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

Lecture 3: Probabilistic Approximation of Loss Functions

Lecturer: Abir De Scribe: Kaustav Prasad, Ayush Kapoor, Harshit Shrivastava, Yash Sanjeev

3.1 Choices for a Hinge Loss Function

Problem Given data $\mathcal{D} \equiv \{(x_i, y_i)\}_{1 \leq i \leq n}$ where $x \in \mathbb{R}^d$ and $y \in \{-1, 1\}$, the objective is to find $w \in \mathbb{R}^d$ such that $w^\top w = 1$ and

$$w^\top x > 1 : y = 1 \quad , \quad w^\top x < -1 : y = -1$$

When $|w^\top x| < 1$, we can assign a loss function which is greater when $\text{sgn}(w^\top x) \neq y$. □

The problem translates to the situation where we wish to separate the data points with $y = 1$ from the data with $y = -1$ using the line with normal vector w. The points with $|w^\top x| < 1$ are within the "margin" of the separator, and contribute to the loss regardless of the value of y. The constraint $w^\top w = 1$ is imposed to avoid linear scaling of w, to artificially decrease the width of the "margin". The most widespread choices for the loss function have been outlined below.

Solution The first choice of w proposed is

$$\min_{w^\top w = 1} \sum_{(x,y) \in \mathcal{D}} \exp(-y \cdot w^\top x)$$

Observe that

$$L = 0 \text{ when } y = 1 \text{ and } w^\top x > 1$$
$$L = 0 \text{ when } y = -1 \text{ and } w^\top x < -1$$

so a prudent choice for the loss function is

$$\text{Loss}(w^\top x, y) = \max(0, 1 - y \cdot w^\top x)$$

also known as the hinge loss, useful in the training of Support Vector Machines (SVMs). □

Note Both the definitions of the loss functions over the entire dataset $\mathcal{D}$ give a measure of

$$\sum_{(x,y) \in \mathcal{D}} \mathbb{I}\left(h(x) \neq y\right)$$

where $h : \mathbb{R}^d \longrightarrow \{-1, 1\}$ is given by $h(x) = \text{sgn}(w^\top x)$

3.2 Probabilistic Approximation of Loss Functions

The probabilistic approach is a much more guided and principled one which provides us with some mathematical guarantee as will be shown later in the course. We try to interpret the same problem as above in a different manner. In this Probabilistic Approximation method, we assume that x and y are drawn from a probability distribution and are independent and identically distributed Random Variables.

$$x \sim P(.)$$
$$y \sim P(./x)$$

Now we need to find a probabilistic way of approximating

$$\sum_{(x,y)\in\mathcal{D}} \mathbb{I}\left(h(x) \neq y\right)$$

We know that, when N is very large,

$$\sum_{i=1}^{N} g\left(X_i\right) \text{converges to } \mathbb{E}[g(X_i)]$$

Similarly,

$$\frac{1}{|D|} \sum_{(x,y)\in\mathcal{D}} \mathbb{I}\left(h(x) \neq y\right) \text{converges almost surely to } \mathbb{E}[\mathbb{I}(h(x) \neq y)] = P(h(X) \neq y)$$

Which is nothing but the Misclassification Probability.
In most problems, we are given with $P_x(.) \sim x$ and we try to find a $\hat{y} \sim P_y(./x)$ such that the misclassification probability is minimised. This is not an explicit function like the h(.) discussed earlier but rather a probabilty density we shall find for an estimator given x such that we achieve the same end goal as before, that is to minimise the loss, or equivalently in this case, misclassification probability.

3.2.1 Estimating w from Labelled Data

In order to estimate the vector w from data, we shall make the following assumption:

Assumption: *The labelled data $D = \{(x_i, y_i)\}$ is derived from a Probability Distribution with some fixed parameter w^*.*

For example, let us state that the probability distribution that generates the data is of the form

$$P(y|x) = \frac{1}{1 + e^{-w^{*T} x \cdot y}} \quad y \in \{-1, 1\}$$

Suppose that we have generated some data i.e a set of pairs (x_i, y_i) from the distribution above. We wish to estimate w^* by *only* using the generated data, and presume no *a priori* knowledge of $w*$. This can be accomplished by using the **Maximum Likelihood Estimator** on the paramter w for the same distribution.

Let

$$P(y|x)_w = \frac{1}{1 + e^{-w^T x \cdot y}} \quad y \in \{-1, 1\}$$

We have taken the sample distribution to be the same as the distribution which generated the data, but in this case w is an unknown parameter.

The Maximum Likelihood Estimate for w is the value of w for which

$$\sum_{x,y \in D} log\, P_w(y|x)$$

is maximized. Let us write the value of w obtained in this way as $\hat{w}$.

Q. **What is the relationship between $\hat{w}$ and w^*?**
Ans. Let us say that we derive n independent data sets from the $w*$ parametrized distribution and apply MLE to each of those data sets to obtain a set $\hat{w_1}, \hat{w_2} \cdots \hat{w_n}$ of estimates. Then, the expectation value of the members of this set is $w*$, i.e.

$$\mathbb{E}[\hat{w_i}] = w^*$$

3.2.2 Use of w in estimating the $P(y|x)$ for $(x, y) \notin D$

For given (x_i, y_i) part of data set D, we use **Maximum Likelihood Estimator(MLE)** to estimate $\hat{w}$ for that particular data set D as the value of w for which underlying expression is maximum.

$$\sum_{x,y \in D} log\, P_w(y|x) \quad where \qquad P_w(y|x) = \frac{1}{1 + e^{-w^{*T} x \cdot y}}$$

And consequently if we consider n independent data sets to obtain $\hat{w_1}, \hat{w_2} \cdots \hat{w_n}$ and using those values determine the expectation value as

$$\mathbb{E}[\hat{w_i}] = w^*$$

As expectation value provides the minimum variance from the set of values $\hat{w_1}, \hat{w_2} \cdots \hat{w_n}$. When a probability for a new pair of (x, y) is evaluated, w* will be the best approximation because it is least biased towards any $\hat{w_i}$.

$$min \quad \mathbb{E}(||\hat{w}_i - w||^2) \,, \qquad when \qquad w = w^*$$

And consequently, we can estimate the probability by

$$\hat{P} \approx P(y|x)_{w*} = \frac{1}{1 + e^{-w^{*T} x \cdot y}} \quad y \in \{-1, 1\}$$

3.3 Group Details and Individual Contribution

Yash Sanjeev (180070068): Wrote Section 3.1 (Choices for a Hinge Loss Function).
Harshit Shrivastava (18D070011): Wrote Section 3.2 (Probabilistic Approximation of Loss Function).
Kaustav Prasad (200260024): Wrote Section 3.2.1 (Estimating w from Labelled Data)
Ayush Kapoor (200020038): Wrote Section 3.2.2 (Use of w in estimating the $P(y|x)$ for $(x, y) \notin D$)