Lecture 1

February 1, 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
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 $\subseteq \{f : X \to R\}$
Assume that F is a normed space ie we can use a complexity measure $\gamma : f \to R$
$\gamma(f)$ measures how complex the hypothesis $f \in Fis$
Ex: F = class of Neural networks of a certain architecture
Then,

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

Φ : 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. Minimize $\hat{L}$ (Constrained form); $\gamma(f) \subseteq \delta$
   δ is the complexity
   It can be considered as a constraint optimization problem

3. This can be relaxed by using a penalised form

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

4. There is another popular method which is a variance of the above 2 equations

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

   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