X Math Crash-Course for Machine Learning 1

Machine Learning 1: Foundations

Billy Joe Franks *(TUK)*

Linear Algebra & Analysis

We will recap the following topics in Linear Algebra and Analysis

- Vectors & Matrices
- Scalar Product & Projection
- Dimension Theorem
- Eigenvalues & Eigenvectors $\leftarrow$ today
- Matrix Decompositions
- Gradient
- Jacobian & Hessian Matrix

Eigenvalues

Let $A \in \mathbb{R}^{d \times d}$. $\lambda \in \mathbb{R}$ is called an **eigenvalue** of A if there is a vector $\mathbf{x} \in \mathbb{R}^d \backslash \{0\}$ such that $A\mathbf{x} = \lambda\mathbf{x}$. In that case $\mathbf{x}$ is an eigenvector corresponding to the eigenvalue λ. For $\lambda \in \mathbb{R}$ and $A \in \mathbb{R}^{d \times d}$ it holds:

$$\begin{aligned}
&\lambda \text{ Eigenvalue of } A \\
&\Leftrightarrow \exists\, \mathbf{x} \neq 0 \in \mathbb{R}^d \text{ with : } A\mathbf{x} = \lambda\mathbf{x} \\
&\Leftrightarrow \exists\, \mathbf{x} \neq 0 \in \mathbb{R}^d \text{ with : } A\mathbf{x} = \lambda I\mathbf{x} \\
&\Leftrightarrow \exists\, \mathbf{x} \neq 0 \in \mathbb{R}^d \text{ with : } \lambda I\mathbf{x} - A\mathbf{x} = 0 \\
&\Leftrightarrow \dim \operatorname{Ker}(\lambda I - A) > 0 \\
&\Leftrightarrow \dim \operatorname{Im}(\lambda I - A) < d \\
&\Leftrightarrow \lambda I - A \text{ not invertible} \\
&\Leftrightarrow \det(\lambda I - A) = 0
\end{aligned}$$

Eigenvalue example

Let us try and calculate the eigenvalues of

$$B = \begin{pmatrix} -6 & 3 \\ 4 & 5 \end{pmatrix}$$

By the last slide we get:

$$\det \begin{pmatrix} \lambda + 6 & -3 \\ -4 & \lambda - 5 \end{pmatrix} = 0 \iff (\lambda + 6)(\lambda - 5) - 12 = 0 \iff$$
$$\lambda^2 + \lambda - 42 = 0 \iff (\lambda + 7)(\lambda - 6) = 0$$

Apparently the eigenvalues of A are -7 and 6. We can also find the corresponding eigenvectors by resubstituting these values back into $B\mathbf{x} = \lambda\mathbf{x}$.

Facts about eigenvalues

Intuition on eigenvectors:

Eigenvectors preserve direction after the linear transformation, but not necessarily their length.(Clear from definition)

Here are some useful facts about eigenvalues and eigenvectors.

- The product of the eigenvalues is equal to the determinant of A
- If the eigenvalues of A are λ_i, and A is invertible, then the eigenvalues of $\mathbf{A}^{-1}$ are simply λ_i^{-1}.
- A can be inverted if and only if all eigenvalues are non-zero: $\lambda_i \neq 0 \quad \forall i$
- The eigenvectors of $\mathbf{A}^{-1}$ are the same as the eigenvectors of $\mathbf{A}$.
- Eigenvectors of real symmetric matrices are orthogonal.

Invertible Matrices

The matrix $I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$ is the identity matrix.

A matrix A is said to be invertible(regular)(non-singular) if there exists a matrix A^{-1} with

$$AA^{-1} = A^{-1}A = I$$

The following characterizations are equivalent.

- A is invertible
- The determinant det(A) is non-zero.
- The row vectors, or coloumn vectors of A are linearly independent.
- The eigenvalues of A are non-zero.

Matrix Properties

Let $A \in \mathbb{R}^{d \times d}$ be a symmetric matrix ($A^\top = A$). We call
A positive definite $:\iff \forall \mathbf{x} \neq 0 \in \mathbb{R}^d: \; \mathbf{x}^T A\mathbf{x} > 0.$
A negative definite $:\iff \forall \mathbf{x} \neq 0 \in \mathbb{R}^d: \; \mathbf{x}^T A\mathbf{x} < 0.$
A positive semi definite $:\iff \forall \mathbf{x} \neq 0 \in \mathbb{R}^d: \; \mathbf{x}^T A\mathbf{x} \geq 0.$
A negative semi definite $:\iff \forall \mathbf{x} \neq 0 \in \mathbb{R}^d: \; \mathbf{x}^T A\mathbf{x} \leq 0.$
A orthogonal $:\iff A^\top A = AA^\top = I$
A diagonal: $\iff$ all values not on the diagonal are zero.

Spectral Decomposition for real valued matrices

Let A be a real valued symmetric matrix. Then we can decompose A as

$$A = Q\Lambda Q^\top$$

where Q is an orthogonal matrix whose columns are the eigenvectors of A, and Λ is a diagonal matrix whose entries are the eigenvalues of A

Singular Value Decomposition

Any real valued matrix $A \in \mathbb{R}^{m \times n}$ can be decomposed as

$$A = UDV^\top$$

with $U \in \mathbb{R}^{m \times m}$, $D \in \mathbb{R}^{m \times n}$, $V^\top \in \mathbb{R}^{n \times n}$. U and V are orthogonal and D is a diagonal matrix.

Gradient(Special Case of Jacobian)

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable. We define the gradient by:

$$\nabla f = \operatorname{grad} f =: \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right)^\top$$

We observe that ∇f is again a function. Each component of the gradient tells us how fast our function is changing in each direction.
To see how fast the change is at a point **p** at direction **v**, we would multiply $\nabla f(p) \cdot \mathbf{v}$.
Observe that this scalar product is maximized if **v** is parallel to $\nabla f(p)$ which shows that ∇f shows in the direction of the steepest ascent.

Jacobian

Suppose $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ is a function such that each of its first-order partial derivatives exist on $\mathbb{R}^n$. This function takes a point $\mathbf{x} \in \mathbb{R}^n$ as input and produces the vector $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^m$ as output. Then the Jacobian matrix of $\mathbf{f}$ is defined to be an $m \times n$ matrix, denoted by $\mathbf{J}$, whose (i, j) th entry is $\mathbf{J}_{ij} = \frac{\partial f_i}{\partial x_j}$, or explicitly

$$\mathbf{J} = \begin{bmatrix} \frac{\partial \mathbf{f}}{\partial x_1} & \cdots & \frac{\partial \mathbf{f}}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \nabla^{\mathrm{T}} f_1 \\ \vdots \\ \nabla^{\mathrm{T}} f_m \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

Hessian (Second Derivative Generalization)

Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is a function taking as input a vector $\mathbf{x} \in \mathbb{R}^n$ and outputting a scalar $f(\mathbf{x}) \in \mathbb{R}$. If all second partial derivatives of f exist and are continuous over the domain of the function, then the Hessian matrix $\mathbf{H}$ of f is a square $n \times n$ matrix, usually defined and arranged as follows:

$$\mathbf{H}_f = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}$$

or, by stating an equation for the coefficients using indices i and j $(\mathbf{H}_f)_{i,j} = \frac{\partial^2 f}{\partial x_i \partial x_j}$.

Hessian Properties

- The Hessian matrix of a function f is the Jacobian matrix of the gradient of the function f; that is: $\mathbf{H}(f(\mathbf{x})) = \mathbf{J}(\nabla f(\mathbf{x}))$.
- The Hessian matrix is symmetric.
- The Hessian matrix of a convex function is positive semi-definite.
- If the Hessian is positive-definite at x, then f attains an isolated local minimum at x.
- If the Hessian is negative-definite at x, then f attains an isolated local maximum at x.

Useful Derivatives

Prove them yourself!

- $\frac{\partial}{\partial \mathbf{x}} \mathbf{c}^\top \mathbf{x} = \mathbf{c}$
- $\frac{\partial}{\partial \mathbf{x}} A\mathbf{x} = A$
- $\frac{\partial}{\partial \mathbf{x}} \mathbf{x}^T A\mathbf{x} = \left(A + A^\top\right) \mathbf{x}$ (Way more useful than it looks !)
- $\frac{\partial}{\partial \mathbf{x}} ||\mathbf{x}||^2 = \frac{\partial}{\partial \mathbf{x}} \mathbf{x}^\top \mathbf{x} = 2\mathbf{x}$

These should suffice to take the derivative of most things!