Lecture 20

Statistical Learning for Neural Networks

Guaranteed Success for ERM with Finite Hypotheses

- **Accuracy ϵ and Tolerance δ**
  - ϵ is the desired maximum error rate of the hypothesis found by ERM.
  - δ is the tolerance for the probability of failure, i.e., the probability that the ERM hypothesis will exceed the error rate ϵ.
- **The Bound**

  With $m > \frac{\log(|H|/\delta)}{\epsilon}$, ERM will find a hypothesis with error less than ϵ with probability greater than $1 - \delta$.

  For a large enough m, ERM produces a hypothesis with error under ϵ at probability $1 - \delta$.

VC Dimension

- **Definition:**
  - VC Dimension is the maximum number of data points that a model class can shatter.
- **Shattering:**
  - A model class can shatter a set of data points if it can perfectly classify every possible arrangement of labels for that set.

VC Dimension of a Line in 2D

- In a two-dimensional space, the VC dimension of a line is 3.

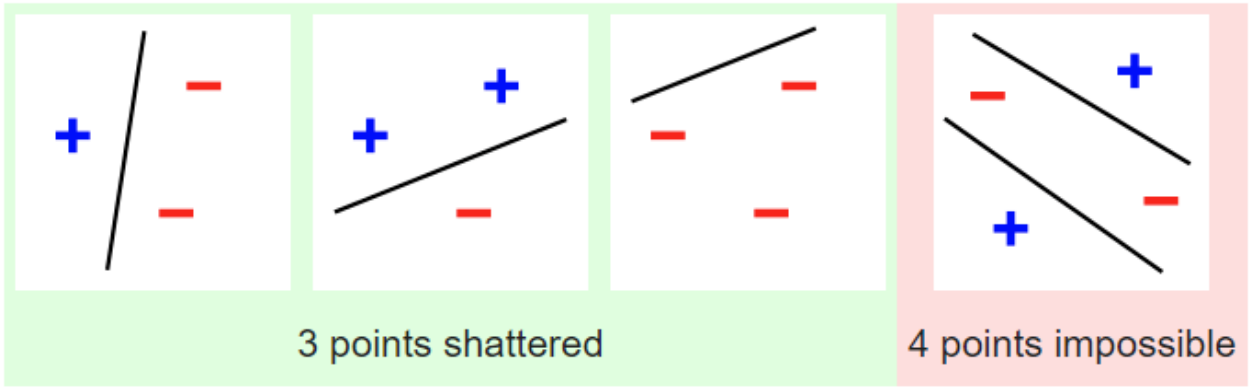

Image: Wikepedia

PAC (Probably Approximately Correct) learnability

- Being able to learn a good-enough hypothesis with high probability given enough examples.

- If with enough data, a model from H can be learned that is probably correct (within ϵ error) with high confidence (probability > $1 - \delta$), then H is PAC Learnable.

H is PAC Learnable if:

- there is a function $m_H : (0,1)^2 \to \mathbb{N}$
- and a learning algorithm A,
- such that for every distribution D over X,
- every $\epsilon, \delta > 0$ and every f in H,
- for samples S of size m ($m = m_H(\epsilon, \delta)$)
- generated by D and labeled by f,
- $\Pr[L_D(A(S)) > \epsilon] < \delta$.

Relaxing the Realizability Assumption

- **Realizability Assumption**:
  - Assumes True Function in Hypothesis Class H
  - Hypotheses Contain Exact Solution
- **Realistic Setup**:
  - **Lack of A Priori Knowledge**:
    - Learner Doesn't Know if True Classifier in H

Unpredictability of Labels:

- Another aspect of more realistic scenarios: labels (the outputs we're trying to predict) might not be fully determined by the instance attributes (the input data).
  - Various reasons: noise in the data, unobserved variables, or inherently stochastic processes.
  - In such cases, even the best possible model in H might not perfectly predict the labels for all instances.

Implications for Learning:

- When these more realistic conditions are assumed, learning becomes more challenging.
- The learner must now find the best possible hypothesis within the class, even if none of the hypotheses can perfectly predict all instances.

General Loss Functions

- **Beyond Classification Errors**:
  - Our learning approach goes beyond counting classification errors.
- **Domain Set Z**:
  - Z represents the set of all possible instances or data points.
- **Loss Function** l:
  - $l : H \times Z \rightarrow \mathbb{R}$
  - Quantifies model h's loss on instance z.

General Loss Functions

- **Probability Distribution P:**
  - P is a probability distribution over Z.
  - Defines the likelihood of each instance.
- **Expected Loss $L_P(h)$:**
  - $L_P(h) = E_{z \sim P}[l(h, z)]$
  - Average loss of model h under P.
- **Flexible Learning Formalism:**
  - Allows us to assess model performance with various loss functions and distributions.

Agnostic PAC Learnability

- H is Agnostic PAC Learnable if:
  - There exists a function $m_H : (0,1)^2 \to \mathbb{N}$ and a learning algorithm A.
  - For every distribution P over $X \times Y$ and every $\epsilon, \delta > 0$.
  - For samples S of size $m > m_H(\epsilon, \delta)$ generated by P.

$$\Pr[L_P(A(S)) \leq \inf_{h \in H} L_P(h) + \epsilon] \geq 1 - \delta$$

Guaranteeing Learnability

- **Question**: Can such learnability be guaranteed?
- **Crucial Factor**: The VC (Vapnik-Chervonenkis) dimension of the class H.

- **Fundamental Theorem of Statistical Learning**:
  - "A class H is PAC (Probably Approximately Correct) learnable if and only if its VC dimension is finite."

Fundamental Theorem - Quantitative Version

- The number of random labeled samples needed for learning a class of predictors H is given by:

$$O\left(\frac{VCdim(H)+\log\left(\frac{1}{\delta}\right)}{\epsilon^2}\right)$$

Complexity-Dependent Error Bound

- $L_P(h) \leq \inf_{h \in H} L_P(h) + f(c(h); m)$
- The true error of a hypothesis h is bounded by a function of its complexity and training sample size.

Function $f(c(h); m)$

- Dependent on the complexity measure $c(h)$ of hypothesis h and sample size m.
- Indicates a relationship between hypothesis complexity, sample size, and the achievable error.

Integrates the concept of hypothesis complexity into the error bound.

Expressive Power of Neural Networks

Theorem: Fix some ϵ in the range $(0, 1)$, and let $s(n)$ denote the minimal number of nodes.

- There exists a neural network with $s(n)$ nodes that can approximate up to ϵ every function from $[0, 1]^n$ to $[0, 1]$.
- $s(n)$ is exponential in n.

Expressive Power of Neural Networks

Theorem: Fix some ϵ in the range $(0, 1)$, and let $s(n)$ denote the minimal number of nodes.

- There exists a neural network with $s(n)$ nodes that can approximate up to ϵ every function from $[0, 1]^n$ to $[0, 1]$.
- $s(n)$ is exponential in n.

- **Key Insight**:
- Neural networks possess remarkable expressive power.
- They can approximate a wide range of functions with high accuracy.
- However, achieving this expressive power may require a large number of nodes, which can grow exponentially with the input dimension n.

Measuring Error Guarantees For Neural Networks

$$L_P(h) \leq \inf_{h \in H} L_S(h) + \sqrt{\frac{|E| + \log(1/\delta)}{m}}$$

Where:

- $|E|$ is the number of edges (parameters) in the hypothesis h.
- m is the sample size.

The authors experiments indicate these reasons dont fully account for the networks generalization success.

Rethinking Generalization in Deep Learning

- **Authors:** Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, Oriol Vinyals
- **Key Insights:**
- A paradox in deep learning: large neural networks manage to generalize well and maintain a minimal gap between training and testing performance, contrary to what might be expected given their size.
- Small generalization errors are commonly linked to model traits or training regularization.
- The authors' experiments indicate these reasons don't fully account for the networks' generalization success.

Experiment on Generalization in Deep Networks

- State-of-the-art convolutional neural networks for image classification, when trained with stochastic gradient methods, can easily fit to randomly labeled training data.
- This ability to fit random labels is not significantly impacted by the use of explicit regularization techniques.
- The phenomenon persists even when true image data is replaced with unstructured random noise.
- Theoretical work supports these findings by showing that neural networks with a sufficient number of parameters relative to data points can perfectly express any finite sample set.

Experimental Details

- **Data Modification:**
  - True labels in standard datasets (CIFAR10, ImageNet) were replaced with random labels.
  - In an extension of the experiment, actual image pixels were replaced with completely random noise.
- **Training Results:**
  - Neural networks reached zero training error on randomly labeled data, suggesting they can memorize the dataset.
  - Test error was equivalent to random guessing due to lack of correlation between training and test labels.
  - Training times were only slightly longer than with true labels, indicating ease of optimization even with random data.

Experimental Details

- **Impact of Noise:**
  - The introduction of noise to the images did not prevent neural networks from fitting the data.
  - A progressive increase in noise led to a corresponding increase in generalization error, yet networks could still capture any signal left in the labels.
- **Implications:**
  - These results challenge the role of VC-dimension, Rademacher complexity, and uniform stability in explaining generalization in neural networks.
  - The networks' ability to fit random labels and noise points to a high capacity for memorization, which is not accounted for by traditional learning complexity measures.

Uniform convergence may be unable to explain generalization in deep learning

Authors:

Vaishnavh Nagarajan, Zico Kolter

Main Claim:

- The paper challenges the adequacy of uniform convergence as a tool for explaining the generalization behavior in overparameterized deep neural networks.
- It highlights a key finding: generalization bounds based on uniform convergence can paradoxically increase with the size of the training dataset.

Understanding Uniform Convergence

- Uniform convergence is a concept in statistical learning theory. It describes how closely the empirical loss (loss on training data) of a learning algorithm converges to the expected loss (loss on the entire data distribution) uniformly over all hypotheses in a hypothesis class.
- This concept is crucial for establishing generalization bounds, which predict how well a model trained on a finite dataset will perform on unseen data.
- The idea is that if the empirical loss converges uniformly to the true loss across all hypotheses, one can be confident that a hypothesis with low empirical loss will also have low true loss, hence will generalize well.

Implications of the Findings

- The paper demonstrates scenarios with overparameterized models (like deep neural networks with more parameters than training data points) where uniform convergence fails to explain generalization.
- It shows that even when considering only hypotheses output by gradient descent with low test errors, uniform convergence provides vacuous (ineffective) generalization guarantees.
- These findings cast doubt on the ability of uniform convergence-based bounds to fully explain why large neural networks generalize effectively, suggesting the need for alternative or additional theoretical frameworks.