IntroML - Lecture Notes Week 6

Ruben Schenk, ruben.schenk@inf.ethz.ch

July 15, 2022

1 Neural Networks

1.1 Features

The success in learning crucially depends on the quality of **features.** But what about kernel methods? Don't they yield "universal" features?

- They provide a set of rich feature maps: can approximate "any function" given infinite data
- Give finite data, the choice of the kernel matters a lot!
- Choosing the "right" kernel can be challenging

1.2 Learning Features

Can we **learn** good features from data directly? Consider learning with m hand-designed features:

$$w^* = \arg\min_w \sum_{i=1}^n l\Big(y_i;\, \sum_{j=1}^m w_j \phi_j(x_i)\Big)$$

The key idea is to parameterize the feature maps, and optimize over the parameters:

$$w^* = \arg\min_{w,\, \theta_j} \sum_{i=1}^n l\Big(y_i;\, \sum_{j=1}^m w_j \phi(x_i;\, \theta_j)\Big)$$

But how do we parameterize feature maps? The design idea is as follows. We build *complex models* out of simple components, for example:

$$\phi(x,\, \theta) = \rho(\theta^T x),$$

where $\theta \in \mathbb{R}^d$ and $\rho : \mathbb{R} \to \mathbb{R}$ is a non-linear (simple) *activation function.*

Examples: Following some simple activation functions:

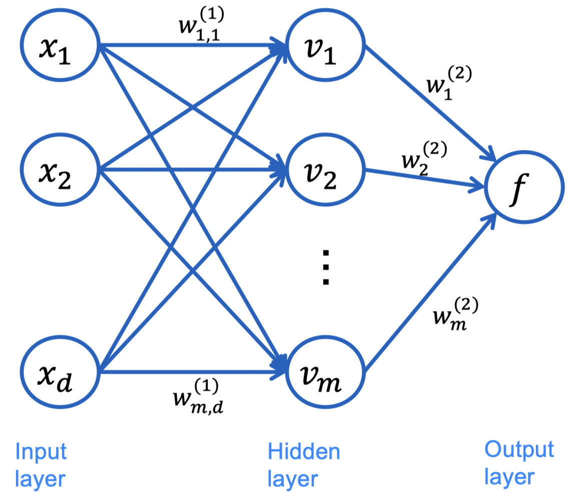
- Identity: $\rho(z) = z$
- Sigmoid: $\rho(z) = \frac{1}{1+\exp(-z)}$
- Tanh: $\rho(z) = \tanh(z)$
- Rectified linear unit (ReLU): $\rho(z) = \max(0,\, z)$

1.3 Artificial Neural Networks (ANNs)

Nested functions of the form

$$f(x;\, w,\, \theta) = \sum_{j=1}^{m} w_j \rho(\theta_j^T x)$$

are examples of **artificial neural networks (ANNs),** also called *multi-layer perceptrons.* More generally, the term artificial neural network refers to non-linear functions which are nested compositions of (learnable) linear functions composed with (fixed) non-linearities.



We can deal with biases by introducing a "constant 1" feature not only to the inputs, but also to the hidden layers. These "constant 1" units have no incoming connections/weights.

1.4 Universal Approximation Theorem

> **Theorem:** Let f be any continuous function on $[0,\, 1]^d$ and ρ any sigmoidal activation function. Then f can be uniformly approximated by a finite sum of the form:
>
> $$\hat{f}(x) = \sum_{j=1}^{m} w_j^{(2)} \rho(w_j^{(1)T} x + w_{j,0}^{(1)})$$