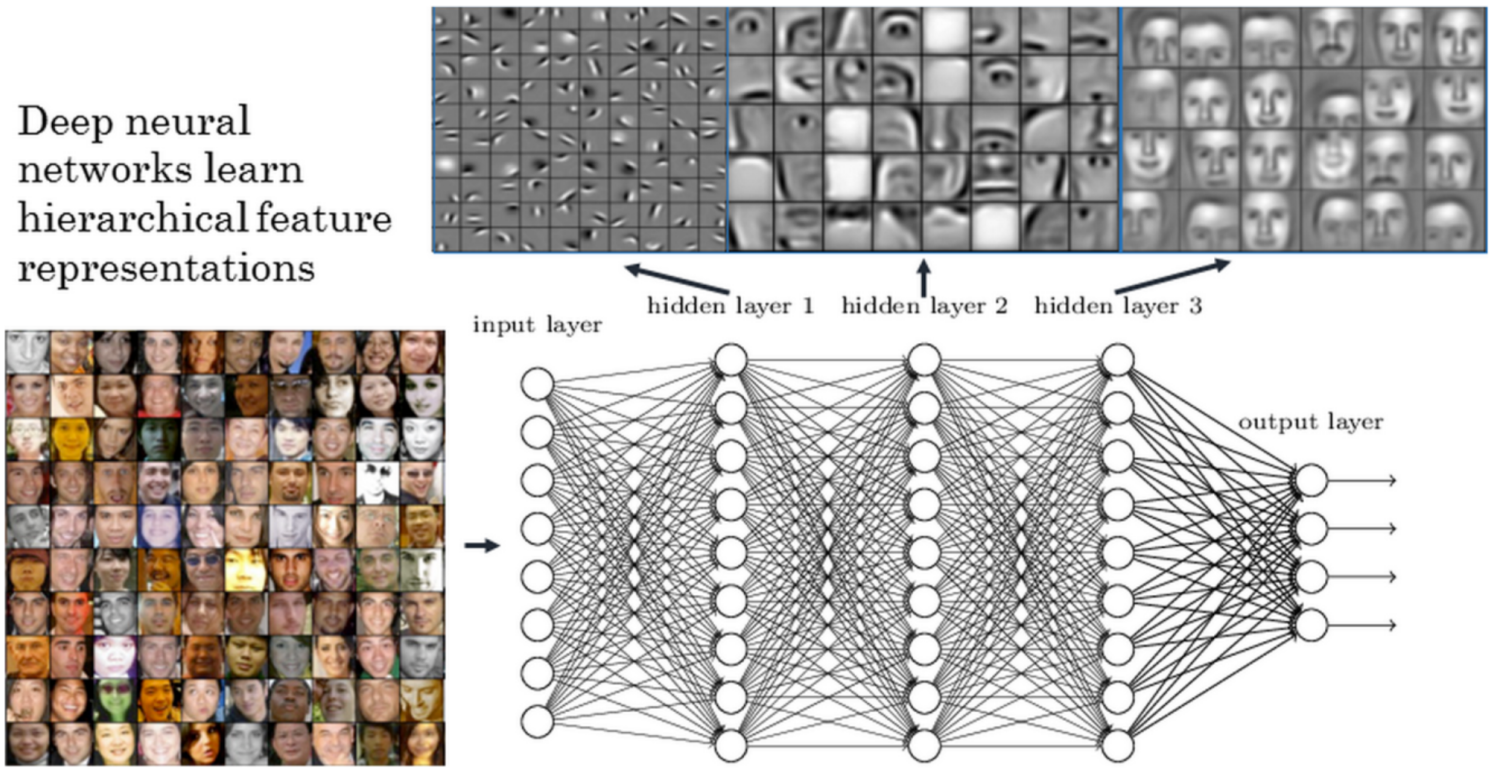
Universal Approximation Theorem, G. Cybenko

Elisa Negrini

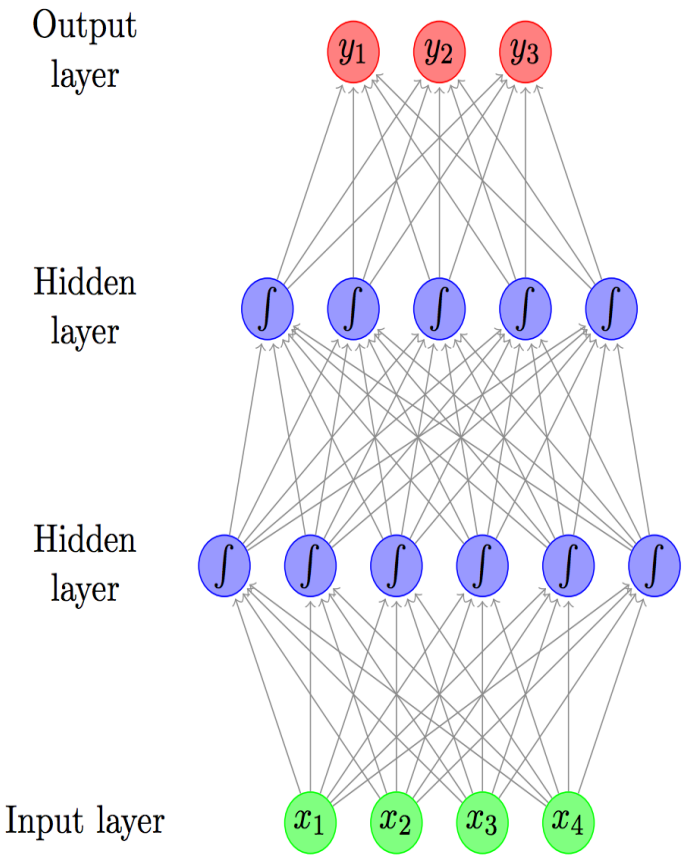
Worcester Polytechnic Institute

October 29 2019

From Randy's Talk



https://nivdul.wordpress.com/2015/11/17/exploring-deep-learning-with-li-zhe/



- A Neural Network is a parametrized family of functions $G(\theta; x) = y$.
- $G(\theta; x)$ is a multiple composition of affine functions and nonlinear functions called **activations**.

More Mathematically

- Let L be the number of layers in the network

More Mathematically

- Let L be the number of layers in the network
- Let the input $x \in \mathbb{R}^n$

More Mathematically

- Let L be the number of layers in the network
- Let the input $x \in \mathbb{R}^n$
- Define the *weight matrices* $W_i \in \mathbb{R}^{n_i \times n_{i-1}}$, $i = 1, \cdots, L$.

More Mathematically

- Let L be the number of layers in the network
- Let the input $x \in \mathbb{R}^n$
- Define the *weight matrices* $W_i \in \mathbb{R}^{n_i \times n_{i-1}},\ i = 1, \cdots, L$.
- Define the *bias vector* $b_i \in \mathbb{R}^{n_i},\ i = 1, \cdots, L$.

More Mathematically

- Let L be the number of layers in the network
- Let the input $x \in \mathbb{R}^n$
- Define the *weight matrices* $W_i \in \mathbb{R}^{n_i \times n_{i-1}},\ i = 1, \cdots, L$.
- Define the *bias vector* $b_i \in \mathbb{R}^{n_i},\ i = 1, \cdots, L$.
- Define the parameters $\theta_i = \{W_i, b_i\}$

More Mathematically

- Let L be the number of layers in the network
- Let the input $x \in \mathbb{R}^n$
- Define the *weight matrices* $W_i \in \mathbb{R}^{n_i \times n_{i-1}},\ i = 1, \cdots, L$.
- Define the *bias vector* $b_i \in \mathbb{R}^{n_i},\ i = 1, \cdots, L$.
- Define the parameters $\theta_i = \{W_i, b_i\}$
- Let $\sigma : \mathbb{R}^{n_i} \to \mathbb{R}^{n_i}$ be the *activation function*

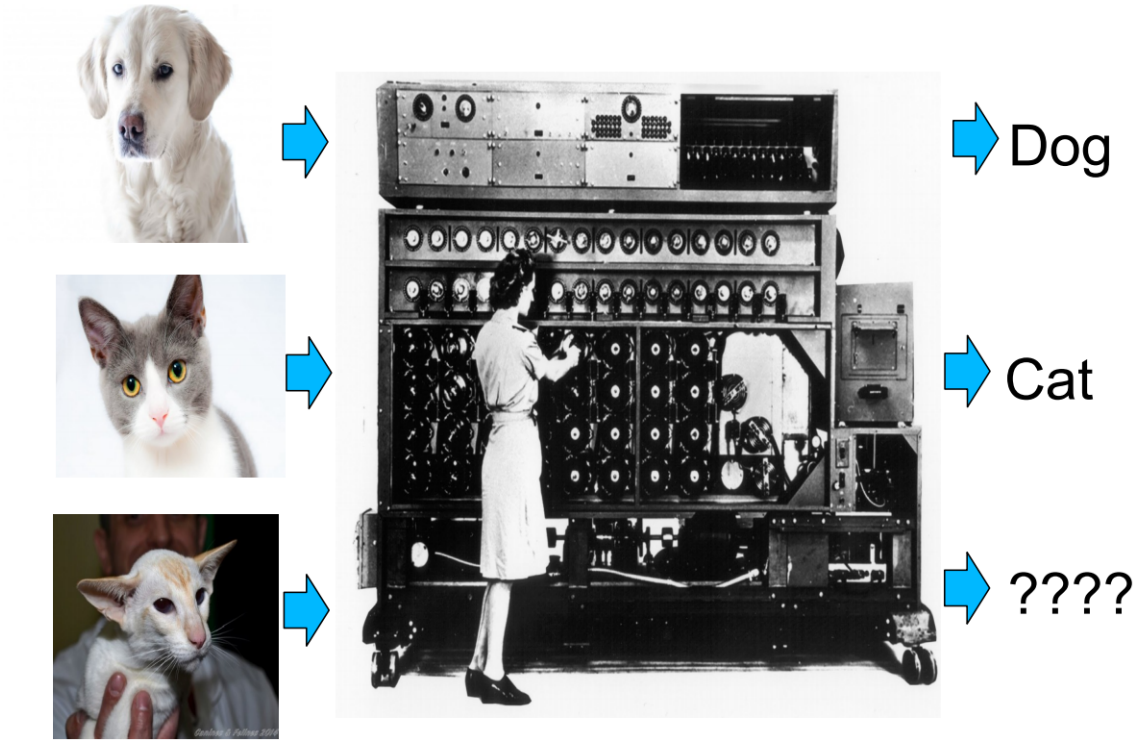
More Mathematically

- Let L be the number of layers in the network
- Let the input $x \in \mathbb{R}^n$
- Define the *weight matrices* $W_i \in \mathbb{R}^{n_i \times n_{i-1}},\ i = 1, \cdots, L$.
- Define the *bias vector* $b_i \in \mathbb{R}^{n_i},\ i = 1, \cdots, L$.
- Define the parameters $\theta_i = \{W_i, b_i\}$
- Let $\sigma : \mathbb{R}^{n_i} \to \mathbb{R}^{n_i}$ be the *activation function*

- Then for $L = 2$

$$y = G(\theta; x) = g_2(\theta_2; g_1(\theta_1; x)) = \sigma(W_2\sigma(W_1x + b_1) + b_2)$$

The Question



https://users.wpi.edu/ msarkis/BrownBag/Randy2.pdf

What functions can $G(\theta; x)$ approximate and how well?

Let's Keep It Simple: One Layer Networks

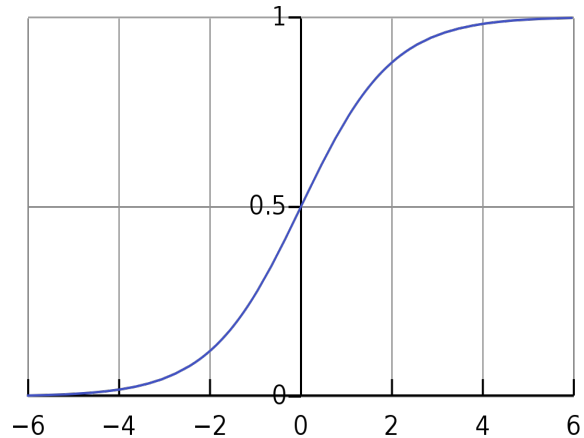
The results presented in Cybenko's paper are for a one layer network with scalar output.

More precisely, if N is the number of neurons in the layer,

$$G(\theta; x) = \sum_{j=1}^{N} \alpha_j \sigma(w_j^T x + b_j), \text{ where } w_j \in \mathbb{R}^n, \ \alpha_j, \ b_j \in \mathbb{R}$$

Moreover we will use **sigmoidal** activation functions, that is:

$$\sigma(t) \to \begin{cases} 1 & \text{as } t \to +\infty \\ 0 & \text{as } t \to -\infty \end{cases}$$



The Result

Theorem (Cybenko)

If σ is any continuous sigmoidal function, then the family of functions parametrized by a one layer neural network is dense, with respect to the supremum norm, in the space of continuous functions on the unit cube.

Notation

- $I_n = [0,1]^n$ is the n-dimensional unit cube.
- $C(I_n)$ space of continuous functions on I_n, with the supremum norm $\|\cdot\|$
- $M(I_n)$ space of finite, signed regular Borel measures on I_n.

Discriminatory Functions

Definition

We say that a function σ is *discriminatory* if given a measure $\mu \in M(I_n)$ such that

$$\int_{I_n} \sigma(w^T x + b) d\mu(x) = 0, \ \forall w \in \mathbb{R}^n, \ b \in \mathbb{R}$$

implies that $\mu = 0$

What functions are NOT discriminatory?

Let $\sigma = \frac{1}{2}$, and μ be a measure with density $g = -\chi_{[0,\frac{1}{2})} + \chi_{(\frac{1}{2},1]}$.

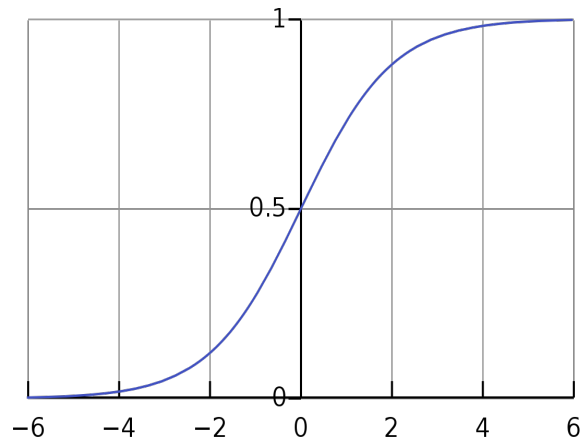
Clearly $\mu \neq 0$, but $\int_{[0,1]} \sigma(w^T x + b) d\mu(x) = 0, \ \forall w, \ b$

What Functions are Discriminatory?

Lemma

Any bounded, measurable, sigmoidal function is discriminatory.
In particular, any continuous sigmoidal function is discriminatory.

Example: A sigmoid $\sigma = \frac{1}{1+e^{-x}}$ is discriminatory.



Universal Approximation Theorem

Theorem (Cybenko)

Let σ be any continuous discriminatory function.
Then finite sums of the form

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(w_j^T x + b_j), \text{ where } w_j \in \mathbb{R}^n, \ \alpha_j, \ b_j \in \mathbb{R}$$

are dense in $C(I_n)$.

In other words, given any $\varepsilon > 0$ and $f \in C(I_n)$, there is a sum $G(x)$ of the above form such that

$$|G(x) - f(x)| < \varepsilon, \quad \forall x \in I_n$$

Tools for the Proof

Theorem (Hahn-Banach)

Let V be a normed vector space, $R \subset V$ a subspace of V.
Let $L \in R^*$. Then there exists $\hat{L} \in V^*$ that extends L to V and satisfies $\|\hat{L}\|_{V^*} = \|L\|_{R^*}$

Corollary

Let V be a normed vector space, $R \subset V$ a subspace of V.
Let $x_0 \in V$ such that $d(x_0, R) = \gamma > 0$.
Then there exists $L \in V^*$ such that

- $\|L\|_{V^*} = 1$
- $L(x_0) = \gamma$
- $L(R) = 0$

Tools for the Proof

Theorem (Riesz Representation Theorem)

Let L be a bounded linear functional on $C(I_n)$.
Then there exists a unique $\mu \in M(I_n)$ such that

$$L(h) = \int_{I_n} h(x) d\mu(x) \quad \forall h \in C(I_n)$$

Proof of Universal Approximation Theorem

Goal: Let $S \subset C(I_n)$ be the set of functions of the form

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(w_j^T x + b_j)$$

We want to to prove that $R := \bar{S} = C(I_n)$.

- S is a linear subspace of $C(I_n)$.

Proof of Universal Approximation Theorem

Goal: Let $S \subset C(I_n)$ be the set of functions of the form

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(w_j^T x + b_j)$$

We want to to prove that $R := \bar{S} = C(I_n)$.

- S is a linear subspace of $C(I_n)$.
- **By contradiction** suppose $R \subsetneq C(I_n)$, that is $\exists f \in C(I_n)$ such that $d(f, R) > 0$.

Proof of Universal Approximation Theorem

Goal: Let $S \subset C(I_n)$ be the set of functions of the form

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(w_j^T x + b_j)$$

We want to to prove that $R := \bar{S} = C(I_n)$.

- S is a linear subspace of $C(I_n)$.
- **By contradiction** suppose $R \subsetneq C(I_n)$, that is $\exists f \in C(I_n)$ such that $d(f, R) > 0$.
- By the Corollary to H-B $\exists L$ bounded linear functional on $C(I_n)$ such that $L \neq 0$, but $L(S) = L(R) = 0$

Proof of Universal Approximation Theorem

- By RRT $\exists!\, \mu \in M(I_n)$ such that

$$L(h) = \int_{I_n} h(x) d\mu(x), \quad \forall h \in C(I_n)$$

Proof of Universal Approximation Theorem

- By RRT $\exists!\, \mu \in M(I_n)$ such that

$$L(h) = \int_{I_n} h(x) d\mu(x), \quad \forall h \in C(I_n)$$

- Since $L(R) = 0$ and since $\sigma(w^T x + b) \in R,\ \forall w, b$, then

$$0 = L(\sigma(w^T x + b)) = \int_{I_n} \sigma(w^T x + b) d\mu(x), \quad \forall w, b \quad (1)$$

Proof of Universal Approximation Theorem

- By RRT $\exists!\, \mu \in M(I_n)$ such that

$$L(h) = \int_{I_n} h(x) d\mu(x), \quad \forall h \in C(I_n)$$

- Since $L(R) = 0$ and since $\sigma(w^T x + b) \in R,\ \forall w, b$, then

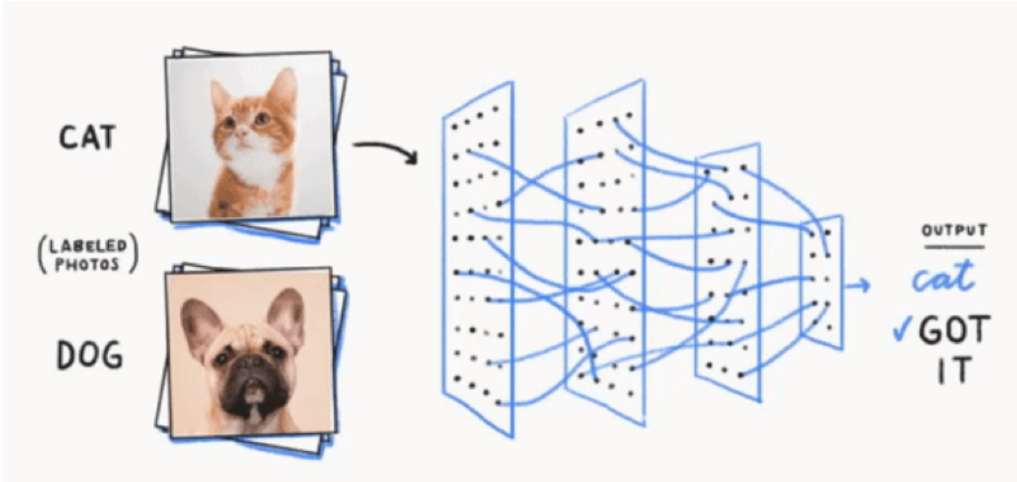
$$0 = L(\sigma(w^T x + b)) = \int_{I_n} \sigma(w^T x + b) d\mu(x), \quad \forall w, b \quad (1)$$

- Since σ is discriminatory, (1) implies $\mu = 0$, which in turn implies $L = 0$, and this is a **contradiction**.

Remarks

- This works in particular for sigmoids which are often used as activation functions.
- We showed that networks with one internal layer and an arbitrary continuous sigmoidal function can approximate continuous functions with arbitrary precision, **providing that no constraints are placed on the number of nodes or size of the weights**

Approximation of Classification Functions



https://becominghuman.ai

Let m be Lebesgue measure on I_n. Let $P_1, \cdots, P_k$ be a partition of I_n into k disjoint subsets. Define the *classification function* f as

$$f(x) = j \iff x \in P_j$$

Can f be approximated by a one layer network?

Approximation of Classification Functions

Theorem

Let σ be a continuous sigmoidal function. Let f be a classification function for any finite measurable partition of I_n.
Then for any $\varepsilon > 0$ there exists a finite sum of the form

$$G(x) = \sum_{j=1}^{N} \alpha_j \sigma(w_j^T x + b_j) \tag{2}$$

and a set $D \subset I_n$ with $m(D) \geq 1 - \varepsilon$ such that

$$|G(x) - f(x)| < \varepsilon, \quad \forall x \in D$$

In other words, there is a network that makes the measure of the set of points incorrectly classified as small as desired!

Tools for the Proof

Theorem (Lusin)

Let $f : I_n \to \mathbb{R}$ be measurable.
Then for any $\varepsilon > 0$ there exists a set $D \subset I_n$ with $m(D) \geq 1 - \varepsilon$ and $h \in C(I_n)$ such that

$$h(x) = f(x), \quad \forall x \in D$$

Proof of the Theorem

- By Lusin's theorem $\exists D \subset I_n$ with $m(D) \geq 1 - \varepsilon$ and $h \in C(I_n)$ such that $h(x) = f(x)$, $\forall x \in D$.

Proof of the Theorem

- By Lusin's theorem $\exists D \subset I_n$ with $m(D) \geq 1 - \varepsilon$ and $h \in C(I_n)$ such that $h(x) = f(x)$, $\forall x \in D$.

- Since $h \in C(I_n)$ by the Universal Approximation Theorem, there exists a network G of the form (2) such that

$$|G(x) - h(x)| < \varepsilon \quad \forall x \in I_n$$

Proof of the Theorem

- By Lusin's theorem $\exists D \subset I_n$ with $m(D) \geq 1 - \varepsilon$ and $h \in C(I_n)$ such that $h(x) = f(x)$, $\forall x \in D$.
- Since $h \in C(I_n)$ by the Universal Approximation Theorem, there exists a network G of the form (2) such that

$$|G(x) - h(x)| < \varepsilon \quad \forall x \in I_n$$

- Then for all $x \in D$ we have

$$|G(x) - f(x)| = |G(x) - h(x)| < \varepsilon$$

Conclusion

- We proved that any continuous function on the unit cube can be uniformly approximated by a one layer network with an arbitrary continuous sigmoidal nonlinearity.

Conclusion

- We proved that any continuous function on the unit cube can be uniformly approximated by a one layer network with an arbitrary continuous sigmoidal nonlinearity.

- We showed that classification functions can be arbitrarily well approximated by a one layer network with continuous sigmoidal nonlinearity and that the set of the incorrectely labeled points has small measure.

Conclusion

- **However,** these are existence theorems.

Conclusion

- **However,** these are existence theorems.
- How many terms in the summation (or how many nodes in the network layer) are needed to yield an approximation of given quality?

Conclusion

- **However,** these are existence theorems.
- How many terms in the summation (or how many nodes in the network layer) are needed to yield an approximation of given quality?
- Some papers try to address this question:

Conclusion

- **However,** these are existence theorems.
- How many terms in the summation (or how many nodes in the network layer) are needed to yield an approximation of given quality?
- Some papers try to address this question:
  1. Baum, Eric B., and David Haussler. "What size net gives valid generalization?." Advances in neural information processing systems. 1989.

Conclusion

- **However,** these are existence theorems.
- How many terms in the summation (or how many nodes in the network layer) are needed to yield an approximation of given quality?
- Some papers try to address this question:
  1. Baum, Eric B., and David Haussler. "What size net gives valid generalization?." Advances in neural information processing systems. 1989.
  2. Makhoul, John, Richard Schwartz, and Amro El-Jaroudi. "Classification capabilities of two-layer neural nets." International Conference on Acoustics, Speech, and Signal Processing,. IEEE, 1989.

Conclusion

- **However,** these are existence theorems.
- How many terms in the summation (or how many nodes in the network layer) are needed to yield an approximation of given quality?
- Some papers try to address this question:
  1. Baum, Eric B., and David Haussler. "What size net gives valid generalization?." Advances in neural information processing systems. 1989.
  2. Makhoul, John, Richard Schwartz, and Amro El-Jaroudi. "Classification capabilities of two-layer neural nets." International Conference on Acoustics, Speech, and Signal Processing,. IEEE, 1989.
  3. Barron, Andrew R. "Approximation and estimation bounds for artificial neural networks." Machine learning 14.1 (1994): 115-133.

References


Robert B. Ash.
Real analysis and probability.
Academic Press, New York, 1972.


George Cybenko.
Approximation by superpositions of a sigmoidal function.
Mathematics of control, signals and systems, 2(4):303–314, 1989.


Walter Rudin.
Real and complex analysis.
McGraw-Hill, New York, 1966.


Walter Rudin.
Functional analysis.
McGraw-Hill, New York, 1973.


Thank You