Lecture 1

February 13, 2021

1 Course Intro & Details

- Course Site
- Grading: 90% Project, 10% participation
- Office Hours: Thu 9:30 am-10:30 am
- References: Matus Telgarkis: Theory Of DL
- Concentrated on supervised learning

2 Supervised Learning

2.1 Parameters

2.1.1 Input domain

X is in Higher dimensions
$X \in R^d$ or X can be space of input images

2.1.2 Data Distribution

ν: data distribution over X;
$\nu \in P(X)$ - set of all probabilities over X

2.1.3 Target Function f*

$X \to R$ regression task
Ex: X - space of all possible molecules
f*(X) - Energy of molecule x

2.1.4 Risk/Loss function

$$L(f) = E_{x \leftarrow \nu}[l(f(x), f^*(x))]$$

In particular,

$$L(f) = E_{x \leftarrow \nu}|f(x) - f^*(x)|^2$$

L is convex w.r.t f so that we can apply gradient descent
we can write it as $||f(x) - f^*(x)||_\nu^2$

2.2 Goal: Predict target f* from finite no of observations

Under the assumption observations are sample from data distribution,

$$\nu : \{X_i, f^*(X_i)\}_i$$

2.3 Empirical Risk Minification

Consider a hypothesis space F $\subseteq \{f : X \rightarrow R\}$
Assume that F is a Normed space ie we can use a complexity measure $\gamma : f \rightarrow R$
$\gamma(f)$ measures how complex the hypothesis $f \in Fis$ Ex: gamma can be the weights of the n/w
Ex: F = class of Neural networks of a certain architecture
Then,

$$F = \{f : X \in R; f(x) = \Phi(w_k, x^k) + .. + \Phi(w_2, x^2) + \Phi(w_1, x)\}$$

$\Phi :$ Activation function
Let's consider a ball,

$$F_\delta = \{f \in F; \gamma(f) \subseteq \delta\}$$

In particular, F_δ is a convex set

2.3.1 Empirical Risk

Recall $L(f) = E_{X \sim \nu}[l(f(x), f^*(x)]$,
E: Expectaion, l: loss function

$$\hat{L}(f) = \frac{1}{L} \sum_{l=1}^{L} l(f(X_l), f^*(X_l)$$

Ex: $\frac{1}{L}\sum_l |f(X_l) - f^*(X_l)|^2$ for MSE

2.4 Structural Risk Minimization

1. We want small empirical risk with small complexity. (simplest hypothesis to solve the problem).

2. There are 3 forms to solve this problem:

   (a) Constrained form

   $$\min_{\gamma(f)\subseteq\delta} \hat{L}$$

   $\gamma(f)$ is the complexity of f
   It can be considered as a constraint optimization problem

   (b) Penalised form(Relaxation of the constraint)

   $$\min_{f\in F} \hat{L}(f) + \lambda.\gamma(f)$$

   (c) Interpolating Form

   $$\min_{\hat{L}(f)=0} \gamma(f)$$

   It is a variation of constrained and penalised form
   If the observations contain true fn & no noise.
   Find hypothesis that agrees with data & smallest complexity
   Makes sense only when labels have no noise

All these formulations are related to each other.
Most Algo implementations use penalised form but we will discuss constrained method for discussing learning methods

3 Basic Decomposition of Error

Suppose we use constrained form and assume we have found $\hat{f}$:

$$\hat{L}(\hat{f}) \leq \min_{f\in F_\delta} \hat{L}(f) + \epsilon_o \quad s.t \;\; \hat{f} \in F_\delta$$

3.1 How good is our $\hat{f}$ at our Goal?

$$
\begin{aligned}
L(\hat{f}) - \min_{f\in F} L(f) &= L(\hat{f}) - \min_{f\in F_\delta} L(f) + \{\min_{f\in F_\delta} L(f) - \min_{f\in F} L(f)\} \\
&= \hat{L}(\hat{f}) - \min_{f\in F_\delta} L(f) + L(\hat{f}) - \hat{L}(\hat{f}) + \epsilon_a \\
&\le \{\min_{f\in F_\delta} \hat{L}(\hat{f}) - \min_{f\in F_\delta} L(f)\} + \{L(\hat{f}) - \hat{L}(\hat{f})\} + \epsilon_a + \epsilon_o \\
&\le \{2 \sup_{f\in F_\delta} |L(f) - \hat{L}(f)|\} + \epsilon_a + epsilon_o \\
&\le \epsilon_s + \epsilon_a + epsilon_o
\end{aligned}
$$

3.1.1 Comparison of errors

- Approximation Error ϵ_a
  Inversely Proportional to δ
  cancel due to Universal Approximation Theorem Any function can be rep by 2 layer NN.

- Statistical Error ϵ_s
  Ensure that hypothesis space is not big.
  Deviation b/w population loss and training loss are under control.
  If we remove the sup and rewrite this becomes

  $$E_\nu($$

  from law of large number(CLT)
  From stats, $\sqrt{\frac{Complexity(\delta)}{L}}$

- Optimization Error ϵ_o
  Ensure this is less in our ML model

In terms of NN,

$$
\begin{aligned}
\epsilon_a &= \min_{f\in F_\delta} ||f - f^*||^2 - \cancel{\min_{f\in F} ||f - f^*||^2} \\
\epsilon_s &= 2 \sup_{f\in F_\delta} |L(f) - \hat{L}(f)|
\end{aligned}
$$

4 Big Question

- How to define functional spacesF with good approximation properties in High dimns?

- Algo to solve effeciently solve ERM

- Balls F_δ should not grow too quickly with dimns
  (stats error under control)

4.1 Part I: Theory (Foundation of Geometric Deep learning)

premise in most applns, dimensionality of X is “hiding” an underlying low dimension structure.
Ex: images $X \in R^d$. X is also X(u), $u \in \Omega$
Math:

- Harmonic Analysis
- Signal Processing
- Graph Spectral Theory

4.2 Part II: Foundations of DL: Optimizations in DL

Stats of fully connected NNs

5 The curse of dimensionality

Focus: Understand how approximation/statistical/optimization error behaves as a function of input dimensionality.
Imagine we are in the space and the input are stars which are very far away.

5.1 Statistical Curse

We observe $f^*(X_1) \cdots f^*(X_L)$

- How many observations L as fn of d and hypothesis f* ?
  - Suppose first that f* is linear:
    $f^*(X) =< X, \theta^* > \quad \theta \in R^d$

    $$F = \{f : R^d \to R; f(x) =< x, \theta >\}$$

    This can be solved by d samples as there d equations to be solved
  - f* is locally linear?
    Then it is a lipschitz function.
    $|f(x) - f(\tilde{x})| \leq \beta ||x - \tilde{x}||$
    Now our hypothesis space F becomes,

$$F = \{f : R^d \to R; f is Lipschitz\}$$

Then our complexity of F becomes the Lipschitz constant
$\gamma(f) = Lip(f) + ||f||_{inf}$
we want:
$\forall \epsilon > 0$, find $f \in F$ s.t
$||f - f^*||_{L^2} \leq \epsilon$
L id samples $\{X_l, f^*(X_l)\}_l$

- ∗ How large L needs to be to achieve error of ϵ ?
  $L \sim \epsilon^{-d}$
  Image a cube of 1 unit.
  For covering a line of 1 unit with distance of ϵ,
  we need to have $\frac{1}{\epsilon}$ points. Extending for
  a cube, we need to have $\frac{1}{\epsilon^3}$ points.

  Given $(X_i, f^*(X_i))$
  Our estimator can be formulated as

$$\hat{f} = arg \min_f Lip(f) \quad s.t \quad f(X_i) = f^*(X_i) \forall i$$

Let us consider a point X where the closest training part in $\bar{X}$
Then,
$|\hat{f}(X) - f^*(X)| \leq |\hat{f}(X) - \hat{f}(\bar{X})| + |\hat{f}(\bar{X}) - f^*(\bar{X})| + |f^*(\bar{X}) - f^*(X)|$