Tsinghua-Berkeley Shenzhen Institute

Learning From Data

Fall 2021

Review Session 1

Weida Wang, Guoqing Zhang September 14, 2021

Aim: This note is to review some basic mathematical knowledge on linear algebra, calculus and probability. We hope it can assist you in your future coursework.

1 Linear Algebra

1.1 Inner Product and trace

Definition 1. (Inner product). A function $\langle \cdot, \cdot \rangle$: $\mathbb{V} \times \mathbb{V} \to \mathbb{F}$ is an inner product if it satisfies [1]:

- **Linearity**: $\langle \alpha \boldsymbol{v} + \beta \boldsymbol{w}, \boldsymbol{x} \rangle = \alpha \langle \boldsymbol{v}, \boldsymbol{x} \rangle + \beta \langle \boldsymbol{w}, \boldsymbol{x} \rangle$;
- **Conjugate symmetry**: $\langle \boldsymbol{v}, \boldsymbol{w} \rangle = \overline{\langle \boldsymbol{w}, \boldsymbol{v} \rangle}$;
- **Positive definiteness**: $\langle \boldsymbol{v}, \boldsymbol{v} \rangle \geq 0$, with the equality iff $\boldsymbol{v} = 0$.

The most common one is the canonical inner product on $\mathbb{R}^n$. It says for vectors $\boldsymbol{x} \triangleq [x_1, \ldots, x_n]^{\mathrm{T}}$ and $\boldsymbol{y} \triangleq [y_1, \ldots, y_n]^{\mathrm{T}}$, we have

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle \triangleq x_1 y_1 + x_2 y_2 + \cdots x_n y_n = \sum_{i=1}^{n} x_i y_i = \boldsymbol{x}^{\mathrm{T}} \boldsymbol{y}.$$

Example 1: (Orthogonal Vectors) Vector $\boldsymbol{x} \in \mathbb{R}^n$ is orthogonal to $\boldsymbol{y} \in \mathbb{R}^n$ when $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$.

Example 2: (Unit Vector) Vector $\boldsymbol{x} \in \mathbb{R}^n$ is of unit length when $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = 1$.

Example 3: (Orthogonal Matrix) The matrix $\mathbf{Q} \in \mathbb{R}^{n \times n}$ is said to be **orthogonal** if

$$\mathbf{Q}\mathbf{Q}^{\mathrm{T}} = \mathbf{Q}^{\mathrm{T}}\mathbf{Q} = I$$

which implies that each column of $\mathbf{Q}$ has unit length and orthogonal to each other.

Definition 2. (Trace). For $\mathbf{M} \in \mathbb{R}^{n \times n}, \mathrm{trace}(\mathbf{M}) = \sum_{i=1}^{n} \mathbf{M}_{ii}$, where $\mathbf{M}_{ii}$ is the diagonal terms of matrix $\mathbf{M}$.

Theorem 1. For any matrices $\mathbf{A}, \mathbf{B}$ of compatible size,

$$\mathrm{trace}(\mathbf{AB}) = \mathrm{trace}(\mathbf{BA}).$$

1.2 Vector Norms

A norm on a vector space $\mathbb{V}$ gives a way of measuring lengths of vectors. Formally:

Definition 3. (Vector norm). A norm on a real vector space $\mathbb{V}$ is a function $||\cdot|| : \mathbb{V} \to \mathbb{R}$ that is:
- **Nonnegatively homogeneous**: $||\alpha\boldsymbol{x}|| = |\alpha|||\boldsymbol{x}||$ for all vectors $\boldsymbol{x} \in \mathbb{V}$, scalars $\alpha \in \mathbb{R}$;
- **Positive definite**: $||\boldsymbol{x}|| \geq 0$, and $||\boldsymbol{x}|| = 0$ iff $\boldsymbol{x} = 0$;
- **Subadditive**: $||\cdot||$ satisfies the triangle inequality $||\boldsymbol{x}+\boldsymbol{y}|| \leq ||\boldsymbol{x}|| + ||\boldsymbol{y}||$, for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{V}$.

One very important family of norms are the ℓ^p norms. If we take $\mathbb{V} = \mathbb{R}^n$, and $p \in [1, \infty)$, for vector $\boldsymbol{x} \triangleq [x_1, \ldots, x_n]^{\mathrm{T}}$, we have

$$||\boldsymbol{x}||_p = \left(\sum_i |\boldsymbol{x}_i|^p\right)^{\frac{1}{p}}. \tag{1}$$

The most frequent used one is the ℓ^2 norm or the "Euclidean norm",

$$||\boldsymbol{x}||_2 = \sqrt{\sum_{i=1}^{n} \boldsymbol{x}_i^2} = \sqrt{\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}}$$

which coincides with our usually way of measuring lengths. Two other cases are of almost equal importance: $p = 1$, and $p \to \infty$. Setting $p = 1$ in (1), we obtain $||\boldsymbol{x}||_1 = \sum_i |\boldsymbol{x}_i|$.
Finally, as p becomes larger, the expression in (1) accentuates the largest $|\boldsymbol{x}_i|$ among $\boldsymbol{x}$ entries. In another words, as $p \to \infty$, $||\boldsymbol{x}||_p \to \max_i |\boldsymbol{x}_i|$. Thus, we can extend the definition of the ℓ^p norm to $p = \infty$ by defining

$$||\boldsymbol{x}||_\infty = \max_i |\boldsymbol{x}_i|.$$

2 Calculus

2.1 Derivatives

Scalar b, vectors $\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{y}$ and matrix $\boldsymbol{A}$, we have :
- $\dfrac{\partial(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b)}{\partial\boldsymbol{x}} = \boldsymbol{w}$
- $\dfrac{\partial(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}\boldsymbol{x}+b)}{\partial\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{A}^{\mathrm{T}}\boldsymbol{x}$
- $\dfrac{\partial(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{A}^{-1}\boldsymbol{y})}{\partial\boldsymbol{A}} = -\boldsymbol{A}^{-\mathrm{T}}\boldsymbol{x}\boldsymbol{y}^{\mathrm{T}}\boldsymbol{A}^{-\mathrm{T}}$

For more derivative calculation, please refer to the `Matrix Cookbook`[2].

3 Probability

3.1 Basic Properties

For events E_1 and E_2, if they are disjoint, i.e. $E_1 \cap E_2 = \emptyset$, then $\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2)$

Definition 4. (Conditional probability) For events A and B, and $\mathbb{P}(A) > 0$,

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}$$

We can define the conditional expectation as

$$\mathbb{E}\left[Y|X = x\right] \triangleq \sum_{y \in \mathcal{Y}} y \cdot p\left(Y = y|X = x\right)$$

Definition 5. (Covariance) For two random variables X and Y, the covariance is defined by

$$\mathrm{Cov}\left[X, Y\right] = \mathbb{E}\left[XY\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]$$

When the covariance of X and Y is 0, we call them **uncorrelated variables**.

Definition 6. (Independent) For two random variables, when the joint pdf can be written as the product of two RVs' pdf

$$f\left(x, y\right) = f_X\left(x\right) f_Y\left(y\right),$$

we call them **independent**.

Theorem 2. We have:

◦ (Multiplication Rule) For events A and B,

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B|A) = \mathbb{P}(B)\mathbb{P}(A|B);$$

◦ (Total probability rule) $B_1, B_2, \ldots, B_k$ form a partition of Ω, $\forall i \neq j, B_i \cap B_j = \emptyset, \cup_{i=1}^{k} B_i = \Omega$, we have:

$$\mathbb{P}(A) = \sum_{i=1}^{k} \mathbb{P}(B_i)\mathbb{P}(A|B_i);$$

∘ (Bayes Rule)

$$\mathbb{P}(B_1|A) = \frac{\mathbb{P}(A \cap B_1)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B_1)\mathbb{P}(B_1)}{\mathbb{P}(A)} = \frac{\mathbb{P}(A|B_1)\mathbb{P}(B_1)}{\sum_{i=1}^{k} \mathbb{P}(A|B_i)\mathbb{P}(B_i)}.$$

3.2 Gaussian Distribution

3.2.1 Normal Distribution

• If random variable $X \in \mathbb{R}$, $X \sim \mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}$, then the density function of it is:

$$p(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

• $\mathbb{E}[X] = \mu$; $\mathrm{var}(X) = \sigma^2$.

3.2.2 Multivariate Gaussian Distribution

• If random variable $\boldsymbol{X} \in \mathbb{R}^n$, $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ is symmetric and positive semi-definite (PSD), then the density function of it is:

$$p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right)$$

• $\mathbb{E}[\boldsymbol{X}] = \boldsymbol{\mu}$; $\mathrm{cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$.

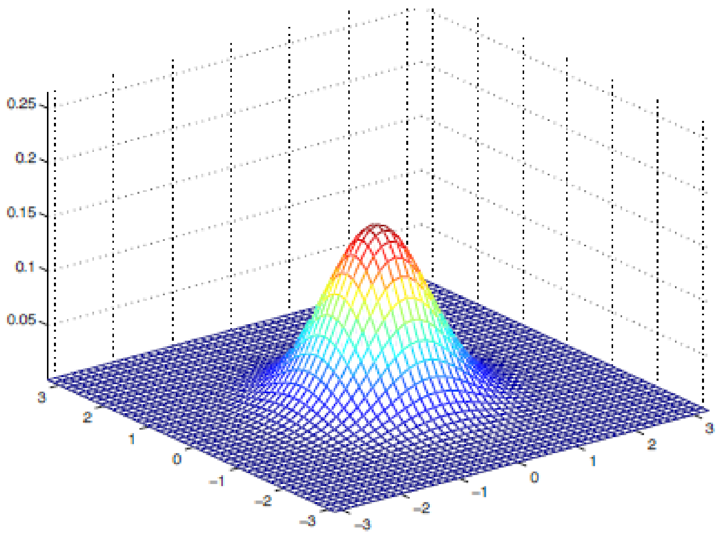

Figure 1: Multivariate Gaussian's p.d.f

References

[1] Strang, Gilbert, et al. Introduction to linear algebra. Vol. 3. Wellesley, MA: Wellesley-Cambridge Press, 1993.

[2] The Matrix Cookbook http://matrixcookbook.com