Academic Journal of Education
ISSN: 3080-9940 (Print), ISSN: 3080-9959 (Online) | Vol. 2, No. 1, 2026

Analytic Geometry in Machine Learning

Xiaolei Jiang

School of Control and Computer Engineering, North China Electric Power University, Baoding, Hebei, 071003, China

Abstract: The applications of analytic geometry in machine learning courses are summarized, covering the definition of hyperplanes, the computation of distances to hyperplanes, their application in support vector machines, and the geometric properties of hyperellipsoids—specifically their centers, axis orientations, and semi-axis lengths—along with the occurrence of such ellipsoids in linear regression. By explicitly addressing these topics, it facilitates both teaching and learning for instructors and students. Moreover, these geometric interpretations render abstract high-dimensional optimization problems more intuitive and help students develop the ability to analyze machine learning algorithms from a geometric perspective. Furthermore, the intrinsic connections between these concepts—such as the role of normal vectors in defining margins and the interpretation of level sets as ellipsoids—can deepen understanding and inspire further exploration.

Keywords: Hyperplanes; Hyperellipsoids; Support Vector Machine; Linear Regression.

1. Introduction

In machine learning courses, in addition to the foundational knowledge of linear algebra, calculus, and probability and statistics, analytic geometry involving hyperplanes and hyperellipsoids is also frequently employed [1]. For example, in support vector machines, hyperplanes are used to construct decision boundaries, where the normal vector determines the classification direction, and the process of maximizing the margin essentially boils down to finding the maximum distance between two parallel hyperplanes [2] [3]. In linear regression, the level sets of the data term in the loss function are hyperellipsoids centered at the least-squares solution, with their principal axis directions and semi-axis lengths determined by the singular values of the design matrix [2] [3]. However, many machine learning textbooks directly apply these geometric conclusions without providing detailed derivations, so students often struggle due to a lack of intuitive geometric understanding. We have systematically clarified the two core concepts—hyperplanes and hyperellipsoids—covering their definitions, geometric properties, and roles in specific algorithms. In particular, we illustrate how they are used to formulate objective functions and interpret optimization outcomes in the contexts of support vector machines and linear regression. We hope that this systematic exposition will help students build a bridge from geometric intuition to the derivation of formulas, lower the barrier to learning, and simultaneously provide instructors with a clear teaching thread that facilitates classroom organization and explanation.

The rest of this article is organized as follows. In Section 2, we introduce the definition of hyperplanes and derive the signed distance from a point to a hyperplane. Section 3 examines the role of hyperplanes in support vector machines, with particular emphasis on maximizing the margin. Section 4 turns to hyperellipsoids, discussing their geometric characterization through basis change and their connection to the eigenvalue decomposition of symmetric positive definite matrices. In Section 5, we show how hyperellipsoids naturally arise in linear regression, where the level sets of the least-squares data term are centered at the optimal solution. Finally, Section 6 concludes the article with a brief summary.

2. Hyperplanes: Distances

Consider the set of points determined by the equation

$$\boldsymbol{a}^T\boldsymbol{x}+\boldsymbol{b}=\boldsymbol{0}, \tag{1}$$

where $\boldsymbol{0}\neq\boldsymbol{a}\in\mathbb{R}^{\boldsymbol{n}}$ and $\boldsymbol{b}\in\mathbb{R}$. Let $\boldsymbol{x_0}$ be a point in this set, then $\boldsymbol{a}^T\boldsymbol{x_0}+\boldsymbol{b}=\boldsymbol{0}$, hence $\boldsymbol{a}^T(\boldsymbol{x_0}-\boldsymbol{x})=\boldsymbol{0}$. Thus, for any point $\boldsymbol{x}$ in this set, the vector $\boldsymbol{x}-\boldsymbol{x_0}$ is orthogonal to $\boldsymbol{a}$. For $\boldsymbol{n}=\boldsymbol{2}$, such a set is a line passing through $\boldsymbol{x_0}$ with $\boldsymbol{a}$ as a normal vector; for $\boldsymbol{n}=\boldsymbol{3}$, such a set is a plane passing through $\boldsymbol{x_0}$ with $\boldsymbol{a}$ as a normal vector. Therefore, the point set determined by equation (1) may be called a hyperplane, and $\boldsymbol{a}$ may be called the normal vector of this hyperplane.

When $\boldsymbol{n}=\boldsymbol{2}$ (or $\boldsymbol{3}$), the equation $\boldsymbol{a}^T\boldsymbol{x}+\boldsymbol{b}=\boldsymbol{0}$ defines a line (or a plane). Let $\boldsymbol{x}$ lie on the line (or plane) $\boldsymbol{a}^T\boldsymbol{x}+\boldsymbol{b}=\boldsymbol{0}$, i.e., $\boldsymbol{a}^T\boldsymbol{x}+\boldsymbol{b}=\boldsymbol{0}$. Then the distance from $\boldsymbol{x_0}$ to the line (or plane) $\boldsymbol{a}^T\boldsymbol{x}+\boldsymbol{b}=\boldsymbol{0}$ is

$$\frac{|\boldsymbol{a}^T(\boldsymbol{x}-\boldsymbol{x_0})|}{\|\boldsymbol{a}\|}=\frac{|\boldsymbol{a}^T\boldsymbol{x_0}+\boldsymbol{b}|}{\|\boldsymbol{a}\|}.$$

Thus, $\frac{\boldsymbol{a}^T\boldsymbol{x_0}+\boldsymbol{b}}{\|\boldsymbol{a}\|}$ is the signed distance from $\boldsymbol{x_0}$ to the line (or plane) $\boldsymbol{a}^T\boldsymbol{x}+\boldsymbol{b}=\boldsymbol{0}$: its absolute value is the distance, and its sign indicates whether $\boldsymbol{x_0}$ lies on the side pointed to by the normal vector $\boldsymbol{a}$ or on the opposite side. For a general positive integer $\boldsymbol{n}$, such a geometric interpretation is not directly available. Consider the optimization problem

$$\min\ \|\mathbf{x}-\mathbf{x}_0\|^2$$

$$\text{s.t. } \mathbf{a}^{\mathrm{T}}\mathbf{x}+b=0.$$

According to the KKT conditions, the minimizer satisfies the system

$$\begin{cases}2(\mathbf{x}-\mathbf{x}_0)+\mu\mathbf{a}=0\\ \mathbf{a}^{\mathrm{T}}\mathbf{x}+b=0,\end{cases}$$

from which it follows that the minimizer $\mathbf{x}=\mathbf{x}_0-\frac{\mathbf{a}^{\mathrm{T}}\mathbf{x}_0+b}{\mathbf{a}^{\mathrm{T}}\mathbf{a}}\mathbf{a}$, and the minimum value is $\left(\frac{\mathbf{a}^{\mathrm{T}}\mathbf{x}_0+b}{\|\mathbf{a}\|}\right)^2$. Therefore, $\frac{\mathbf{a}^{\mathrm{T}}\mathbf{x}_0+b}{\|\mathbf{a}\|}$ is the signed distance from $\mathbf{x}_0$to the hyperplane $\mathbf{a}^{\mathrm{T}}\mathbf{x}+b=0$. In other words, the geometric meaning of the affine function $\mathbf{a}^{\mathrm{T}}\mathbf{x}+b$ is $\|\mathbf{a}\|$ times the signed distance to the hyperplane $\mathbf{a}^{\mathrm{T}}\mathbf{x}+b=0$.

Consider the optimization problem

$$\min \|\mathbf{x}_1 - \mathbf{x}_2\|^2$$
$$\text{s.t. } \mathbf{a}^{\mathrm{T}}\mathbf{x}_1 + b_1 = 0,$$
$$\mathbf{a}^{\mathrm{T}}\mathbf{x}_2 + b_2 = 0,$$

where $0 \neq \mathbf{a} \in \mathbb{R}^n$ and $b_1, b_2 \in \mathbb{R}$. According to the KKT conditions, the minimizer satisfies the system

$$\begin{cases} 2(\mathbf{x}_1 - \mathbf{x}_2) + \mu_1\mathbf{a} = 0 \\ 2(\mathbf{x}_1 - \mathbf{x}_2) - \mu_2\mathbf{a} = 0 \\ \mathbf{a}^{\mathrm{T}}\mathbf{x}_1 + b_1 = 0 \\ \mathbf{a}^{\mathrm{T}}\mathbf{x}_2 + b_2 = 0, \end{cases}$$

from which it follows that the minimizer $(\mathbf{x}_1, \mathbf{x}_2)$ satisfies $\mathbf{x}_1 - \mathbf{x}_2 = \frac{b_1 - b_2}{\mathbf{a}^{\mathrm{T}}\mathbf{a}}\mathbf{a}$, and the minimum value is $\left(\frac{b_1 - b_2}{\|\mathbf{a}\|}\right)^2$. Therefore, the distance between hyperplane $\mathbf{a}^{\mathrm{T}}\mathbf{x}_1 + b_1 = 0$ and hyperplane $\mathbf{a}^{\mathrm{T}}\mathbf{x}_2 + b_2 = 0$ is $\frac{|b_1 - b_2|}{\|\mathbf{a}\|}$.

3. Hyperplanes in Supporting Vector Machines

Suppose the data set is $\{(\mathbf{x}_i, t_i)\}_{i=1}^m$, where $\mathbf{x}_i \in \mathbb{R}^n$, and $t_i \in \{-1,1\}$. Consider the case where $\{(\mathbf{x}_i, t_i)\}_{i=1}^m$ is linearly separable, that is, there exists a hyperplane $\mathbf{a}^{\mathrm{T}}\mathbf{x} + b = 0$ such that $t_i(\mathbf{a}^{\mathrm{T}}\mathbf{x}_i + b) \geq 0$ holds for $i = 1, \ldots, m$. Once we have a separable hyperplane $\mathbf{a}^{\mathrm{T}}\mathbf{x} + b = 0$, a new data $\mathbf{x}_0$ can be classified according to the sign of $\mathbf{a}^{\mathrm{T}}\mathbf{x}_0 + b$.

Usually, there exist infinitely many separating hyperplanes that can correctly classify the samples in the training set. The support vector machine selects the separating hyperplane that maximizes the distance from the samples to it, a distance known as the margin. This yields the following optimization problem:

$$\max\ r$$
$$\text{s.t. } t_i(\mathbf{a}^{\mathrm{T}}\mathbf{x}_i + b) \geq 0,\ i = 1, \ldots, m \qquad (2)$$
$$\frac{t_i(\mathbf{a}^{\mathrm{T}}\mathbf{x}_i + b)}{\|\mathbf{a}\|} \geq r.\ i = 1, \ldots, m$$

Note that $t_i(\mathbf{a}^{\mathrm{T}}\mathbf{x}_i + b) \geq 0$, hence $\frac{t_i(\mathbf{a}^{\mathrm{T}}\mathbf{x}_i + b)}{\|\mathbf{a}\|} = \frac{|\mathbf{a}^{\mathrm{T}}\mathbf{x}_i + b|}{\|\mathbf{a}\|}$ is the distance from $\mathbf{x}_i$ to the separating hyperplane $\mathbf{a}^{\mathrm{T}}\mathbf{x} + b = 0$. Since the data set is linearly separable, the optimization problem (2) can be rewritten as

$$\max\ r$$
$$\text{s.t. } \frac{t_i(\mathbf{a}^{\mathrm{T}}\mathbf{x}_i + b)}{\|\mathbf{a}\|} \geq r.\ i = 1, \ldots, m \qquad (3)$$

The optimization problem (3) can be reformulated into the following canonical form:

$$\min \|\mathbf{a}\|^2$$
$$\text{s.t. } t_i(\mathbf{a}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1.\ i = 1, \ldots, m$$

The sample points for which the equality holds are called support vectors. Note that the distance between the hyperplane $\mathbf{a}^{\mathrm{T}}\mathbf{x} + b = 1$ and the hyperplane $\mathbf{a}^{\mathrm{T}}\mathbf{x} + b = -1$ is $\frac{2}{\|\mathbf{a}\|}$.

4. Hyperellipsoids: Basis Change

Consider the hypersurface determined by the equation

$$(\boldsymbol{x} - \boldsymbol{\mu})^{\boldsymbol{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}) = \boldsymbol{c}, \qquad (4)$$

where $\boldsymbol{\Sigma}$ is a positive definite $n \times n$ matrix, $\mathbf{x}, \boldsymbol{\mu} \in \mathbb{R}^n$, and $c > 0$. Let the eigenvalues and normalized eigenvectors of $\boldsymbol{\Sigma}$ be given by $\boldsymbol{\Sigma}\mathbf{u}_i = s_i\mathbf{u}_i$, $i = 1, \ldots, n$. Let $\mathbf{U} = [\mathbf{u}_1 \cdots \mathbf{u}_n]$ be the eigenvector matrix and $\mathbf{S} = \mathrm{diag}(s_1, \ldots, s_n)$ be the eigenvalue matrix. Then $\boldsymbol{\Sigma}\mathbf{U} = \mathbf{U}\mathbf{S}$, hence $\mathbf{U}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{U} = \mathbf{S}^{-1}$. Set $\mathbf{x} - \boldsymbol{\mu} = \mathbf{U}\mathbf{y}$. Then

$$(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = \mathbf{y}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{y} = \sum_{i=1}^n \frac{y_i^2}{s_i}.$$

Thus, under the coordinate transformation $\mathbf{x} - \boldsymbol{\mu} = \mathbf{U}\mathbf{y}$, the hypersurface defined by equation (4) becomes $\mathbf{y}^{\mathrm{T}}\mathbf{S}^{-1}\mathbf{y} = c$. Note that when $\mathbf{y} = \mathbf{e}_i$, we have $\mathbf{x} = \boldsymbol{\mu} + \mathbf{u}_i$, $i = 1, \ldots, n$, where $\mathbf{e}_i$ is the i-th standard basis vector in $\mathbb{R}^n$. Consequently, the hypersurface given by equation (4) is a hyperellipsoid centered at $\boldsymbol{\mu}$, with its axes oriented along the eigenvectors $\mathbf{u}_i$ of $\boldsymbol{\Sigma}$, and the semi-axis length along each direction is proportional to the square root $\sqrt{s_i}$ of the corresponding eigenvalue of $\boldsymbol{\Sigma}$.

The geometric characterization of hyperellipsoids provided above is particularly valuable because many optimization problems in machine learning give rise to level sets of precisely this form. In the context of linear regression, as we shall see in the next section, the least-squares data term can be expressed as a quadratic form in the weight space, and its level sets are hyperellipsoids centered at the optimal solution. The geometric interpretation established here therefore provides a direct bridge to understanding the structure of such loss surfaces and the behavior of optimization algorithms.

5. Hyperellipsoids in Linear Regression

Let the input vector be $\mathbf{x} \in \mathbb{R}^d$ and the target variable be $t \in \mathbb{R}$. Define a feature map $\boldsymbol{\phi}: \mathbb{R}^d \longrightarrow \mathbb{R}^n$ from the original input space to an n-dimensional feature space:

$$\mathbf{x} \mapsto \boldsymbol{\phi}(\mathbf{x}) = \begin{bmatrix} \phi_1(\mathbf{x}) \\ \phi_2(\mathbf{x}) \\ \vdots \\ \phi_n(\mathbf{x}) \end{bmatrix},$$

where the component function $\phi_j(\cdot)$ is the j-th basis function (or feature function), $j = 1, \ldots, n$. The linear regression model assumes that the output y is a linear combination of these basis functions:

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=1}^n w_j\,\phi_j(\mathbf{x}) = \boldsymbol{\phi}(\mathbf{x})^{\mathrm{T}}\mathbf{w},$$

where

$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} \in \mathbb{R}^n$$

is the weight vector.

Suppose the training set consists of m samples, with input vectors $\{\mathbf{x}_i\}_{i=1}^m$ and corresponding target values $\{t_i\}_{i=1}^m$. Assemble all target values into a column vector

$$\mathbf{t} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_m \end{bmatrix} \in \mathbb{R}^m.$$

Applying the feature map $\boldsymbol{\phi}$ to each input sample $\mathbf{x}_i$ yields the corresponding feature vector $\boldsymbol{\phi}(\mathbf{x}_i) \in \mathbb{R}^n$. The design matrix $\boldsymbol{\Phi}$ is defined with these feature vectors as its columns:

$$\boldsymbol{\Phi} = [\boldsymbol{\phi}(\mathbf{x}_1) \quad \boldsymbol{\phi}(\mathbf{x}_2) \quad \cdots \quad \boldsymbol{\phi}(\mathbf{x}_m)] = \begin{bmatrix} \phi_1(\mathbf{x}_1) & \phi_1(\mathbf{x}_2) & \cdots & \phi_1(\mathbf{x}_m) \\ \phi_2(\mathbf{x}_1) & \phi_2(\mathbf{x}_2) & \cdots & \phi_2(\mathbf{x}_m) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_n(\mathbf{x}_1) & \phi_n(\mathbf{x}_2) & \cdots & \phi_n(\mathbf{x}_m) \end{bmatrix} \in \mathbb{R}^{n \times m}.$$

Using these notations, the model's prediction vector for all training samples is $\mathbf{\Phi}^{\mathrm{T}}\mathbf{w}$, whose i-th component is exactly $\boldsymbol{\phi}(\mathbf{x}_i)^{\mathrm{T}}\mathbf{w}$.

In linear regression, the data term in the loss function is usually taken as $\|\mathbf{\Phi}^{\mathrm{T}}\mathbf{w} - \mathbf{t}\|^2$. If $\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}$ is invertible, the least-squares solution is $\mathbf{w}_{\mathrm{LS}} = (\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}})^{-1}\mathbf{\Phi}\mathbf{t}$. The data term can be decomposed as

$$\|\mathbf{\Phi}^{\mathrm{T}}\mathbf{w} - \mathbf{t}\|^2 = \|\mathbf{\Phi}^{\mathrm{T}}(\mathbf{w} - \mathbf{w}_{\mathrm{LS}})\|^2 + \|\mathbf{\Phi}^{\mathrm{T}}\mathbf{w}_{\mathrm{LS}} - \mathbf{t}\|^2$$

$$= (\mathbf{w} - \mathbf{w}_{\mathrm{LS}})^{\mathrm{T}}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}(\mathbf{w} - \mathbf{w}_{\mathrm{LS}}) + \mathrm{const.}$$

Thus, the level set $\|\mathbf{\Phi}^{\mathrm{T}}\mathbf{w} - \mathbf{t}\|^2 = c$ is is equivalent to $(\mathbf{w} - \mathbf{w}_{\mathrm{LS}})^{\mathrm{T}}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}(\mathbf{w} - \mathbf{w}_{\mathrm{LS}}) = c'$. This equation represents a hyperellipsoid centered at $\mathbf{w}_{\mathrm{LS}}$ in the weight space $\mathbb{R}^n$. The principal axes of this hyperellipsoid are oriented along the eigenvectors of the matrix $(\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}})^{-1}$, and the semi-axis lengths are proportional to the square roots of the corresponding eigenvalues of $(\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}})^{-1}$.

The geometric interpretations derived above can be used for comparing the sparsity of solutions with ℓ_1 and ℓ_2 regularization terms, for illustrating the sequential updating of the posterior distribution of the parameter vector, and for providing further explanation of the evidence approximation.

6. Conclusion

We have presented a systematic exposition of the geometric interpretations of hyperplanes and hyperellipsoids and their roles in two fundamental machine learning algorithms. By explicitly linking the geometric properties—such as signed distances and axis orientations—to the optimization formulations of support vector machines and linear regression, students are able to develop a clearer and more intuitive understanding of how analytic geometry underpins these methods. This approach not only simplifies the derivation of key results, such as the margin maximization and the least-squares solution, but also provides students with a more structured perspective on how geometric concepts translate into algorithmic formulations. Ultimately, this exposition helps students bridge the gap between abstract geometry and practical machine learning methods, deepening their understanding of the underlying optimization principles and enhancing both their theoretical foundation and their ability to analyze related models. For instructors, this exposition offers a concise and coherent reference that can be easily integrated into lecture preparation.

Acknowledgments

The author thanks his colleges and students for their kindness and support.

References

[1] Deisenroth, M. P., Faisal, A. A., & Ong, C. S. (2020). *Mathematics for machine learning*. Cambridge University Press. https://doi.org/10.1017/9781108679930.

[2] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer. https://doi.org/10.1007/978-0-387-21572-9.

[3] Murphy, K. P. (2022). *Probabilistic machine learning: An introduction*. MIT Press. https://doi.org/10. 7551/mitpress/ 12290. 001.0001.