Online Learning : Preliminaries

Sunday, 23 April 2023 6:23 PM

Traditional ML Methods assume that entire dataset D is available before training starts. In this case we perform "batch learning".

In many cases data arrives sequentially as an unbounded stream. In this case we want to perform Online learning

Let θ_{t-1} be our parameter estimate given data points from time $1, 2, \ldots t-1$. We want to update our parameter in constant time when we see the t^{th} point We have to find a rule

$\theta_t = f(\theta_{t-1} | y_t)$. This is called a recursive update.

One way to think about some online prediction tasks is to think of it as a repetitive game between the predictor and the environment

In each round of game

Env: Choose an instance of problem

Predictor: Make a prediction for this instance

Env: Calculate loss for the prediction and send feedback to predictor.

Predictor: Learn and record feedback.

(Lec. Notes from course by Ofer Dekel)

Some Examples :-

MLE for univariate Gaussian

Mean is given as

$$\hat{\mu}_t = \frac{1}{t} \sum_{n=1}^{t} y_n$$

Now

$$\frac{1}{t} \sum_{n=1}^{t} y_n$$

$$= \frac{1}{t} \left((t-1) \hat{\mu}_{t-1} + y_t \right)$$

$$= \hat{\mu}_{t-1} + \frac{1}{t} \left(y_t - \hat{\mu}_{t-1} \right)$$

This is known as moving average

Suppose the distribution of data is changing, we want to give more weightage to more recent data examples

Exponentially Weighted Moving Average

$$\hat{\mu}_t = \beta \mu_{t-1} + (1-\beta) y_t$$

$$0 < \beta < 1$$

$$\hat{\mu}_t = \beta \mu_{t-1} + (1-\beta) y_t$$

$$= \beta^2 \mu_{t-2} + \beta(1-\beta) y_{t-1} + (1-\beta) y_t$$

$$= \beta^t y_0 + (1-\beta) \beta^{t-1} y_1 + \ldots + (1-\beta) \beta y_{t-1} + (1-\beta) y_t$$

So the contribution of k^{th} data point is weighted by $\beta^k (1-\beta)$.

Now

$$\beta^t + \beta^{t-1} + \cdots \beta^1 + \beta_0 = \frac{1-\beta^{t+1}}{1-\beta}$$

$$\therefore (1-\beta) \sum_{k=0}^{t} \beta^k = (1-\beta) \frac{1-\beta^{t+1}}{(1-\beta)}$$

$$= 1 - \beta^{t+1}$$

As $t \to \infty$, $\beta^{t+1} \to 0$ $(\because 0 < \beta < 1)$

(SMALLER β forgets past more quickly)

Since $\hat{\mu}_0 = 0$, there is some bias To correct it we can scale

$$\tilde{\mu}_t = \frac{\hat{\mu}_t}{1-\beta^t}$$

Ex. 2 :-

Algorithm for online linear Regression

(Lec. Notes from course by Rob Schapire)

- Initialize $W_1 = 0$ (Widrow Hoff Algorithm)
- Choose $\eta > 0$
- For $t = 1, \ldots T$

  Get x_t

  $\hat{y}_t = w_t \cdot x_t$

  Observe y_t

  Update $w_{t+1} = w_t - \eta (w_t \cdot x_t - y_t) \cdot x_t$

$$L_{WH} \leq \min_{u} \sum_{t=1}^{T} (u \cdot x_t - y_t)^2 + \text{small number}$$

This is just the tip of the iceberg.

Online learning covers topics like

- Online convex Opt
- FTL & FTRL Algos.
- Online Learning with Expert Advice
- Adaptive Algorithms