Jinchao Xu

Deep Learning Algorithms and Analysis

Summer 2020

Contents

1

Nonlinear Models

1.1 Nonlinear classifiable sets

In the section, we will extend the linearly separable sets to nonlinear case. A natural extension is like what kernel method does in SVM for binary case, we will introduce the so-called feature mapping.

Thus, we have the following natural extension for linearly separable by using feature mapping and original definition of linearly separable.

Definition 1 (nonlinearly separable sets). *These data sets $A_1, A_2, \cdots, A_k \subset \mathbb{R}^d$ are called nonlinearly separable, if there exist a feature space $\mathbb{R}^{\tilde{d}}$ and a smooth (if it has derivatives of all orders) feature mapping*

$$\varphi : \mathbb{R}^d \mapsto \mathbb{R}^{\tilde{d}} \tag{1.1}$$

such that

$$\tilde{A}_i := \varphi(A_i) = \{\tilde{x} \mid \tilde{x} = \varphi(x), x \in A_i\}, \quad i = 1, 2, \ldots, k, \tag{1.2}$$

are linearly separable.

Remark 1. 1. This definition is also consistent with the definition of linearly separable as we can just take $\tilde{d} = d$ and $\varphi = \mathrm{id}$ if $A_1, A_2, \cdots, A_k$ are already linearly separable.
2. The kernel method in SVM is mainly based on this idea for binary case (k=2) where they use kernel functions to approximate this $\varphi(x)$.
3. For most commonly used deep learning models, they are all associated with a softmax mapping which means that we can interpret these deep learning models as the approximation for feature mapping φ.

However, softmax is not so crucial for this definition actually as we have the next equivalent result.

Theorem 1. *$A_1, A_2, \cdots, A_k \subset \mathbb{R}^d$ are nonlinearly separable is equivalent that there there exist a smooth classification function*

$$\psi : \mathbb{R}^d \mapsto \mathbb{R}^k \tag{1.3}$$

such that for all $i = 1 : k$ and $j \neq i$

$$\psi_i(x) > \psi_j(x), \quad \forall x \in A_i. \tag{1.4}$$

Proof. On the one hand, it is easy to see that if $A_1, A_2, \cdots, A_k \subset \mathbb{R}^d$ are nonlinearly separable then they we can just take

$$\psi(x) = \boldsymbol{p}(\varphi(x); \theta), \tag{1.5}$$

where the $\boldsymbol{p}(y; \theta)$ is the softmax function for linearly separable sets $\varphi(A_i)$ for $i = 1, 2, \cdots, k$.

On the other hand, let assume that ψ is the smooth classification functions for $A_1, A_2, \cdots, A_k \subset \mathbb{R}^d$. We claim that, we can take $\varphi(x) = \psi(x)$ and then

$$\varphi(A_1), \varphi(A_2), \cdots, \varphi(A_k) \subset \mathbb{R}^k \quad (\tilde{d} = k), \tag{1.6}$$

will be linearly separable. Actually, if you take $\theta = (I, 0)$ in softmax mapping $\boldsymbol{p}(x; \theta)$, then the monotonicity of e^x show that for all $i = 1 : k$ and $j \neq i$

$$\boldsymbol{p}_i(\varphi(x); \theta) = \frac{e^{\psi_i(x)}}{\sum_{i=1}^k e^{\psi_i(x)}} > \frac{e^{\psi_j(x)}}{\sum_{i=1}^k e^{\psi_i(x)}} = \boldsymbol{p}_j(\varphi(x); \theta), \quad \forall x \in A_i. \tag{1.7}$$

□

Similarly to linearly separable sets, we have the next lemme for $k = 2$.

Lemma 1. *A_1 and A_2 are nonlinearly separable is equivalent that there exists a function $\varphi : \mathbb{R}^d \mapsto \mathbb{R}$ such that*

$$\varphi(x) > 0 \quad \forall x \in A_1 \quad \textit{and} \quad \varphi(x) < 0 \quad \forall x \in A_2. \tag{1.8}$$

Proof. Based the equivalence of nonlinearly separable sets, there exists $\psi_1(x)$ and $\psi_2(2)$ such that for all $i = 1 : 2$ and $j \neq i$

$$\psi_i(x) > \psi_j(x), \quad \forall x \in A_i. \tag{1.9}$$

Then, we can just take

$$\varphi(x) = \psi_1(x) - \psi_2(x). \tag{1.10}$$

On the other hand, if there exist $\varphi(x)$, then we can construct $\psi_1(x)$ and $\psi_2(2)$ as

$$\psi_1(x) = \frac{1}{2}\varphi(x) \quad \text{and} \quad \psi_2(x) = -\frac{2}{2}\varphi(x). \tag{1.11}$$

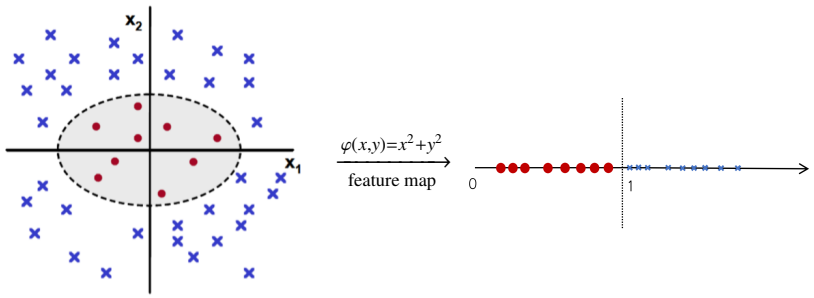
□

Remark 2. Here we mention that, we only assume that for all $i = 1 : k$ and $j \neq i$ we have $\psi_i(x) > \psi_j(x)$, $\forall x \in A_i$ for nonlinearly separable. We do not assume that $\psi_i(x) \geq 0$ or $\sum_{i=1}^k \psi_i(x) = 1$, which means that $\psi(x) = \begin{pmatrix} \psi_1(x) \\ \psi_2(x) \\ \vdots \\ \psi_k(x) \end{pmatrix}$ is not a discrete probability distribution over all k classes.

The previous theorem shows that softmax function is not so crucial in nonlinearly separable case. Combined with deep learning models, we have the following understanding about what deep learning models are approximating.

1. If a classification model is followed with a softmax, then the it is approximating the feature mapping $\varphi : \mathbb{R}^d \mapsto \mathbb{R}^{\tilde{d}}$.
2. If the classification model dose not followed by softmax, then it is approximating $\psi : \mathbb{R}^d \mapsto \mathbb{R}^k$ directly.

Example 1. Consider $k = 2$ and $A_1 \subset \{(x, y)|x^2 + y^2 < 1\}, \quad A_2 \subset \{(x, y)|x^2 + y^2 > 1\}$, then we can have the following nonlinear feature mapping:



Here we have the following comparison for linear and nonlinear models from the viewpoint of loss functions:

Linear case (Logistic regression):

$$L_\lambda(\theta) = \sum_{j=1}^{N} \ell(y_j, p(x_j; \theta)) + \lambda R(\|\theta\|)$$

Nonlinear case:

$$L_\lambda(\theta) = \sum_{j=1}^{N} \ell(y_j, p(\varphi(x_j; \theta_1); \theta_2)) + \lambda R(\|\theta\|)$$

Remark 3. We have the following remarks.

1. $\ell(q, p) = \sum_{i=1}^k -q_i \log p_i \leftrightarrow$ cross-entropy

2. $p(x;\theta) = \text{softmax}(Wx + b)$ where $\theta = (W, b)$
3. $\theta = (\theta_1, \theta_2)$ for nonlinear case
4. $\lambda R(\|\theta\|) \leftrightarrow$ regularization term

In general, we have the following popular nonlinear models for $\varphi(x;\theta)$

1. Polynomials.
2. Piecewise polynomials (finite element method).
3. Kernel functions in SVM.
4. Deep neural networks.

References

[1] L. Chen, P. Sun, and J. Xu. Optimal anisotropic meshes for minimizing interpolation errors in l^p-norm. *Mathematics of Computation*, 76(257):179–204, 2007.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision*, pages 630–645. Springer, 2016.

[4] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural networks*, 6(6):861–867, 1993.

[5] A. Pinkus. Approximation theory of the mlp model in neural networks. *Acta numerica*, 8:143–195, 1999.

[6] L. R. Scott and S. Zhang. Finite element interpolation of nonsmooth functions satisfying boundary conditions. *Mathematics of Computation*, 54(190):483–493, 1990.

[7] J. Xu. Finite element methods. Lecture notes, 2020.