06/05/23

Machine Learning

Assignment - 2

Ans1. $X = \mathbb{R}^d$ $\quad D = \{(x_i, y_i)\}_{i=1}^{n}$

$y = \{0, 1\}$

- As given in the question, since there are only 2 labels / target variables, the class conditional probabilities would be:

. $P(y=1|x) = \sigma(w^T x) = \frac{1}{1+e^{-w^T x}}$

. $P(y=0|x) = 1 - P(y=1|x)$

$= 1 - \frac{1}{1+e^{-w^T x}}$

- the likelihood for a single observation would be:

. $P(y|x,w) = Ber(y \mid \sigma(w^T x))$

- using the formula $p^k \cdot (1-p)^{(1-k)}$, we get:

. $P(y|x,w) = \sigma(w^T x)^y \times (1 - \sigma(w^T x))^{1-y}$

- likelihood of the training data

. $P(y|X,w) = \prod_{i=1}^{N} \sigma(w^T x_i)^{y_i} \times (1 - \sigma(w^T x_i))^{1-y_i}$

↳ given that our data satisfies the IID constraints

- the log likelihood then looks like:

$$LL = \sum_{i=1}^{N} \{ y_i \ln \sigma(w^T x_i) + (1-y_i) \ln \sigma(w^T x_i) \}$$

- we maximize the likelihood by:

$$w^* = \arg\max_w \sum_{i=1}^{N} \{ y_i \ln \sigma(w^T x_i) + (1-y_i) \ln (1-\sigma(w^T x_i)) \}$$

- Now for simplifying the differentiation, we use the chain rule and consider the following things:

→ $a_i = w^T x$

→ $o_i = \sigma(w^T x)$
$= \sigma(a_i)$

- substituting these values we get the log loss as:

$$LL = \sum_{i=1}^{N} [y_i \log o_i + (1-y_i) \log (1-o_i)]$$

- the chain rule for differentiation would look like:

$$\frac{\partial LL}{\partial w} = \frac{\partial a_i}{\partial w} \cdot \frac{\partial o_i}{\partial a_i} \cdot \frac{\partial a_i}{\partial w} \cdot \frac{\partial LL}{\partial o_i}$$

- calculating each differentiation term:

→ $\frac{\partial a_i}{\partial w} = \frac{\partial}{\partial w} (w^T x_i) = x_i$ — ①

→ $\frac{\partial o_i}{\partial a_i} = \frac{\partial}{\partial a_i} \left(\frac{1}{1+e^{-a_i}} \right)$

$= (-1) \left(1+e^{-a_i}\right)^{-2} \cdot e^{-a_i} (-1)$

$$\frac{\partial o_i}{\partial a_i} = \frac{e^{-a_i}}{1+e^{-a_i}} \cdot \frac{1}{1+e^{-a_i}}$$

$$\frac{\partial o_i}{\partial a_i} = \left(1 - \frac{1}{1+e^{-a_i}}\right) \frac{1}{1+e^{-a_i}}$$

$$\frac{\partial o_i}{\partial a_i} = (1-o_i)\, o_i \quad \text{— (2)}$$

$$\rightarrow LL = \sum_{i=1}^{N} \{ y_i \ln o_i + (1-y_i) \ln (1-o_i) \}$$

- differentiating over w

$$\frac{\partial LL}{\partial w} = \sum_{i=1}^{N} \{ y_i \frac{1}{o_i} \cdot \frac{\partial o_i}{\partial a_i} \cdot \frac{\partial a_i}{\partial w} + (1-y_i) \frac{1}{1-o_i} (-1) \cdot \frac{\partial o_i}{\partial a_i} \cdot \frac{\partial a_i}{\partial w} \}$$

- substituting the differentiation values from eq ① and ②

$$\frac{\partial LL}{\partial w} = \sum_{i=1}^{N} \{ y_i \cdot \frac{1}{o_i} \cdot (1-o_i) o_i \cdot x_i + (1-y_i) \frac{-1}{(1-o_i)} (1-o_i) o_i \cdot x_i \}$$

$$\frac{\partial LL}{\partial w} = \sum_{i=1}^{N} \{ y_i (1-o_i) x_i - (1-y_i) o_i x_i \}$$

$$\frac{\partial LL}{\partial w} = \sum_{i=1}^{N} \{ y_i x_i - x_i y_i o_i - o_i x_i + x_i y_i o_i \}$$

$$\frac{\partial LL}{\partial w} = \sum_{i=1}^{N} (y_i - o_i) x_i$$

$$\nabla_w LL = g(w) \rightarrow \text{closed solution}$$

Hence, the loss function of logistic regression would look like:

$$E(w) = -\sum_{i=1}^{N} \{ y_i \ln \sigma(w^T x) + (1-y_i) \ln (1-\sigma(w^T x)) \}$$

this is the negative log loss function which the model has to minimize to get the best decision boundary.

Ans 2.

$$\sigma(a) = \frac{1}{1+e^{-a}}$$

Given:

1. $a = w^T x$

$w, x \in R^m$

$\frac{\delta \sigma(a)}{\delta w} = ?$

- To find $\frac{\delta \sigma(a)}{\delta w}$, we would have to use the chain rule of differentiation.

- Hence, $\frac{\delta(\sigma(a))}{\delta w} = \frac{\delta \sigma(a)}{\delta a} \cdot \frac{\delta a}{\delta w}$

- Calculating each differentiation term:

$$\rightarrow \frac{\delta \sigma(a)}{\delta a} = \frac{\delta}{\delta a} (1+e^{-a})^{-1}$$

$$= (-1) \cdot (1+e^{-a})^{-2} \cdot e^{-a} \cdot (-1)$$

$$= e^{-a} \cdot (1+e^{-a})^{-2}$$

$$\frac{\delta \sigma(a)}{\delta a} = \frac{e^{-a}}{(1+e^{-a})^2}$$

$$\frac{\delta \sigma(a)}{\delta a} = \frac{e^{-a}}{(1+e^{-a})} \cdot \frac{1}{(1+e^{-a})} \quad \text{— (1)}$$

$$\rightarrow \frac{\delta a}{\delta w} = \frac{\delta}{\delta w}(w^T x)$$

$$= x \quad \text{— (2)}$$

$$\therefore \frac{\delta \sigma(a)}{\delta w} = \frac{\delta \sigma(a)}{\delta a} \cdot \frac{\delta a}{\delta w}$$

- using equations (1) and (2)

$$\frac{\delta \sigma(a)}{\delta w} = \frac{e^{-a}}{(1+e^{-a})^2} \cdot x$$

- now substituting the value of $a = w^T x$

$$\frac{\delta \sigma(a)}{\delta w} = \frac{e^{-w^T x}}{(1+e^{-w^T x})^2} \cdot x$$

2. target variable $y \in \{0, 1\}$

$P(y|x, w) = ?$

- The equation we use to represent the posterior of y with given x and w for logistic regression would be:

$$P(y|x, w) = \prod_{i=1}^{N} \sigma(w^T x_i)^{y_i} \times \left(1 - \sigma(w^T x_i)\right)^{1-y_i}$$

- using this formula we can get the equation for both the classes logistic regression tries to seperate the data into:

→ for $y = 1$

$$P(y=1|x, w) = \prod_{i=1}^{N} \sigma(w^T x_i)^{0} \times \left(1 - \sigma(w^T x_i)\right)^{1-0}$$

$$P(y=1|x, w) = \prod_{i=1}^{N} 1 \cdot \left(1 - \sigma(w^T x_i)\right)^{1}$$

$$P(y=1|x, w) = \prod_{i=1}^{N} \left(1 - \sigma(w^T x)\right)$$

→ for $y = 0$

$$P(y=0|x, w) = \prod_{i=1}^{N} \sigma(w^T x_i)^{1} \times \left(1 - \sigma(w^T x_i)\right)^{1-1}$$

$$P(y=0|x, w) = \prod_{i=1}^{N} \sigma(w^T x_i) \times \left(1 - \sigma(w^T x_i)\right)^{0}$$

$$P(y=0|x, w) = \prod_{i=1}^{N} \sigma(w^T x_i)$$

3. The loss function for logistic regression is:

→ the negative log loss function

$$E(w) = -\sum_{i=1}^{N} \{ y_i \ln \sigma(w^T x_i) + (1-y_i) \ln (1 - \sigma(w^T x_i)) \}$$

- this is also referred to as the cross entropy loss function. logistic regression will try to minimize this $E(w)$ value for getting the best decision boundary.

- The value of w is learnt iteratively using gradient descent. Here we initialize $w = [0, 0 \ldots 0]^T$. Till we do not reach convergence we use the following formula to update w value.

$$w^k \leftarrow w^{k-1} \times \eta \nabla E(w)$$

$$w^k \leftarrow w^{k-1} \times \eta X^T (O - Y)$$

- convergence criteria is:
  1. reduction in error between successive iteration is below $0 < \varepsilon << 1$
  2. max no. of iterations.

Ans 3. Submitted as a pdf and ipynb file with the required implementation.

Ans 4.

Gaussian Discriminant Analysis (GDA)	Logistic Regression
• GDA is a generative model that basically means that they build a model of what each of the classes looks like and at test time it evaluates the model against the two types - the one which matches closely. This means that they learn the $P(x_i \mid y_i)$ - features and then classify which class the new data point would belong to.	• Logistic regression is a discriminative model which basically means that they try to have the best fit decision boundary to classify the two classes. These models consider the $P(y \mid x)$ value for getting the best decision boundary for classification
• GDA makes stronger assumptions on the data (i.e, it is distributed - gaussian). If the assumption is correct it works very well especially on smaller data. It might work badly on bigger data where the assumptions are wrong.	• Logistic regression makes weaker assumption hence might work better with data with Poission distribution as compared to GDA. This also might work on bigger data.