CS 419M Introduction to Machine Learning **Spring 2021-22**

Lecture 3: Probabilistic estimators

Lecturer: Prof Abir De *Scribe: Team G2*

3.1 Topics covered in lecture

1. Surrogate functions in deterministic space
2. Probabilistic interpretation of training problem
3. Maximum likelihood Estimator
4. Approximating function in general situation

3.1.1 Shift from Deterministic functions to Probabilistic estimators

In the last lecture, we saw the approach of defining different deterministic surrogate functions that would potentially help us predict y from x

1. Today's lecture began with discussing two such deterministic surrogate functions-

$$e^{-y\cdot(w^T x)} \tag{3.1}$$

and

$$\max(1-(w^T x)y, 0) \tag{3.2}$$

2. The latter, also called the **Hinge Function** has an alternate nomenclature of

$$(1-(w^T x)y)_+ \tag{3.3}$$

and this is what is largely known as the **ReLU function**

$$\text{ReLU}(1-(w^T x)y) \tag{3.4}$$

in the field of ML

3. Another approach to the same problem would be a probabilistic one, where X and Y are sampled from P, an unknown distribution function

3.1.2 Formulation

X is sampled from a probability distribution and Y is sampled from a probability distribution conditioned on x respectively

Now, we try to estimate $\sum_{x,y\in\ D} I(h(x)\neq \text{y})$ in terms of probability.

Using, Law of large numbers

$[\sum_{x,y\in\ D} I(h(x)\neq \text{y})]/|D| \approx E[I(h(x)\neq \text{y})]] = \text{P(h(x)} \neq \text{y)}$

3.2 Probabilistic Distributions

In the previous lecture were attempting to deduce a deterministic functional relationship between the sample data $\mathbf{x}$ and it's associated label y. However it is also possible, and much more probable, that the data is actually probabilistic in nature, due to the nature of errors, ambiguity, etc.
Thus we need to translate our error-function minimization condition into a statement involving probabilities. Expression in probabilities also allows for the assignment of a confidence value to the output of our algorithm.

Q. **How can the expression for error of the predictor function be written in a probabilistic manner?**
Given that $\mathbf{x}$ and y are sampled from a distribution with pdf $f_{XY}(\mathbf{x}, y)$,

$$\frac{1}{N_D}\sum_{(x,y)\in D} \mathbb{1}(h(\mathbf{x}) \neq y) \longrightarrow \mathbb{E}(\mathbb{1}(h(\mathbf{x}) \neq y)) \qquad \text{(Central Limit Theorem)}$$

$$= \int_D \mathbb{1}(h(\mathbf{x}) \neq y) f_{XY}(\mathbf{x}, y)\, d(\mathbf{x}, y)$$

$$= \int_{\{h(\mathbf{x})\neq y\}\subset D} \underbrace{\mathbb{1}(h(\mathbf{x}) \neq y)}_{1} f_{XY}(\mathbf{x}, y)\, d(\mathbf{x}, y)$$

$$+ \int_{\{h(\mathbf{x})= y\}\subset D} \underbrace{\mathbb{1}(h(\mathbf{x}) \neq y)}_{0} f_{XY}(\mathbf{x}, y)\, d(\mathbf{x}, y)$$

$$= \int_{\{h(\mathbf{x})\neq y\}\subset D} f_{XY}(\mathbf{x}, y)\, d(\mathbf{x}, y)$$

$$\implies \text{Error}(h) \simeq N_D \cdot \mathbb{P}(h(\mathbf{x}) \neq y)$$

3.2.1 y Prediction

Since in most problems the instance of interest is known, f_X is not of much interest. We instead aim to approximate the probability distribution of the label y given $\mathbf{x}$. Since y is a binary random variable, it must have Bernoulli distribution. Assuming a sigmoid variation of the success

probability,

$$\begin{aligned} y &\sim \text{Bernoulli}\left(\frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}}\right) \\ \implies \mathbb{P}(y=1 \mid \mathbf{x}) &= \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}} \\ \implies \mathbb{P}(y=-1 \mid \mathbf{x}) &= 1 - \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}} \\ &= \frac{1}{1+e^{\mathbf{w}^T\mathbf{x}}} \\ \implies \mathbb{P}(y \mid \mathbf{x}) &= \frac{1}{1+e^{-y\cdot\mathbf{w}^T\mathbf{x}}} \end{aligned}$$

Note that we have chosen a sigmoid form for h since we need to trim the maximal space over which $h(x)$ exists, otherwise it is too large to conduct any meaningful search. The sigmoid function is good for this purpose becuase it's binary nature in the asymptotic limit.
Thus with this probabilisitc nature of y, we can devise a maximum likelihood estimator for it, instead of a deterministic $h(x)$.

3.2.2 Maximum likelihood estimator

- In general, we fix the function upto a parameter, $P(.|x)$=f(θ,x).
- Now we find $max_\theta \prod_{(x,y)\in D} P_\theta(y|x)$ and get theta which maximises $P(y|x)$ over the function set f(θ,x) and set of data points D.
- Given the distribution of a statistical model f(θ;x) with unkown deterministic parameter θ, MLE is to estimate the parameter θ by maximizing the probability f(θ; x) with observations x
- Instead of product, we take logarithm, ie $max_\theta \sum_{(x,y)\in D} log(P_\theta(y|x))$ (log-likelihood estimate)
- The key properties of maximum likelihood estimator are its unbiasedness and minimum variance achieved over all estimators.
- That is $E(\hat{\theta_i}) = \theta*$
- (We generate θ_i by sampling (x, y) from D)
- Also this method minimises $E(|\hat{\theta} - \theta*|^2$ ie minimises the variance over all estimators.

3.2.3 Approximating function in general situation

- Underlying concept in deep learning is that there is a generic form of distribution.
- As mentioned in class, a linear-ReLU-linear can approximate any non linear function.
- But no such approximation in probability so we come up with a base distribution : Normal, Logistic depending on output.

3.3 Group Details and Individual Contribution

Group Members

Varad Mahashabde - (**200260057**)
Contribution:Formulation and Probabilistic Distributions, y-prediction
Shirish Chinchanikar - (**19B090012**)
Contribution: Maximum Likelihood Estimators and Approximating functions in general situation
Prakriti Shetty - (**200020095**)
Contribution: Introduction, Surrogate functions in Deterministic Space
Yuvraj Singh - (**200070093**)
Contribution: Surrogate functions in Deterministic Space and Formulation
Parth Vijay Dange - (**200020091**)