Interpreting Neural Networks through Mahalanobis Distance

Alan Oursland

Independent Researcher

October 2024

Abstract

This paper establishes a novel connection between neural network architectures and the Mahalanobis distance, a statistical measure that accounts for data covariance structure. We present a robust mathematical framework bridging neural networks with this statistical distance measure, demonstrating how Absolute Value (Abs) activations facilitate distance-based interpretations. This distance-based interpretation has the potential to enhance model robustness, improve generalization, and provide more intuitive explanations of neural network decisions.

1 Introduction

Neural networks have revolutionized machine learning, achieving remarkable success across diverse applications. Central to their efficacy is the use of activation functions, which introduce non-linearity and enable the modeling of complex relationships within data. While Rectified Linear Units (ReLU) have gained prominence due to their simplicity and effectiveness [Nair and Hinton, 2010], the exploration of alternative activation functions remains an open and valuable area of research [Ramachandran et al., 2018].

Neural network units are often viewed as linear separators that define decision boundaries between classes [Minsky and Papert, 1969], with larger activation values suggesting stronger contributions of features to those decisions. Our work challenges this perspective, exploring how individual neurons can be understood through the lens of statistical distance measures. Clustering techniques aim to minimize the distance between data points and feature prototypes, with smaller values indicating stronger membership to the feature or cluster [MacQueen, 1967a]. Our work explores the intersection between these perspectives, leveraging the distance-minimization approach of clustering techniques to lay the groundwork for novel neural network designs based on statistical distance measures.

This paper establishes a novel connection between neural network architectures and the Mahalanobis distance, a statistical measure that accounts for the covariance structure of data [Mahalanobis, 1936]. We present robust mathematical framework that bridges neural networks to this statistical distance measure and lay the groundwork for future research into neural network interpretability and design [Lipton, 2016]. This distance-based interpretation has the potential to enhance model robustness, improve generalization, and offer more intuitive explanations of neural network decisions.

Our key contributions are:

1. We establish a mathematical connection between neural network linear layers and the Mahalanobis distance, demonstrating how Absolute Value (Abs) activations facilitate distance-based interpretations.

2. We analyze the implications of activation function choices on neural network behavior, comparing Abs and ReLU activations in terms of feature learning and training dynamics.

3. We discuss the broader implications of this framework for neural network design and interpretability, laying the groundwork for more interpretable and stable models.

2 Background and Related Work

2.1 Activation Functions

Activation functions introduce non-linearity in neural networks, enabling them to model complex data relationships. The field has evolved from early sigmoid and hyperbolic tangent functions [Rosenblatt, 1958] to the widely adopted Rectified Linear Unit (ReLU) [Nair and Hinton, 2010], which mitigates the vanishing gradient problem in deep networks [Glorot and Bengio, 2010, Krizhevsky et al., 2012].

ReLU variants intended to address its shortcomings include Leaky ReLU maas2013rectifier; Parametric ReLU (PReLU) [He et al., 2015]; and Exponential Linear Unit (ELU) [Clevert et al., 2016]. Additionally, newer activation functions like Swish [Ramachandran et al., 2018] and GELU [Hendrycks and Gimpel, 2016] have been proposed to further enhance network performance and training dynamics.

Tanh and Sigmoid activations are still used in many architectures such as recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) networks [Hochreiter and Schmidhuber, 1997].

The variety of activation functions used in modern networks reflects the diverse needs of different architectures. The exploration of activation functions remains an active area of research, with ongoing investigations into their impact on neural network performance, generalization, and interpretability [Ramachandran et al., 2018]. Despite extensive research, the interpretation of activation functions in terms of statistical measures remains an open area of investigation.

2.2 Overview of Distance Metrics in Clustering

Distance metrics are fundamental in clustering algorithms, determining how similarity between data points is measured. Various clustering methods employ different distance measures:

- K-Means typically uses Euclidean distance (ℓ_2 norm), assuming spherical clusters and equal feature importance [MacQueen, 1967b].
- Gaussian Mixture Models (GMMs) employ Mahalanobis distance, accounting for data covariance and modeling elliptical clusters [Reynolds, 2009].
- Hierarchical and Agglomerative Clustering can use various metrics (Euclidean, Manhattan, correlation-based), affecting dendrogram shape [Murtagh, 1983].
- DBSCAN, while often using Euclidean distance, can employ any metric for density-based clustering [Ester et al., 1996].
- Spectral Clustering incorporates similarity measures like Gaussian kernel functions [Von Luxburg, 2007].

The Mahalanobis distance stands out for its ability to account for feature correlations and scale differences, making it particularly useful in multivariate analysis [Mahalanobis, 1936, De Maesschalck et al., 2000]. It provides a scale-invariant measure that adjusts for the covariance structure of the data, offering advantages in high-dimensional spaces.

Understanding these distance metrics and their properties is crucial for selecting appropriate clustering algorithms and interpreting their results. As we explore the connection between neural networks and distance-based interpretations, these insights from clustering algorithms provide valuable context and inspiration.

2.3 Neural Network Interpretability and Statistical Models

The interpretability of neural networks remains a critical challenge, often referred to as the "black-box" nature of these models [Lipton, 2016]. In applications requiring transparency, such as healthcare and finance, understanding the decision-making processes of neural networks is paramount [Rudin, 2019]. Various approaches have been developed to enhance interpretability, including feature visualization, saliency maps, and prototype-based methods [Erhan et al., 2009, Simonyan and Zisserman, 2013, Kim et al., 2018].

Recent research has aimed to bridge neural networks with statistical and probabilistic models to provide a more principled understanding of their internal mechanisms [Bengio et al., 2013, Goodfellow et al., 2016]. Bayesian neural networks, for instance, incorporate uncertainty estimates that align network outputs with probabilistic interpretations [Neal, 1996, Blundell et al., 2015]. Additionally, connections between neural networks and kernel methods have been explored, highlighting how deep architectures can implicitly perform kernel-based feature transformations [Rahimi and Recht, 2008].

However, there is a notable gap in establishing direct mathematical connections between neural network components, specifically activation functions, and statistical distance measures like the Mahalanobis distance. Addressing this gap can provide deeper insights into feature learning and decision-making processes, thereby enhancing the interpretability and robustness of neural network models.

3 Mathematical Framework

In this section, we develop the mathematical foundation that connects neural networks to the Mahalanobis distance, thereby providing a framework for interpreting neural network operations through the lens of statistical distance metrics. We begin by revisiting key concepts related to Gaussian distributions and the Mahalanobis distance, followed by a detailed exploration of how neural network components, particularly linear layers and activation functions, can approximate these distance metrics. This framework not only enhances our understanding of neural network behavior but also lays the groundwork for leveraging statistical principles to improve network interpretability and training dynamics.

3.1 Mahalanobis Distance for a Multivariate Gaussian Distribution

A multivariate Gaussian (Normal) distribution is a fundamental concept in statistics, describing a d-dimensional random vector $\mathbf{x} \in \mathbb{R}^d$ with a mean vector $\boldsymbol{\mu} \in \mathbb{R}^d$ and a covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{d \times d}$ [Bishop, 2006]. We denote this distribution as $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

The Mahalanobis distance quantifies the distance between a point $\mathbf{x}$ and the mean $\boldsymbol{\mu}$ of a distribution, while considering the covariance structure of the data [Mahalanobis, 1936, De Maesschalck et al., 2000]. It is defined as:

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})}. \tag{1}$$

This metric adjusts for variance across dimensions by effectively whitening the data, resulting in a spherical distance measure.

3.2 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms data into a new coordinate system, emphasizing directions (principal components) that capture the most variance [Jolliffe, 2002]. When performing PCA on the covariance matrix $\mathbf{\Sigma}$, it is decomposed using eigenvalue decomposition:

$$\mathbf{\Sigma} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top, \tag{2}$$

where:

- $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_d]$ is a matrix whose columns are the orthogonal unit eigenvectors of $\mathbf{\Sigma}$.
- $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$ is a diagonal matrix of the corresponding eigenvalues λ_i, representing the variance along each principal component.

Substituting $\mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$ for $\mathbf{\Sigma}$ in the Mahalanobis distance equation (1), we obtain:

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^\top(\mathbf{x} - \boldsymbol{\mu})}. \tag{3}$$

To further simplify, we can express the Mahalanobis distance in terms of the principal components:

$$\begin{aligned} D_M(\mathbf{x}) &= \sqrt{(\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{V}\mathbf{\Lambda}^{-1}\mathbf{V}^\top(\mathbf{x} - \boldsymbol{\mu})} \\ &= \sqrt{(\mathbf{V}^\top(\mathbf{x} - \boldsymbol{\mu}))^\top \mathbf{\Lambda}^{-1}(\mathbf{V}^\top(\mathbf{x} - \boldsymbol{\mu}))} \\ &= \sqrt{\sum_{i=1}^{d} \lambda_i^{-1} \left(\mathbf{v}_i^\top(\mathbf{x} - \boldsymbol{\mu})\right)^2} \\ &= \left\| \lambda_i^{-1/2} \mathbf{v}_i^\top(\mathbf{x} - \boldsymbol{\mu}) \right\|_2 . \end{aligned} \tag{4}$$

where $\|\cdot\|_2$ denotes the Euclidean (ℓ_2) norm.

This shows that the Mahalanobis distance can also be expressed as the ℓ_2 norm of the number of standard deviations of $\mathbf{x}$ along each principal component.

3.3 Connecting Neural Networks to Mahalanobis Distance

We consider the Mahalanobis distance along a single principal component.

$$D_{M,i}(\mathbf{x}) = \left| \lambda_i^{-1/2} \mathbf{v}_i^\top(\mathbf{x} - \boldsymbol{\mu}) \right|, \tag{5}$$

This equation projecting the centered data $(\mathbf{x} - \boldsymbol{\mu})$ onto the direction of variance defined by the principal component eigenvector and scales by the inverse square root of the eigenvalue (variance).

Let

$$\mathbf{W} = \lambda_i^{-1/2} \mathbf{v}_i^\top, \tag{6}$$

$$\mathbf{b} = -\lambda_i^{-1/2} \mathbf{v}_i^\top \boldsymbol{\mu}. \tag{7}$$

We can simplify Equation (5) to

$$D_{M,i}(\mathbf{x}) = |\mathbf{W}\mathbf{x} - \mathbf{b}|, \tag{8}$$

This is identical to the equation for a linear layer where $\boldsymbol{W}$ represents the weight matrix, $\boldsymbol{b}$ the bias vector, and the Abs function serves as the activation function.

Linear nodes with Abs activations can be interpreted as 1d Gaussians along a direction of variance with the decision boundary passing through the mean of the modelled cluster.

Extending this to all principal components, the Mahalanobis distance can be represented as a multiple linear nodes, each corresponding to a principal component, followed by an Abs activation. This layered structure inherently accounts for the covariance structure of the data, effectively 'whitening' the input.

3.4 Non-Uniqueness of Whitening

Calculating the Mahalanobis distance along each principal component results in a whitened data set.

Whitening data is defined as follows:

$$\mathbf{x}_w = \boldsymbol{\Lambda}^{-1/2}\mathbf{V}^\top(\mathbf{x} - \boldsymbol{\mu}),$$

where $\mathbf{x}_w$ follows a distribution $\mathcal{N}(0, I)$.

The Mahalanobis distance can then be expressed as:

$$D_M(\mathbf{x}_w) = \|\mathbf{x}_w\|_2.$$

However the vectors to generate whitened data are not unique.

We can select a rotation matrix $\mathbf{R} \in SO(n)$ such that:

$$\begin{aligned} D_M(\mathbf{x}_w) &= D_M(\mathbf{R}\mathbf{x}_w) && (9)\\ &= D_M\left((\mathbf{R}\boldsymbol{\Lambda}^{-1/2}\mathbf{V}^\top)(\mathbf{x} - \boldsymbol{\mu})\right) && (10)\\ &= D_M\left(\boldsymbol{\Lambda}^{-1/2}(\mathbf{V}^\top\mathbf{R})(\mathbf{x} - \boldsymbol{\mu})\right). && (11) \end{aligned}$$

The Mahalanobis distance for whitened data remains invariant under the application of rotation matrices. If we transform the rotated basis back to the original space, we end up with a whitening basis whose component vectors may not be orthogonal.

This implies that, although linear nodes can represent principal components, they are unlikely to precisely learn the principal components when estimating Mahalanobis distances. Instead, they are expected to learn a basis that effectively whitens the data. However, the decision boundary of the learned hyperplane should still pass through the data mean.

3.5 Approximating Mahalanobis Distance with Neural Networks

In the neural network architecture, each Abs-activated linear node computes the deviation along a specific whitening component. The next linear node sums these values, thereby approximating the overall Mahalanobis distance through an ℓ_1 norm.

$$D_{M,\ell_1}(x) = \sum |\mathbf{W_i}\mathbf{x} - \mathbf{b_i}|, \tag{12}$$

This observation is motivated by a desire to analyze and maintain the behavior of well-studied standard neural network architectures. It offers several additional benefits over ℓ_2 including computational efficiency and robustness to outliers and sparse data, especially in high dimensional spaces Boyd and Vandenberghe [2004], Horn and Johnson [2012].

ℓ_1 norms have been utilized as an effective alternative to ℓ_2 norms in various high-dimensional applications [Bernhardsson, 2018, Tibshirani, 1996]. It is well-known that the ℓ_1 and ℓ_2 norms are related by the inequality:

$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2,$$

where $x \in \mathbb{R}^n$.

In high-dimensional spaces, the ℓ_1 norm is a reasonable approximation of the ℓ_2 norm Vershynin [2018].

4 Implications and Discussion

We discuss implications, potential impact and future work of this reframing of linear layers in neural networks. While this paper provides a robust theoretical foundation for interpreting neural networks through Mahalanobis distance and Abs activation functions, it does not include empirical results. Future work will involve validating these theoretical insights with empirical data to further assess their applicability and performance in real-world scenarios.

4.1 Expected Value Interpretation

The concept of expected value or mean is fundamental in statistics and machine learning, providing a central tendency that characterizes a distribution. In the context of neural networks, identifying an expected value for each neuron can offer insights into the features it recognizes and how it processes information.

By interpreting linear nodes as principal components of Gaussian approximations of data clusters, we can potentially identify a mean value for the data points recognized by each neuron. This probabilistic interpretation of neural network operations enhances our understanding of their internal representations.

As a linear separator, a neuron with inputs in $\mathbb{R}^n$ defines a hyperplane with $(n-1)$ dimensions and a normal vector in 1 dimension. The normal vector represents the Gaussian direction, with the mean being somewhere on the hyperplane. There are many techniques that could be successful at this including projecting data onto the normal vector and clustering, weighted averages using functions of the inverse distances, boundary constraints, optimization searches and manifold analysis.

This expected value or mean effectively acts as a prototype for the feature that the neuron has learned to recognize [Li et al., 2018]. It represents the 'ideal' or 'typical' input for that neuron, around which the neuron's response varies. This prototype interpretation can provide valuable insights into the feature extraction process of neural networks and may lead to more interpretable models and improved network architectures.

4.2 Equivalence between Abs and ReLU

This work stems from a desire to interpret neural network internals. While our analysis of standard MLP architecture leads to an ℓ_1 approximation of the Mahalanobis distance, and necessitates an Abs activation, we posit that ReLU can provide comparable information and may be interpretable within the same framework.

Given a confidence bound δ, we observe:

- For Abs activation: The preceding linear layer learns to output $\{-\delta, +\delta\}$, with the decision boundary intersecting the cluster mean.
- For ReLU activation: The preceding linear layer can position its decision boundary just outside the cluster, learning to output values between $\{0, 2\delta\}$.

Subsequent layers in the network can adapt to either output range. This suggests a functional equivalence between Abs and ReLU in this context.

Techniques developed to enhance learning and interpretation with Abs activation functions may be adaptable to the more commonly used ReLU, potentially bridging theoretical insights with practical neural architectures.

4.3 Activations as Distance Metrics

Traditional neural networks typically employ an "intensity metric model," where larger activation values indicate stronger feature presence. In contrast, a "distance metric model" interprets smaller activation values as indicating closer proximity to a learned feature or prototype [Broomhead and Lowe, 1988]. The following observations suggest directions for future work:

- Distance and intensity metrics can be interconverted through negation.
- Subsequent layer weights can apply their own negation, obscuring the metric type learned by internal nodes.
- Distance metrics are incompatible with sparse layer output.
- Most error functions (e.g., Cross Entropy Loss, Hinge Loss) are designed for intensity metrics. Output layers using Abs activation should be modified accordingly.
- The Gaussian connection suggests transforming distance metrics through exponential ($y = e^{-x^2}$) or Laplace ($y = e^{-|x|}$) functions. These may suffer from vanishing gradients. An alternative approximation could combine Abs and ReLU: $y = \text{ReLU}(-\text{Abs}(x) + \text{confidence\_bound})$.
- Some neural network architectures, such as Radial Basis Function networks [Broomhead and Lowe, 1988], have employed distance metrics, but they are not widely adopted.
- There may exist regularization techniques that encourage distance metric learning [Weinberger and Saul, 2009].

4.4 Model Initialization and Pretraining

The interpretation of neurons as learning distances from cluster means suggests novel approaches to model initialization and pretraining. This perspective offers a potential alternative to standard random initialization techniques [Kamilov et al., 2017].

Given randomly initialized weights $\mathbf{W}$, we propose setting the bias b such that the decision boundary passes through a data point or cluster centroid:

$$b = -\mathbf{W} \cdot \boldsymbol{\mu} \tag{13}$$

where $\boldsymbol{\mu}$ represents a chosen data point or cluster centroid.

This initialization strategy could offer several potential advantages:

- Faster convergence by starting with meaningful decision boundaries
- Improved interpretability of initial network states
- Potential for better generalization by incorporating data distribution information from the start

Future work could explore:

- Empirical comparisons with standard initialization techniques (e.g., He, Xavier)
- Extensions to deep networks, considering layerwise or global clustering approaches
- Combination with other pretraining methods, such as autoencoders or contrastive learning
- Theoretical analysis of the impact on gradient flow and optimization dynamics

4.5 Model Translation and Componentization

The interpretation of neurons as principal components of Gaussians suggests a potential homomorphism between neural networks and hierarchical Gaussian Mixture Models (GMMs) [Jacobs et al., 1991]. This perspective opens up several promising avenues for research and application.

It may be possible to directly convert between neural networks and GMMs. This translation could offer several benefits:

- Enhanced interpretability of neural networks through their GMM counterparts
- Ability to leverage well-established statistical techniques for GMMs in neural network analysis
- Potential for hybrid models that combine the strengths of both paradigms

The locality properties of Gaussians suggest a novel approach to managing large neural networks:

- Decomposition of large networks into smaller, context-specific subnetworks
- Offline storage of these subnetworks, with dynamic loading based on data context
- Potential for improved memory efficiency and faster inference in large-scale applications

This componentization approach could address challenges in deploying large models on resource-constrained devices or in latency-sensitive applications.

5 Conclusion

This paper establishes a novel connection between neural network architectures and the Mahalanobis distance, providing a fresh perspective on neural network interpretability. By demonstrating how linear layers with Abs activations can approximate Mahalanobis distances, we bridge the gap between statistical distance measures and neural network operations. This framework offers several key insights:

- It provides a probabilistic interpretation of neural network nodes as learning principal components of Gaussian distributions.
- It suggests new approaches for model initialization, pretraining, and componentization.
- It establishes a potential homomorphism between neural networks and hierarchical Gaussian Mixture Models.

These findings lay the groundwork for future research into more interpretable and robust neural network architectures. By leveraging statistical principles in neural network design, we open new avenues for enhancing model transparency, improving generalization, and developing more efficient training techniques. As the field of AI continues to evolve, such interpretable frameworks will be crucial in building trustworthy and explainable AI systems.

References

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8): 1798–1828, 2013.

Erik Bernhardsson. Annoy: Approximate nearest neighbors in c++/python optimized for memory usage and loading speed. *GitHub repository*, 2018. URL `https://github.com/spotify/annoy`.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 1613–1622. PMLR, 2015.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

David S Broomhead and David Lowe. Radial basis functions, multi-variable functional interpolation and adaptive networks. *Royal Signals and Radar Establishment Malvern (United Kingdom) RSRE Memorandum*, 4148, 1988.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In *4th International Conference on Learning Representations (ICLR)*, 2016. URL `https://arxiv.org/abs/1511.07289`.

R. De Maesschalck, D. Jouan-Rimbaud, and D. L. Massart. The mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50(1):1–18, 2000. doi: 10.1016/S0169-7439(99)00047-7.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. In *Technical Report, University of Montreal*, 2009.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231. AAAI Press, 1996.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 249–256, 2010.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. URL `https://arxiv.org/abs/1606.08415`.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9 (8):1735–1780, 1997.

Roger A Horn and Charles R Johnson. *Matrix Analysis*. Cambridge University Press, 2012.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Ian T. Jolliffe. *Principal Component Analysis*. Springer, 2 edition, 2002.

Ulugbek S. Kamilov, Hassan Mansour, and Brendt Wohlberg. A survey of computational imaging methods for inverse problems. *IEEE Signal Processing Magazine*, 34(6):85–95, 2017.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 2668–2677. PMLR, 2018.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105, 2012.

Oscar Li, Hao Liu, Chaofan Chen, and Cynthia Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Zachary C. Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016. URL `https://arxiv.org/abs/1606.03490`.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967a.

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297. University of California Press, 1967b.

Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2(1):49–55, 1936.

Marvin L. Minsky and Seymour A. Papert. *Perceptrons: An Introduction to Computational Geometry*. MIT Press, Cambridge, MA, USA, 1969. ISBN 9780262631832.

Fionn Murtagh. A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*, 26(4):354–359, 1983.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pages 807–814, 2010.

Radford M. Neal. *Bayesian Learning for Neural Networks*, volume 118 of *Lecture Notes in Statistics*. Springer-Verlag, 1996.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, pages 1177–1184, 2008.

Prajit Ramachandran, Barret Zoph, and Quoc V. Le. Searching for activation functions. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=Hk4_qw5K-`. arXiv preprint arXiv:1710.05941.

Douglas A. Reynolds. Gaussian mixture models. In *Encyclopedia of Biometrics*, pages 659–663. Springer, 2009.

Frank Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Karen Simonyan and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. URL `https://arxiv.org/abs/1312.6034`.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

Kilian Q. Weinberger and Lawrence K. Saul. Distance metric learning for large margin nearest neighbor classification. In *International Conference on Machine Learning (ICML)*, pages 1317–1324, 2009. URL `https://proceedings.mlr.press/v5/weinberger09a.html`.