ECE 469/ECE 568 Machine Learning

Textbook:
Machine Learning: a Probabilistic Perspective by Kevin Patrick Murphy

Southern Illinois University

October 1, 2024

Logistic regression for classification

- Consider a two-class classification problem.
- Then, the posterior probability of $\mathbf{C}_1$ can be represented as a logistic sigmoid acting on a linear function of the feature vector $\boldsymbol{x}$:

$$p(\mathcal{C}_1|\boldsymbol{x}) = \sigma(\mathbf{w}^T\boldsymbol{x})$$

- Since there are only two cases, we have

$$p(\mathcal{C}_2|\boldsymbol{x}) = 1 - p(\mathcal{C}_1|\boldsymbol{x}) = 1 - \sigma(\mathbf{w}^T\boldsymbol{x}),$$

  where $\sigma(\cdot)$ is the logistic sigmoid function, and hence, the name "logistic regression".

- But please do not be confused with regression/curve-fitting that we discussed during last several lectures. Logistic regression is actually used for classification in ML.

Logistic regression

Interpretation of the Model / Hypothesis function

15

09/29/2021

$g(\underline{x}, \underline{\omega})$ = estimated posteriori probability that $y=1$ on input $\underline{x}$

Ex : If $\underline{x} = \begin{bmatrix} x_0 \\ x_1 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{Tumor size} \end{bmatrix}$

Let $g(\underline{x}, \underline{\omega}) = 0.7$

$\Rightarrow$ 70% chance of tumor being malignant (cancer)

Let $g(\underline{x}, \underline{\omega}) = P(Y=1 \mid \underline{x}, \underline{\omega})$

$\Rightarrow$ Probability that $Y=1$ given $\underline{x}$ and parameter vector $\underline{\omega}$.

For binary classification , $y \in \{0, 1\}$.

$$P(Y=0 \mid \underline{x}, \underline{\omega}) + P(Y=1 \mid \underline{x}, \underline{\omega}) = 1$$

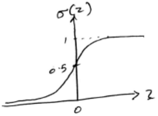
$$P(Y=0 \mid \underline{x}, \underline{\omega}) = 1 - P(Y=1 \mid \underline{x}, \underline{\omega}) = 1 - \sigma(\underline{\omega}^T \underline{x})$$

Logistic regression

Decision boundaries for logistic regression.

Logistic sigmoid function $= g(\underline{w}, \underline{x}) = \sigma(\underline{w}^T \underline{x})$

where $\sigma(z) = \dfrac{1}{1 + e^{-z}}$



Let $g(\underline{x}, \underline{w}) = \sigma(\underline{w}^T \underline{x}) = P(Y=1 \mid \underline{x}, \underline{w})$

$$y = \begin{cases} 1 & \text{if } g(\underline{x}, \underline{w}) \geq 0.5 \\ 0 & \text{if } g(\underline{x}, \underline{w}) < 0.5 \end{cases}$$

$\sigma(z) \geq 0.5$ if $z \geq 0$

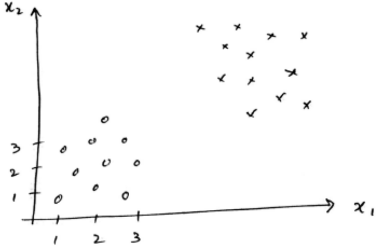
$\sigma(\underline{w}^T \underline{x}) \geq 0.5$ if $\underline{w}^T \underline{x} \geq 0$

$\sigma(z) < 0.5$ if $z < 0$

$\sigma(\underline{w}^T \underline{x}) < 0.5$ if $\underline{w}^T \underline{x} < 0$

Decision boundaries for two input features

09/24



$x \Rightarrow y = 1$

$0 \Rightarrow y = 0$

Let $g(\underline{\omega}, \underline{x}) = \sigma(\omega_0 + \omega_1 x_1 + \omega_2 x_2)$

Our goal is to train $\underline{\omega} = [\omega_0, \omega_1, \omega_2]$ to minimize a cost function.

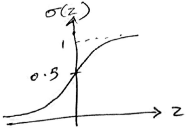
Assume $\underline{\omega} = [-3, 1, 1]$

$$g(\underline{\omega}, \underline{x}) = \sigma(-3 + x_1 + x_2)$$

We predict $Y=1$ if $P(Y=1 \mid \underline{x}, \underline{w}) \geq 0.5$

We interpret $P(Y=1 \mid x, w)$ by $\sigma(-3+x_1+x_2)$

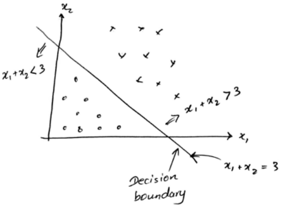
If $\sigma(-3+x_1+x_2) \geq 0.5$, then $Y=1$



$$\Rightarrow \quad -3+x_1+x_2 \geq 0$$

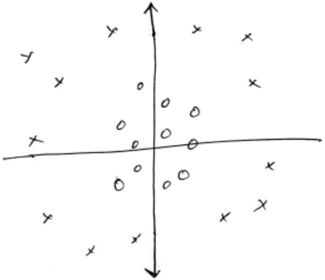
$$x_1+x_2 \geq 3$$

- Note that $\sigma(z) \geq 0.5$ for all $z \geq 0$, and $\sigma(z) < 0.5$ for all $z < 0$.
- Thus, the above decision rule produces a linear decision boundary.



Logistic regression

<u>Non-Linear decision boundary</u>.



Let $g(\underline{\omega}, \underline{x}) = \sigma(\omega_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_1^2 + \omega_4 x_2^2)$

Assume $\underline{\omega} = [\underset{\uparrow \atop \omega_0}{-1}, \underset{\uparrow \atop \omega_1}{0}, \underset{\uparrow \atop \omega_2}{0}, \underset{\uparrow \atop \omega_3}{1}, \underset{\uparrow \atop \omega_4}{1}]^T$

$g(\underline{\omega}, \underline{x}) = \sigma(-1 + x_1^2 + x_2^2)$

Logistic regression

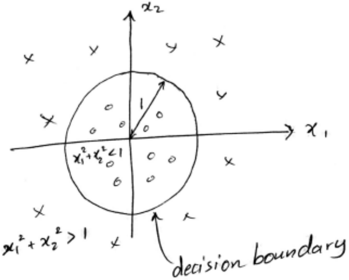
We predict $y=1$ if $P(y=1|\underset{\sim}{x},\underset{\sim}{\omega}) \geq 0.5$

$$\Rightarrow \sigma(-1 + x_1^2 + x_2^2) \geq 0.5$$

$$\Rightarrow -1 + x_1^2 + x_2^2 \geq 0$$

$$x_1^2 + x_2^2 \geq 1$$

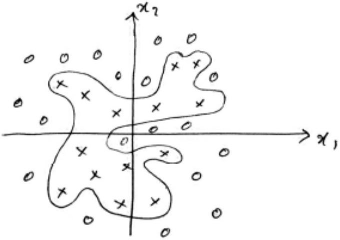
$x_1^2 + x_2^2 = 1$ is a circle with a radius equals to 1.



Logistic regression

More Complicated Hypothesis Functions

$$g(\underline{\omega}, \underline{x}) = \sigma\left(\omega_0 + \omega_1 x_1 + \omega_2 x_2 + \omega_3 x_1^2 + \omega_4 x_2^2 + \omega_5 x_1^2 x_2 + \omega_6 x_1 x_2^2\right)$$



when the model is more complicated, the algorithm may suffer from overfitting.

$\Rightarrow$ Regularization can help in this case.

Logistic regression

Cost Function for Logistic Regression.

A cost function is needed to train weights.

Optimal weights can be found by minimizing the cost function.

Let the training set be

$$\{(\underline{x}^{(1)}, y^{(1)}), \ldots, (\underline{x}^{(i)}, y^{(i)}), \ldots, (\underline{x}^{(m)}, y^{(m)})\}$$

There are m elements = size of the training set.

Let $x_0 = 1$, $y \in \{0, 1\}$ for logistic regression with two classes.

$$\underset{\substack{\uparrow \\ \text{feature vector}}}{\underline{x}} = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^{n+1}$$

Logistic regression

Our hypothesis function or model is

$$g(\underline{w}, \underline{x}) = \sigma\,(\underbrace{\underline{w}^T \underline{x}}_{\text{linear argument function}})$$

logistic sigmoid function

How to find $\underline{w}$?

Recall that for linear regression, we found $\underline{w}$ by minimizing the sum squared error cost function.

Logistic regression

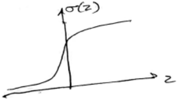
$$J(\underline{\omega}) = \frac{1}{2} \sum_{i=1}^{m} \underbrace{\left(g(\underline{\omega}, \underline{x}^{(i)}) - y^{(i)} \right)^2}_{cost\left(g(\underline{\omega}, \underline{x}^{(i)}),\ y^{(i)}\right)}$$

09

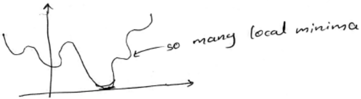
For linear regression, $g(\underline{\omega}, \underline{x})$ was a linear function.
So the sum-squared error cost function was convex.
$\Rightarrow$ we can find a global minima.

But for logistic regression, $g(\underline{\omega}, \underline{x}) = \sigma(\underline{\omega}^T \underline{x})$

This is a non-linear function



Sum-squared error for logistic regression with sigmoid function will not be convex.
$\Rightarrow$ cannot find a global minima.



Logistic regression

Cost Function for Logistic Regression

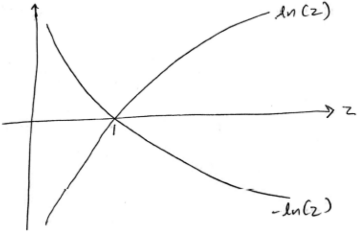
$$\text{cost}(g(\underline{\omega}, \underline{x}), y) = \begin{cases} -\ln\left[g(\underline{\omega}, \underline{x})\right] & \text{for } Y=1 \\ -\ln\left[1 - g(\underline{\omega}, \underline{x})\right] & \text{for } Y=0 \end{cases}$$

We can obtain this cost function by using maximum likelihood estimation.

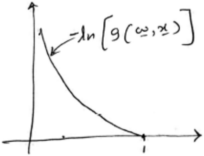
$\Rightarrow$ Yet to be described for logistic regression.

Logistic regression

Interpretation of this cost function:



If $Y = 1$



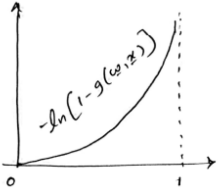
$0 \leq g(\underline{\omega}, \underline{x}) \leq 1$

If $Y = 1$ and $g(\underline{\omega}, \underline{x}) = 1$,
then cost $= 0$.

But as $g(\underline{\omega}, \underline{x}) \to 0$, cost $\to \infty$

So if we incorrectly predict the output variable to be "0" instead of "1", we get the maximum penalty from our cost function. (19) 09/29/2021

If $Y = 0$



$0 \le g(\underline{w}, \underline{x}) \le 1$

~~when than then~~

when $g(\underline{w}, \underline{x}) = 0$ and $Y = 0$,
then cost $= 0$.

If Y is incorrectly predicted to be 1, then we got maximum penalty

Logistic regression

We can show that this cost function is also convex.

$\Rightarrow$ We have a global minima.

Note that $Y = 1$ or $Y = 0$ always.

$$\text{cost}\left(g(\underline{\omega}, \underline{x}), y\right) = -y \ln\left[g(\underline{\omega}, \underline{x})\right] - (1-y)\ln\left[1 - g(\underline{\omega}, \underline{x})\right]$$

Thus, our $J(\underline{\omega})$ becomes

$$J(\underline{\omega}) = \frac{1}{m}\sum_{i=1}^{m} \text{cost}\left(g(\underline{\omega}, \underline{x}^{(i)}, y^{(i)})\right)$$

$$J(\underline{\omega}) = \frac{-1}{m}\sum_{i=1}^{m}\left(y^{(i)}\ln\left[g(\underline{\omega}, \underline{x}^{(i)})\right] + (1-y^{(i)})\ln\left[1 - g(\underline{\omega}, \underline{x}^{(i)})\right]\right)$$

Logistic regression

To find $\underline{\omega}$,

$$\min_{\underline{\omega}} J(\underline{\omega}) \rightarrow \hat{\underline{\omega}} \leftarrow \text{estimate for } \underline{\omega}$$

Note that for logistic regression,

$$g(\underline{\omega}, \underline{x}) = \frac{1}{1 + e^{-\underline{\omega}^T \underline{x}}} = \sigma(\underline{\omega}^T \underline{x})$$

$\Rightarrow$ Use gradient descent algorithm and find $\hat{\underline{\omega}}$ numerically.