVD $\mathbf{R} = \mathbf{U}\mathbf{u}\mathbf{U}^T$ it is seen that the eigenvalues λ_i are equal to the squared singular values σ_i of $\mathbf{A}$, obtained from the SVD of $\mathbf{A}$. Specifically, assume that the eigenvalues λ_i of $\mathbf{R}$ have been rank ordered so that $\lambda_1 \geq \ldots \geq \lambda_p$. Then it is easily verified that

$$\lambda_i(\mathbf{R}) = \sigma_i^2(\mathbf{A}).$$

Furthermore the eigenvectors $\mathbf{U}$ of $\mathbf{R}$ are identical to the left singular vectors $\mathbf{U}$ of $\mathbf{A}$.

1.16 Positive definite (pd) and positive semi-definite (psd) matrices

For a square symmetric matrix $\mathbf{R}$ and a compatible vector $\mathbf{x}$, a quadratic form is the scalar defined by $\mathbf{x}^T\mathbf{R}\mathbf{x}$. The matrix $\mathbf{R}$ is positive semidefinite (psd)[4] for any $\mathbf{x}$

$$\mathbf{x}^T\mathbf{R}\mathbf{x} \geq 0. \tag{10}$$

$\mathbf{R}$ is positive definite (pd) if it is psd and the only $\mathbf{x}$ that will give "=" in (10) is $\mathbf{x} = \mathbf{0}$. More explicitly $\mathbf{R}$ is pd if

$$\mathbf{x}^T\mathbf{R}\mathbf{x} > 0, \quad \mathbf{x} \neq \mathbf{0}. \tag{11}$$

Examples of psd (pd) matrices:

- $\mathbf{R} = \mathbf{B}^T\mathbf{B}$ for arbitrary (pd) matrix $\mathbf{B}$
- $\mathbf{R}$ symmetric with only non-negative (positive) eigenvalues

Rayleigh-quotient: If $\mathbf{A}$ is a psd $n\times n$ matrix with eigenvalues $\{\lambda_i\}_{i=1}^n$ the Rayleigh quotient quadratic defined as the ratio of quadratic forms $\frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{x}}$. The Rayleigh quotient satisfies

$$\min(\lambda_i) \leq \frac{\mathbf{x}^T\mathbf{A}\mathbf{x}}{\mathbf{x}^T\mathbf{x}} \leq \max(\lambda_i)$$

where the lower bound is attained when $\mathbf{x}$ is the eigenvector of $\mathbf{A}$ associated with the minimum eigenvalue of $\mathbf{A}$ and the upper bound is attained by the eigenvector associated with the maximum eigenvalue of $\mathbf{A}$.

[4] Positive semidefinite has an equivalent terminology: non-negative definite

1.17 Square root factors of pd symmetric matrices

For any symmetric positive definite covariance matrix $\mathbf{R}$ there exists a positive definite square root factor $\mathbf{R}^{\frac{1}{2}}$ and a positive definite square root inverse factor $\mathbf{R}^{-\frac{1}{2}}$ which satisfy:

$$\mathbf{R} \;=\; \mathbf{R}^{\frac{1}{2}}\mathbf{R}^{T/2}, \text{ and } \mathbf{R}^{-1} = \mathbf{R}^{-T/2}\mathbf{R}^{-\frac{1}{2}}.$$

There are many possible factorizations of this type. This includes Cholesky factorization which yields upper and lower triangular factors, respectively. Another factorization yields symmetric (and therefore identical) factors given by the eigendecomposition of $\mathbf{R} = \mathbf{U}\mathbf{D}\mathbf{U}^T$, where we recall

- $\mathbf{D} = \text{diag}(\lambda_i)$ are (positive) eigenvalues of $\mathbf{R}$
- $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_p]$ are (orthogonal) eigenvectors of $\mathbf{R}$

As $\mathbf{U}^T\mathbf{U} = \mathbf{I}$

$$\mathbf{R} \;=\; \mathbf{U}\mathbf{D}\mathbf{U}^T = \mathbf{U}\mathbf{D}^{\frac{1}{2}}\mathbf{D}^{\frac{1}{2}}\mathbf{U}^T = \mathbf{U}\mathbf{D}^{\frac{1}{2}}\mathbf{U}^T\,\mathbf{U}\mathbf{D}^{\frac{1}{2}}\mathbf{U}^T$$

Corresponding we identify the square root factor

$$\mathbf{R}^{\frac{1}{2}} = \mathbf{U}\mathbf{D}^{\frac{1}{2}}\mathbf{U}^T.$$

Due to orthogonality, $\mathbf{U}^{-1} = \mathbf{U}^T$ we can also identify an inverse square root factor

$$\mathbf{R}^{-\frac{1}{2}} = \mathbf{U}\mathbf{D}^{-\frac{1}{2}}\mathbf{U}^T$$

It is immediately verifiable that these have the desired properties of square root factors and are in addition symmetric.

1.18 Partition matrix identities for symmetric matrices

If $\mathbf{A}$ is a symmetric matrix with partition representation (2) then it is easily shown that

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & -\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ \mathbf{O} & \mathbf{I} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{O}^T \\ \mathbf{O} & \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{O}^T \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{I} \end{bmatrix}^{-1}, \quad (12)$$

as long as $\mathbf{A}_{22}^{-1}$ exists. Here $\mathbf{O}$ denotes a block of zeros. This implies: if $\mathbf{A}$ is positive definite the matrices $\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$ and $\mathbf{A}_{22}$ are pd. By using an analogous identity we can conclude that $\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}$ and $\mathbf{A}_{11}$ are also pd.

1.19 Determinant of a matrix

If $\mathbf{A}$ is any square matrix its determinant is the product of its eigenvalues

$$|\mathbf{A}| = \prod_i \lambda_i$$

Note: a square matrix is non-singular iff its determinint is non-zero.

If $\mathbf{A}$ is partitioned as in (2) and $\mathbf{A}_{11}^{-1}$ and $\mathbf{A}_{22}^{-1}$ exist then

$$|\mathbf{A}| = |\mathbf{A}_{11}||\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12}| = |\mathbf{A}_{22}||\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}| \quad (13)$$

This follows from the decomposition (12).

1.20 Trace of a matrix

For any square matrix $\mathbf{A} = ((a_{ij}))$ the trace of $\mathbf{A}$ is defined as

$$\text{tr}\{\mathbf{A}\} = \sum_i a_{ii} = \sum_i \lambda_i$$

The following property of the trace is useful. For compatible matrices $\mathbf{A}$ and $\mathbf{B}$

$$\text{tr}\{\mathbf{AB}\} = \text{tr}\{\mathbf{BA}\}.$$

This has the following implication for quadratic forms:

$$\mathbf{x}^T\mathbf{R}\mathbf{x} = \text{tr}\{\mathbf{x}\mathbf{x}^T\ \mathbf{R}\},$$

1.21 Matrix norms

Matrix norms arise frequently in optimization formulations of machine learning methods. A few of the most important ones are the following (For more information see Strang [15, Ch. 1.11]). In the following we assume $\mathbf{A} = (a_{ij})_{ij}$ is a $p \times n$ matrix with entries a_{ij} where $p < n$. Let $\mathbf{A}$ have singular values $\sigma_1, \ldots, \sigma_p$.

All valid matrix norms satisfy the properties

- $\|\mathbf{A}\| \geq 0$ (non-negativity)
- $\|c\mathbf{A}\| = c\|\mathbf{A}\|$ (scaling property)
- $\|\mathbf{A}+\mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$ (triangle inequality)
- $\|\mathbf{AB}\| \leq \|\mathbf{A}\|\ \|\mathbf{B}\|$ (Cauchy-Schwarz inequality)

Some matrix norms

- Any vector norm $\|\cdot\|$ on $\mathbb{R}^n$ induces a matrix norm on $\mathbb{R}^{p\times n}$ as follows

$$\|\mathbf{A}\| = \max_{\mathbf{u}\neq\mathbf{0}} \frac{\|\mathbf{Au}\|}{\|\mathbf{u}\|}$$

  Thus, in particular, the matrix-l_2 norm[5] is obtained by $\|\mathbf{u}\| = \|\mathbf{u}\|_2 = \sqrt{\mathbf{u}^T\mathbf{u}}$ and corresponds to $\|\mathbf{A}\|_2 = \sigma_1$, the maximum singular value of $\mathbf{A}$.

- The Frobenius norm is defined as

$$\|\mathbf{A}\|_F = \sqrt{\sum_{ij} a_{ij}^2} = \sqrt{\sum_{i=1}^p \sigma_i^2}$$

  Thus, the matrix-Frobenius norm is equal to the vector-l2 norm of the vector of singular values $[\sigma_1, \ldots, \sigma_p]^T$ of $\mathbf{A}$.

- The trace norm (also called the nuclear norm) is defined as

$$\|\mathbf{A}\|_* = \text{tr}\left((\mathbf{A}\mathbf{A}^T)^{1/2}\right) = \sum_{i=1} \sigma_i,$$

  where $(\mathbf{A}\mathbf{A}^T)^{1/2}$ denotes a square root factor of the psd matrix $\mathbf{A}\mathbf{A}^T$. Thus, the matrix-trace norm is equal to the l_1 norm of the vector of singular values of $\mathbf{A}$.

[5] Also called the spectral norm.

1.22 Vector differentiation for linear and quadratic forms

Differentiation of functions of a vector variable often arise in machine learning. If $\mathbf{h} = [h_1, \ldots, h_n]^T$ is an $n \times 1$ vector and $g(\mathbf{h})$ is a scalar function then the gradient of $g(\mathbf{h})$, denoted $\nabla g(\mathbf{h})$ or $\nabla_{\mathbf{h}} g(\mathbf{h})$ when necessary for conciseness, is defined as the (column) vector of partials

$$\nabla g = \left[\frac{\partial g}{\partial h_1}, \ldots, \frac{\partial g}{\partial h_n}\right]^T.$$

In particular, if c is a constant

$$\nabla_{\mathbf{h}} c = \mathbf{0},$$

if $\mathbf{x} = [x_1, \ldots, x_n]^T$

$$\nabla_{\mathbf{h}}(\mathbf{h}^T\mathbf{x}) = \nabla_{\mathbf{h}}(\mathbf{x}^T\mathbf{h}) = \mathbf{x},$$

and if $\mathbf{B}$ is an $n \times n$ matrix

$$\nabla_{\mathbf{h}}(\mathbf{h}-\mathbf{x})^T\mathbf{B}(\mathbf{h}-\mathbf{x}) = 2\mathbf{B}(\mathbf{h}-\mathbf{x}).$$

For a vector valued function $\mathbf{g}(\mathbf{h}) = [g_1(\mathbf{h}), \ldots, g_m(\mathbf{h})]^T$ the gradient of $\mathbf{g}(\mathbf{h})$ is an $m \times n$ matrix. In particular, for a scalar function $g(\mathbf{h})$, the two applications of the gradient $\nabla(\nabla g)^T$ gives the $n \times n$ Hessian matrix of g, denoted as $\nabla^2 g$. This yields useful and natural identities such as:

$$\nabla^2_{\mathbf{h}}(\mathbf{h}-\mathbf{x})^T\mathbf{B}(\mathbf{h}-\mathbf{x}) = 2\mathbf{B}.$$

For a more detailed discussion of vector and matrix differentiation the reader is referred to Boyd and Vandenberghe [2, Appendix A.4], Kay [8].

1.23 References

There are many useful textbooks that cover elements of linear algebra including the classic book by Noble and Daniel [11]. A recent book with many worked examples and applications to machine learning is the book by Strang [15]. More advanced books focused on computational linear algebra is Golub and Van Loan [4] which covers many fast and numerically stable algorithms arising in machine learning. Another nice book on linear algebra with emphasis on statistical applications is Graybill [5] that contains lots of useful identities.

2 BIBLIOGRAPHY

References

[1] D. Barber, *Bayesian reasoning and machine learning*, Cambridge University Press, 2012.

[2] S. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge university press, 2004.

[3] M. A. Garcia-Pérez, “On the confidence interval for the binomial parameter,” *Quality and quantity*, vol. 39, no. 4, pp. 467–481, 2005.

[4] G. H. Golub and C. F. Van Loan, *Matrix Computations (2nd Edition)*, The Johns Hopkins University Press, Baltimore, 1989.

[5] F. A. Graybill, *Matrices with Applications in Statistics*, Wadsworth Publishing Co., Belmont CA, 1983.

[6] J. A. Gubner, *Probability and Random Processes for Electrical and Computer Engineers*, Cambridge University Press, New York, 2006.

[7] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, "The elements of statistical learning: data mining, inference and prediction," *The Mathematical Intelligencer*, vol. 27, no. 2, pp. 83–85, 2005.

[8] S. M. Kay, *Statistical Estimation*, Prentice-Hall, Englewood-Cliffs N.J., 1991.

[9] E. D. Kolaczyk and G. Csárdi, *Statistical analysis of network data with R*, volume 65, Springer, 2014.

[10] K. P. Murphy, *Machine learning: a probabilistic perspective*, MIT press, 2012.

[11] B. Noble and J. W. Daniel, *Applied Linear Algebra*, Prentice Hall, Englewood Cliffs, NJ, 1977.

[12] S. Ross, *A first course in probability*, Prentice-Hall, Englewood Cliffs, N.J., 1998.

[13] L. L. Scharf, *Statistical Signal Processing: Detection, Estimation, and Time Series Analysis*, Addison-Wesley, Reading, MA, 1991.

[14] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*, Cambridge university press, 2014.

[15] G. Strang, *Linear algebra and learning from data*, Wellesley-Cambridge Press, 2019.