SNS COLLEGE OF TECHNOLOGY

Coimbatore-35
An Autonomous Institution

Accredited by NBA – AICTE and Accredited by NAAC – UGC with 'A++' Grade
Approved by AICTE, New Delhi & Affiliated to Anna University, Chennai

DEPARTMENT OF COMPUTER APPLICATIONS

23CAT702 – Machine Learning

II YEAR III SEM

UNIT I – FOUNDATION OF LEARNING

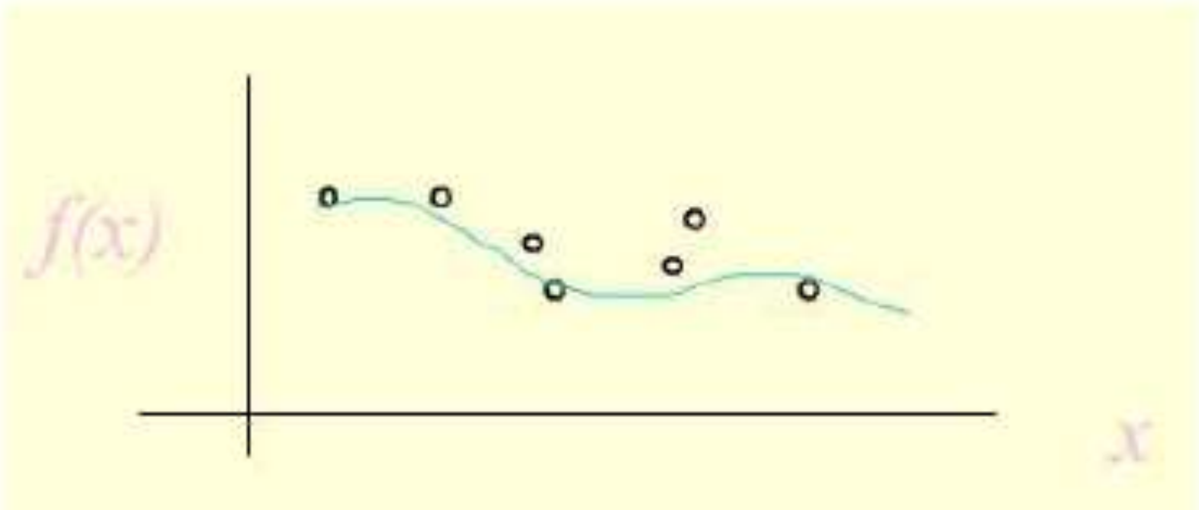
TOPIC 7– Theory of Generalization

Theory of Generalization

- In machine learning, generalization usually refers to the ability of an algorithm to be effective across a range of inputs and applications
- Our key working assumption is that data is generated by an underlying, unknown distribution D. Rather than accessing the distribution directly, statistical learning assumes that we are given a training sample S, where every element of S is i.i.d and generated according to D. A learning algorithm chooses a function (hypothesis h) from a function space (hypothesis class) H where $H = \{f(x, \alpha)\}$ where α is the parameter vector
- We can then define the generalization error of a hypothesis h as the difference between the expectation of the error on a sample x picked from the distribution D and the empirical loss

Generalization

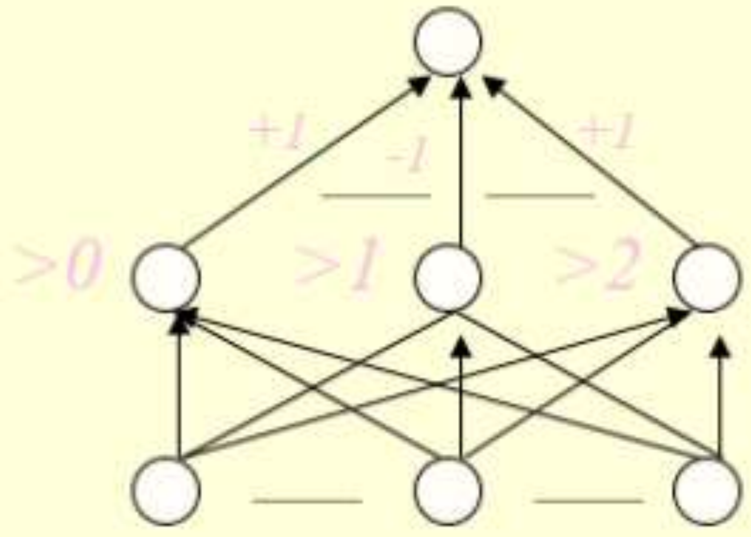
- The objective of learning is to achieve good generalization to new cases, otherwise just use a look-up table
- Generalization can be defined as a mathematical interpolation or regression over a set of training points:



Generalization

An Example: Computing Parity

Parity bit value



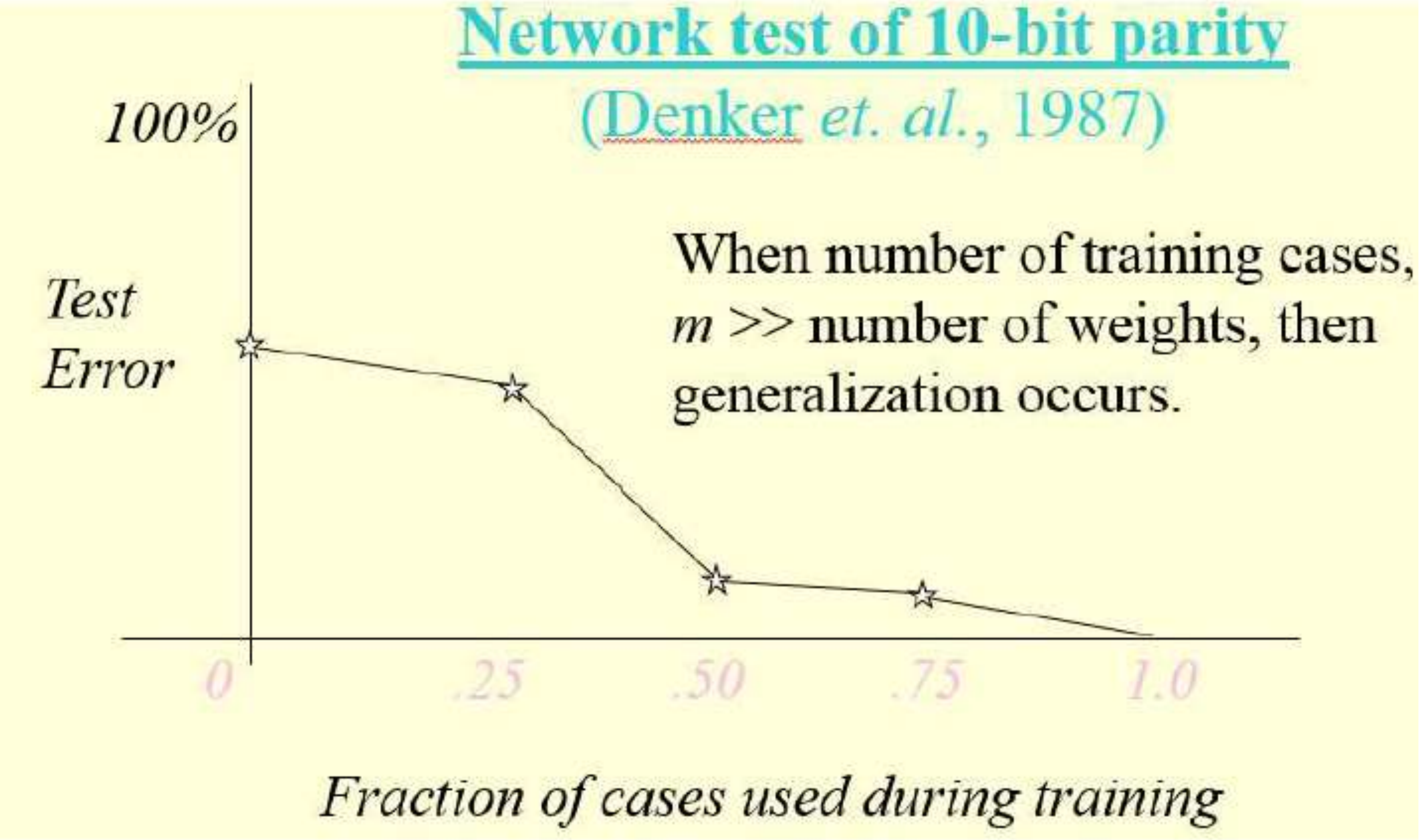
(n+1)^2 weights

n bits of input

⇧

2^*n* possible examples

Can it learn from *m* examples to generalize to all 2^*n* possibilities?


Network test of 10-bit parity
(Denker et. al., 1987)
100%
Test
Error
When number of training cases, m >> number of weights, then generalization occurs.
0
.25
.50
.75
1.0
Fraction of cases used during training

A Probabilistic Guarantee

N = # hidden nodes m = # training cases

W = # weights ε = error tolerance (< 1/8)

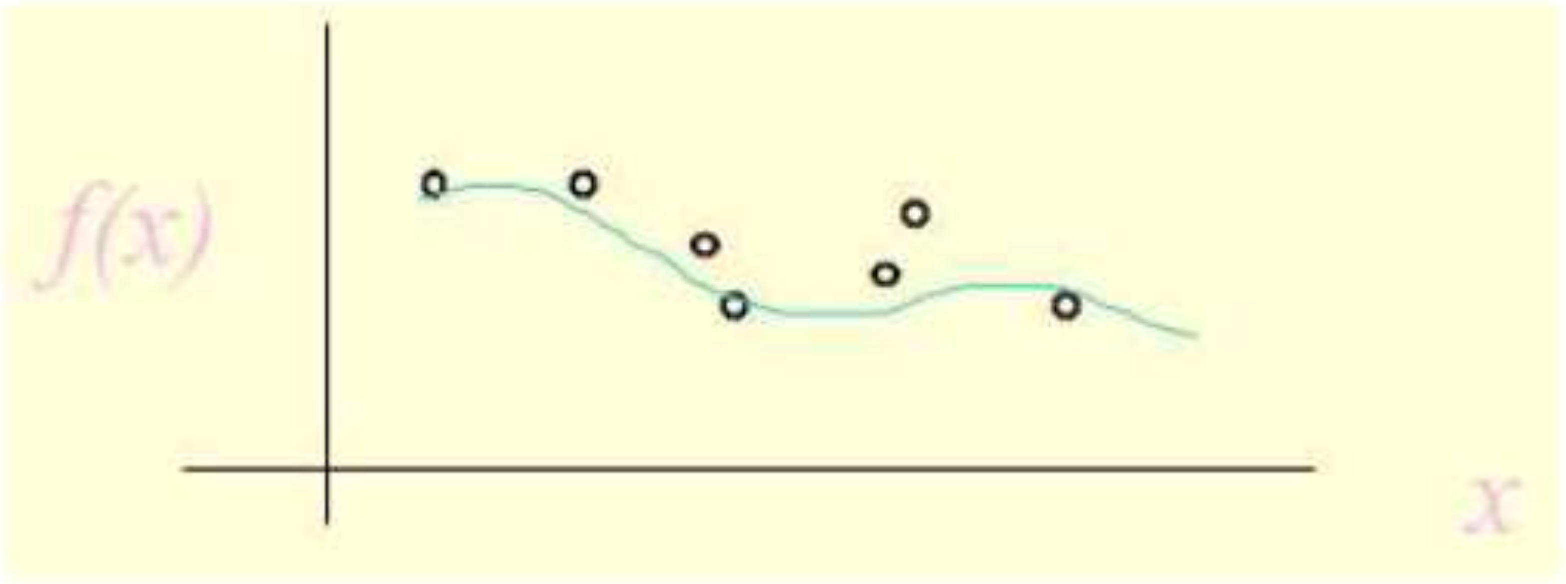
Network will generalize with 95% confidence if:

1. Error on training set < $\varepsilon / 2$
2. $$m > O(\frac{W}{\varepsilon}\log_2\frac{N}{\varepsilon}) \approx m > \frac{W}{\varepsilon}$$

Based on PAC theory => provides a good rule of practice.

Generalization

- The objective of learning is to achieve good generalization to new cases, otherwise just use a look-up table
- Generalization can be defined as a mathematical interpolation or regression over a set of training points:



Over-Training

- Is the equivalent of over-fitting a set of data points to a curve which is too complex
- Occam's Razor (1300s): "plurality should not be assumed without necessity"
- The simplest model which explains the majority of the data is usually the best

Preventing Over-training

- Use a separate test or tuning set of examples
- Monitor error on the test set as network trains
- Stop network training just prior to over-fit error occurring- early stopping or tuning
- Number of effective weights is reduced
- Most new systems have automated early stopping methods

How can we control number of effective weights?

- Manually or automatically select optimum number of hidden nodes and connections
- Prevent over-fitting = over-training
- Add a weight-cost term to the bp error equation

In order for the entire hypothesis space to have a generalization gap bigger than ϵ, at least one of its hypothesis: h_1 **or** h_2 **or** h_3 **or** ... etc should have. This can be expressed formally by stating that:

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}|R(h) - R_{\text{emp}}(h)| > \epsilon\right] = \mathbb{P}\left[\bigcup_{h\in\mathcal{H}}|R(h) - R_{\text{emp}}(h)| > \epsilon\right]$$

Where $\bigcup$ denotes the union of the events, which also corresponds to the logical **OR** operator. Using the union bound inequality, we get:

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}}|R(h) - R_{\text{emp}}(h)| > \epsilon\right] \leq \sum_{h\in\mathcal{H}}\mathbb{P}[|R(h) - R_{\text{emp}}(h)| > \epsilon]$$

We exactly know the bound on the probability under the summation from our analysis using the Heoffding's inequality, so we end up with:

Generalization Bound

$$\mathbb{P}\left[\sup_{h\in\mathcal{H}} |R(h) - R_{\text{emp}}(h)| > \epsilon\right] \leq 2|\mathcal{H}| \exp(-2m\epsilon^2)$$

Where $|\mathcal{H}|$ is the size of the hypothesis space. By denoting the right hand side of the above inequality by δ, we can say that with a confidence $1 - \delta$:

$$|R(h) - R_{\text{emp}}(h)| \leq \epsilon \Rightarrow R(h) \leq R_{\text{emp}}(h) + \epsilon$$

And with some basic algebra, we can express ϵ in terms of δ and get:

$$R(h) \leq R_{\text{emp}}(h) + \sqrt{\frac{\ln |\mathcal{H}| + \ln \frac{2}{\delta}}{2m}}$$

Generalization Bound

- There are two types of bound
  - VC generalization bound
  - Distributed function based bound

VC generalization bound

$$R(h) \lessapprox \widehat{R}_n(h) + \epsilon(\mathcal{H}, n)$$

Reference

- Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, ―Learning from Data, AML Book Publishers, 2012.
- P. Flach, ―Machine Learning: The art and science of algorithms that make sense of data‖, Cambridge University Press, 2012.
- W3school.com


Thank you