06-20416 and 06-12412
(Intro to) Neural Computation

09 – Universal Approximation

Per Kristian Lehre

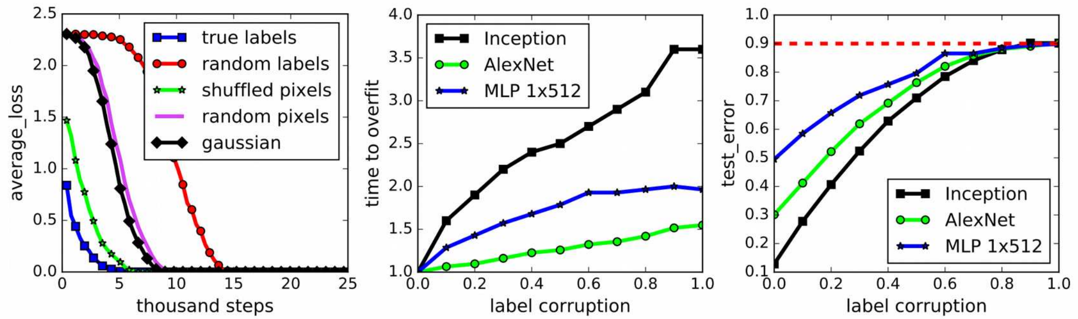
Last lectures

- Variants of stochastic gradient descent
  - Momentum, Nesterov Momentum
  - AdaGrad
  - Adam
- The biological brain (guest lecture by Hidalgo)

Generalisation in Neural Networks

- Hypothesis:
  - Neural networks generalise from the training data, i.e., by learning inherent “structure” in the data.
  - Test of hypothesis: Removing structure should reduce the network performance.
- Zhang et al. (ICLR 2017) trained a neural network on the CIFAR10 dataset in these settings:
  - True labels: original training set
  - Random labels: all the labels are replaced with random ones
  - Shuffled pixels: a random permutation of pixels
  - Gaussian: A Gaussian distribution is used to generate random pixels for each image

Results



(a) learning curves (b) convergence slowdown (c) generalization error growth

Figure 1: Fitting random labels and random pixels on CIFAR10. (a) shows the training loss of various experiment settings decaying with the training steps. (b) shows the relative convergence time with different label corruption ratio. (c) shows the test error (also the generalization error since training error is 0) under different label corruptions.

- Deep neural networks easily fit random labels.
- The effective capacity of neural networks is sufficient for memorizing the entire data set.
- Training time increases only by a small constant factor.

Outline

- Rethinking generalisation in deep learning
- Approximation capability
  - Perceptrons (Rosenblatt, 1957)
  - Perceptrons and the XOR function (Minsky & Papert, 1969)
  - Universal Approximation theorem (Cybenko, 1989)
  - Power of depth (Eldan & Shamir, 2016)

Perceptron (Rosenblatt, 1957)

Precursor to today's neural networks

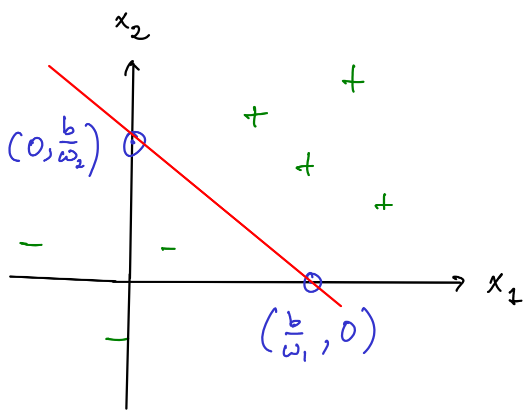
$$f(x) = \text{sign}(w^T x - b) \quad \text{where} \quad \text{sign}(x) = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{otherwise} \end{cases}$$

Linear Decision Boundary

The boundary between positive and negative output of a single-layer perceptron can be described by a linear hyperplane.

Example: In two dimensions, we have

$$f(x_1, x_2) = \text{sign}(w_1 x_1 + w_2 x_2 - b)$$



$$w_1 x_1 + w_2 x_2 - b > 0$$

$$\Rightarrow \quad x_2 > \frac{b - w_1 x_1}{w_2}$$

Minsky & Papert (1969)

A single layer perceptron cannot learn the XOR function.

Caused controversy, and contributed to "AI winter"

- reduced research funding to neural network research
- reduced interest among researchers

Lebesgue-integration

The Lebesgue-integral

$$\int f(x)\, d\mu(x)$$

an alternative to the Riemann integral which is defined for more complex functions f.
It is defined with respect to a measure μ which measures the "size" of subsets of the domain of f.

"Simplified" definition using the Riemann integral

Given $f : X \to \mathbb{R}$, let

$$f^*(t) = \mu\left(\{ x \mid f(x) > t \} \right)$$

then

$$\underbrace{\int f(x)\, d\mu(x)}_{\text{Lebesgue-integral.}} = \underbrace{\int_0^\infty f^*(t)\, dt}_{\text{Riemann integral}}$$

Discriminatory Functions

Def $\sigma : \mathbb{R} \to \mathbb{R}$ is called discriminatory if

$$\int_{I_n} \sigma(y^T x + \theta)\, d\mu(x) = 0$$

for all $y \in \mathbb{R}^n$ and $\theta \in \mathbb{R}$ implies $\mu = 0$.

Lemma

Any function $\sigma : \mathbb{R} \to \mathbb{R}$ where

$$\sigma(x) \to \begin{cases} 1 & \text{for} \quad x \to \infty \\ 0 & \text{for} \quad x \to -\infty \end{cases}$$

is discriminatory.

Example: The sigmoid function $\sigma(x) = \dfrac{1}{1 + e^{-x}}$ is discriminatory.

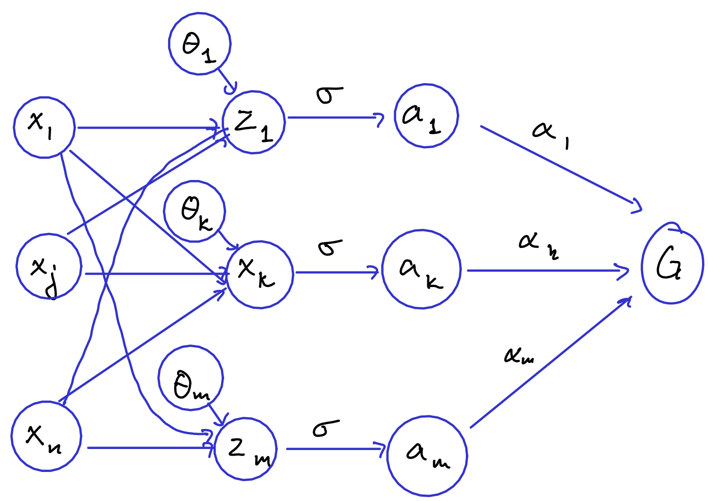
Theorem (Cybenko, 1989)

Let σ be any continuous discriminatory function, then for any $f \in C(I_n)$ (i.e., continuous function on $I_n = [0,1]^n$), and any $\varepsilon > 0$, there exists a finite sum on the form

$$G(x) = \sum_{j=1}^{m} \alpha_j \sigma\left(w_j^T x + \theta_j\right)$$

such that

$$|G(x) - f(x)| < \varepsilon \quad \text{for all } x \in I_n.$$



Definition: Normed Linear Space

is a vector space X over $\mathbb{R}$
and a function $\|\cdot\| : X \to \mathbb{R}$ satisfying

(i) $\|x\| \geq 0$ for all $x \in X$
(ii) $\|x\| = 0$ if and only if $x = 0$
(iii) $\|\alpha x\| = |\alpha| \cdot \|x\|$ for all $\alpha \in \mathbb{R}$ and $x \in X$
(iv) $\|x+y\| \leq \|x\| + \|y\|$ for all $x, y \in X$

Definition (supremum norm)

Define

$$\|f\| := \sup \{ |f(x)| \mid x \in X \}.$$

We can now measure the "distance" between two functions g and f by

$$\|f - g\|$$

Closure

Let Y be a subset of a normed vector space X.
The closure of Y, denoted $\overline{Y}$, consists of all $x \in X$ such that for each $\varepsilon > 0$, we can find an element $y \in Y$ such that

$$\|y - x\| < \varepsilon.$$

Example

The closure of the set of rational numbers $\mathbb{Q}$ is the set of real numbers $\mathbb{R}$.

Definition

A linear functional on a real vector space X is a function $L : X \to \mathbb{R}$ such that

(i) $L(x+y) = L(x) + L(y)$ for all $x, y \in X$

(ii) $L(\alpha x) = \alpha L(x)$ for all $x \in X, \alpha \in \mathbb{R}$

Definition

A subset $Y \subset X$ of a vector space X is called a subspace of X if Y is a vector space, i.e., $0 \in Y$, and

$$\alpha x + \beta y \in Y \text{ for all } x, y \in Y \text{ and } \alpha, \beta \in \mathbb{R}.$$

Theorem

Let $(X, \|\cdot\|)$ be a normed linear space, Y be a subspace of X, and $f \in X$.

If f does not belong to the closure of Y, then there exists a bounded linear functional $L : X \to \mathbb{R}$ such that

$$L(x) = 0 \quad \text{if } x \in Y, \text{ and}$$
$$L(f) \neq 0$$

Proof

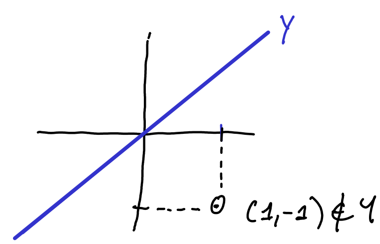
The theorem is a corollary to the Hahn-Banach theorem

Example

$X = \mathbb{R}^2$

$Y = \{ (x, x) \mid x \in \mathbb{R} \}$

$L((x,y)) = x - y$



Proof (by contradiction)

Let $S \subset C(I_n)$ be the set of functions that can be described on the form G.

The statement of the theorem is equivalent to the claim that the closure of S equals $C(I_n)$

Assume by contradiction that the closure of S is a strict subset R of $C(I_n)$.

It follows by the previous theorem that there must exist a linear functional $L : C(I_n) \to \mathbb{R}$ such that

$$L(g) = 0 \qquad \text{for all} \quad g \in R$$

$$L(h) \neq 0 \qquad \text{for some} \quad h \in C(I_n) \setminus R$$

By the Riesz representation theorem, there exists a signed measure $\mu \neq 0$ such that

$$L(f) = \int_{I_n} f(x)\, d\mu(x) \qquad \text{for all} \quad f \in C(I_n)$$

Since $\sigma(w_j^T x + \theta_j) \in R$ for all w_j, θ_j we have

$$L\left(\sigma\left(w_j^T x + \theta_j\right)\right) = \int_{I_n} \sigma\left(w_j^T x + \theta_j\right) d\mu(x) = 0$$

However, this contradicts our assumption that σ is discriminatory.

The theorem now follows.

$\square$

The Power of Depth

Cybenko's result does not tell us

- how many units are needed in the hidden layer
- how difficult it is to train the network to approximate the function

(see the paper on canvas)

Theorem (Eldan & Shamir, 2016)

If the activation function σ satisfies some weak assumptions then there is a function $g : \mathbb{R}^n \to \mathbb{R}$ and a probability measure μ on $\mathbb{R}^n$ such that

polynomial width

1) g is "expressible" by a 3-layer network of width $O(n^{19/4})$,

2) every function f expressed by a 2-layer network of width at most ce^{cn} satisfies

$$\mathbb{E}_{x \sim \mu} \left(f(x) - g(x) \right)^2 \geq c$$

exponential width

Summary

- Rethinking generalisation in deep learning
- Universal Approximation
  - Perceptrons (Rosenblatt, 1957)
  - Perceptrons and the XOR function (Minsky & Papert, 1969)
  - Universal Approximation theorem (Cybenko, 1989)
  - Power of depth (Eldan & Shamir, 2016)

Next week

- Guest lecture by Dr Fernando (Deepmind)
  - Wednesday November 22nd, 13-14 in Mech Eng G31
- No Tuesday lecture
- Friday lecture
  - Regularisation techniques