Notes of Deep Learning [1]

Part I (chapter 2-5)

[1] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. The MIT Press.

Part I: Applied Math and Machine Learning Basics

- Chapter 2. Linear algebra
- Chapter 3. Probability and information theory
- Chapter 4. Numerical computation
- Chapter 5. Machine learning basics

Part I: Applied Math and Machine Learning Basics

- Chapter 2. Linear algebra

Linear Algebra

- Scalars, Vectors, Matrices and Tensors
- Notations:
  - s
  - $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$ $S = \{1,3\}, \boldsymbol{x_S}, \boldsymbol{x_{-S}}$
  - $\boldsymbol{A}$ $\boldsymbol{A}_{:,j}$ $\boldsymbol{A}_{i,:}$ $A_{i,j}$ $f(\boldsymbol{A})_{i,j}$
  - $\mathbf{A}$ $A_{i,j,k}$

Linear Algebra

- Operations:
  - Transpose:
  - Addition
  - Multiply by a scalar
  - Broadcasting
  - Multiply matrices and vectors:
    - matrix product: $\boldsymbol{C} = \boldsymbol{AB}$
    - element-wise product: $\boldsymbol{C} = \boldsymbol{A} \odot \boldsymbol{B}$
    - distributive, associative, not commutative for matrices

Linear Algebra

- Identity and inverse matrices:
  - Identity matrix
  - Matrix inverse:
    - $\boldsymbol{A^{-1}A = I_n}$
    - $\boldsymbol{Ax = b,\ x = A^{-1}b}$
    - (Should not actually be used in practice for limited precision on a digital computer)

Linear Algebra

- Linear dependence and span:
  - $\boldsymbol{A}_{m*n}\boldsymbol{x} = \boldsymbol{b}_m$:
    - 1 solution ($m = n$ $linear\ independent\ vectors$)
    - no solutions ($n < m$ $linear\ independent\ vectors$)
    - infinitely many solutions ($n > m$ $linear\ independent\ vectors$)
  - Linear combination:
    - $\boldsymbol{A}\boldsymbol{x} = \sum_i x_i \boldsymbol{A}_{:,i} = \sum_i c_i \boldsymbol{v}^{(i)}$
  - Span: a set of vectors all points obtainable by linear combination of the original vectors.
  - Column space/ Range of $\boldsymbol{A}$: span of the columns of $\boldsymbol{A}$.
  - Linear independent vectors: no vector in the set is a linear combination of the other vectors.
  - The column space of the matrix to encompass all of R^m $\Leftrightarrow$ the matrix contain at least one set of exactly m linearly independent columns
  - Singular: A square matrix with linearly dependent columns
  - it can still possible to solve $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ while $\boldsymbol{A}$ is not square or is singular, but we can not use matrix inversion to find the solution

Linear Algebra

- Norms: the size of vectors:
  - $f(\boldsymbol{x}) = 0 \Rightarrow \boldsymbol{x} = \boldsymbol{0}$
  - $f(\boldsymbol{x}+\boldsymbol{y}) \leq f(\boldsymbol{x}) + f(\boldsymbol{y})$ $(\textit{the triangle inequality})$
  - $\forall \alpha \in R,\ f(\alpha\boldsymbol{x}) = |\alpha| f(\boldsymbol{x})$
- L^p norm: $\|\boldsymbol{x}\|_p = \left(\sum_i |x_i|^p\right)^{\frac{1}{p}}$, for $p \in R, p \geq 1$
- For vectors:
  - 'L^0': # of nonzero elements
  - L^1: Manhattan distance
  - L^2: Eculidean distance
  - L^∞: Max norm $\|\boldsymbol{x}\|_\infty = \max_i |x_i|$
- For matrices:
  - Frobenius norm: $\|\boldsymbol{A}\|_F = \sqrt{\sum_{i,j} A_{i,j}^2}$
  - L^1: $\|\boldsymbol{A}\|_1 = \max_j \sum_i \|A_{i,j}\|$
  - L^2: $\|\boldsymbol{A}\|_2 = \sqrt{\lambda}$, where λ is the maximum eigen value of $\boldsymbol{A}^T\boldsymbol{A}$
  - L^∞: $\|\boldsymbol{A}\|_\infty = \max_i \sum_j \|A_{i,j}\|$

Linear Algebra

- Special kinds of matrices and vectors:
  - Diagonal matrices (not necessary to be square), $diag(\boldsymbol{v})$
  - Symmetric matrices
  - Unit vectors
  - Orthogonal, Orthonormal (orthogonal + unit norm)
  - Orthogonal matrix:
    - $\boldsymbol{A^T A = A A^T = I}$
    - $\boldsymbol{A^T = A^{-1}}$
    - There is no special term for a matrix whose rows or columns are orthogonal but not orthonormal.

Linear Algebra

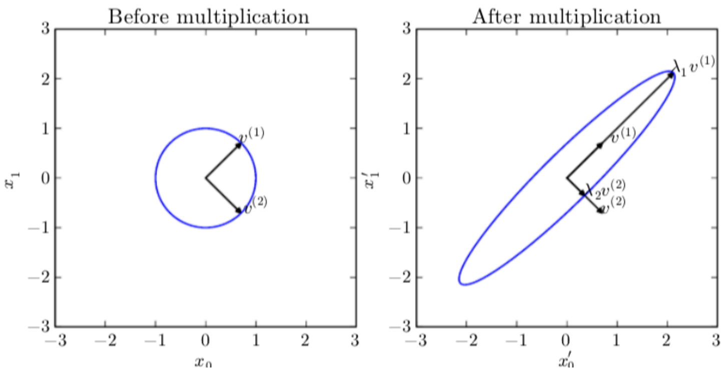

Figure 2.3: An example of the effect of eigenvectors and eigenvalues. Here, we have a matrix $\boldsymbol{A}$ with two orthonormal eigenvectors, $\boldsymbol{v}^{(1)}$ with eigenvalue λ_1 and $\boldsymbol{v}^{(2)}$ with eigenvalue λ_2. *(Left)* We plot the set of all unit vectors $\boldsymbol{u} \in \mathbb{R}^2$ as a unit circle. *(Right)* We plot the set of all points $\boldsymbol{Au}$. By observing the way that $\boldsymbol{A}$ distorts the unit circle, we can see that it scales space in direction $\boldsymbol{v}^{(i)}$ by λ_i.

- Eigendecomposition
  - (right) Eigenvector, eigenvalue: $\boldsymbol{Av} = \lambda \boldsymbol{v}$
  - Eigendecomposition (for square matrix): $\boldsymbol{A} = \boldsymbol{V} diag(\boldsymbol{\lambda})\boldsymbol{V}^{-1}$
  - Real symmetric matrix:
    - $\boldsymbol{A} = \boldsymbol{Q\Lambda Q}^T$, where $\boldsymbol{Q}$ is an orthogonral matrix composed of eigenvectors of $\boldsymbol{A}$ and$\boldsymbol{\Lambda}$ is a doagonal matrix
    - If any two or more eigenvectors share the same eigenvalue, then any set of orthogonal vectors lying in their span are also eigenvectors with that eigenvalue, and we could equivalently choose a $\boldsymbol{Q}$ using those eigenvectors instead
    - The matrix is singular if and only if any of the eigenvalues are 0
- Optimize quadratic expressions of the form

$$f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{Ax}, s.t. \|\boldsymbol{x}\|_2 = 1$$

- Positive definite/Positive semidefinite (Negative definite/ Negative semidefinite):
  - Positive definite:
    - Eigenvalues are all positive
    - $\forall \boldsymbol{x}, \boldsymbol{x}^T \boldsymbol{Ax} > 0$
    - $\boldsymbol{x}^T \boldsymbol{Ax} = 0 \Rightarrow \boldsymbol{x} = 0$
  - Positive semidefinite:
    - Eigenvalues are all positive or zero-valued
    - $\forall \boldsymbol{x}, \boldsymbol{x}^T \boldsymbol{Ax} \geq 0$

Linear Algebra

› Singular value decomposition(SVD)

$$\boldsymbol{A}_{m*n} = \boldsymbol{U}_{m*m}\boldsymbol{D}_{m*n}\boldsymbol{V}_{n*n}^{T}$$

Left-singular vectors:
Eigenvectors of $\boldsymbol{A}\boldsymbol{A}^T$

Elements along
the diagonal:
Singular Values

Right-singular vectors:
Eigenvectors of $\boldsymbol{A}^T\boldsymbol{A}$

Linear Algebra

- The Moore-Penrose Pseudoinverse
  - For solve $\boldsymbol{Ax} = \boldsymbol{y}$ by left-multiplying each side to obtain $\boldsymbol{x} = \boldsymbol{By}$
  - Definition: $\boldsymbol{A}^{+} = \lim_{\alpha \to 0} (\boldsymbol{A}^{T}\boldsymbol{A} + \alpha\boldsymbol{I})^{-1}\boldsymbol{A}^{T}$
  - In practice: $\boldsymbol{A} = \boldsymbol{UD}V^{T} \Rightarrow \boldsymbol{A}^{+} = V\boldsymbol{D}^{+}\boldsymbol{U}^{T}$
    - $\boldsymbol{D}^{+}$ of a diagonal matrix D is obtained by taking the reciprocal of its non-zero elements then taing the transpose of the resulting matrix
  - $\#\, of\ columns > \#\, of\ rows$:
    - There could be multiple possible solutions
    - $\boldsymbol{x} = \boldsymbol{A}^{+}\boldsymbol{y}\ with\ minimal\ \|\boldsymbol{x}\|_2$
  - $\#\, of\ columns < \#\, of\ rows$:
    - It is possible to have no solutions
    - $\boldsymbol{Ax}\ with\ minimal\ \|\boldsymbol{Ax} - \boldsymbol{y}\|_2$

Linear Algebra

- The trace operator: $Tr(\boldsymbol{A}) = \sum_i A_{i,i}$
  - $\|A\|_F = \sqrt{Tr(\boldsymbol{A}\boldsymbol{A}^T)}$
  - $Tr(\boldsymbol{A}) = Tr(\boldsymbol{A}^T)$
  - Invariance to cyclic permutation (even if the resulting product has a different shape):
$$Tr(\boldsymbol{ABC}) = Tr(\boldsymbol{CAB}) = Tr(\boldsymbol{BCA})$$

Linear Algebra

- The determinant: $det(\boldsymbol{A})$
  - $det(\boldsymbol{A}) = \text{product of all the eigenvalues of the matrix}$
  - A measure of how much multiplication by the matrix expands or contracts space:
    - If it is 0, then space is contracted completely along at least one dimension, causing it to lose all of its volume
    - If it is 1, then the transformation is volume-preserving

To be continued...