DEEP AUTOENCODERS AS A KERNEL METHOD

Aidan Rocke
aidanrocke@gmail.com

September 14, 2021

1 The general intuition for kernel methods

Cover's theorem implicitly states that given a non-linearly separable dataset we may transform it into a linearly separable dataset by projecting it into a higher-dimensional space via a nonlinear transformation. How might we make use of this theorem using kernel methods?

If we have a mapping $\phi : \mathbb{R}^n \to \mathbb{R}^m$ that maps vectors in $\mathbb{R}^n$ to a Hilbert space with states in $\mathbb{R}^m$ then the dot product of $x, y \in \mathbb{R}^n$ is $\phi(x)^T \cdot \phi(y)$. A kernel is a function K that corresponds to the dot product:

$$K(x, y) = \phi(x)^T \cdot \phi(y) \tag{1}$$

This is useful because kernels give us a way to compute dot products in the feature space without explicit knowledge of the nonlinear map ϕ. By allowing us to compute dot product, we increase the probability that the data is linearly separable as predicted by Cover's theorem.

2 Deep Autoencoders as data-driven kernel methods

The objective of the deep autoencoder is to map the input space X to a higher-dimensional Hilbert space Y by identifying eigenfunctions that capture the intrinsic geometry of $X \subset \mathbb{R}^n$. The challenge of identifying eigenfunctions is simplified using the Universal Approximation property of fully-connected networks as their latent space naturally defines an orthogonal function space.

Given a relu network:

$$F_\theta : X \to Y \tag{2}$$

F_θ is a linear combination of functions ϕ_i that are affine on $X_i \subset X$ and zero on $X \setminus X_i$. Since ϕ_i may be modelled with active nodes in F_θ and ϕ_i is continuous on X_i, X_i must be compact and X_i and X_j must be pair-wise disjoint so:

$$\forall i, j \neq i, X_i \cap X_{j \neq i} = \emptyset \tag{3}$$

$$\forall i, j \neq i, \langle \phi_i, \phi_{j \neq i} = \int_X \phi_i(x) \cdot \phi_j(x) dx = 0 \tag{4}$$

$$\exists A \in \mathbb{R}^{m \times n} B \in \mathbb{R}^n, \phi_i(x) = Ax + B \tag{5}$$

The orthogonality criterion (3) is actually a corollary of (2). In fact, let's suppose there exists $x \in X$ such that $x \in X_i$ and $x \in X_j$. Since ϕ_i and ϕ_j correspond to different activation patterns in F_θ there must be a node in F_θ with value $\alpha \in \mathbb{R}$ such that $\alpha \geq 0$ and $\alpha < 0$.

Thus, ϕ_i represent the eigenfunctions of a linear map approximated by F_θ.

References

References:

[1] Ian Goodfellow and Yoshua Bengio and Aaron Courville. Deep Learning. MIT Press. 2016.

[2] Andrey Kolmogorov S.V. Fomin. Elements of the Theory of Functions and Functional Analysis: Metric and normed spaces. Dover. 1954.

[3] Cover, T.M. "Geometrical and Statistical properties of systems of linear inequalities with applications in pattern recognition". IEEE. 1965.

[4] Hofmann, Thomas; Scholkopf, Bernhard; Smola, Alexander J. Kernel Methods in Machine Learning. 2008.

[5] Welling, Max; Kingma, Diederik P. An Introduction to Variational Autoencoders. 2019.