Machine Learning 2018 – Logistic Regression

Kien C Nguyen

November 28, 2018

Contents

Binary Classification

- Recall that in supervised learning, if the targets (labels) are categorical, the problem is called classification.
- Classify an email as Not Spam / Spam
- In credit scoring, classify a customer as Good / Bad
- In network intrusion detection, classify a connection as Normal / Attack
- Detect the gender (Male / Female) using profile pictures

- Recall that in linear regression, $\hat{y} = \mathbf{w}^T \mathbf{x}$.
- This model can only be used if y is not upper-bounded and not lower-bounded.
- In logistic regression, we predict the probability of a Positive Class (vs a Negative Class).
- E.g. Probability that an email is Spam, probability that a customer is a Bad customer.

Probability of passing an exam versus hours of study

- A group of 20 students spend between 0 and 6 hours studying for an exam. How does the number of hours spent studying affect the probability that the student will pass the exam?
- We predict the probability that a student passes the exam ($y = 1$) using the number of hours that student spent.

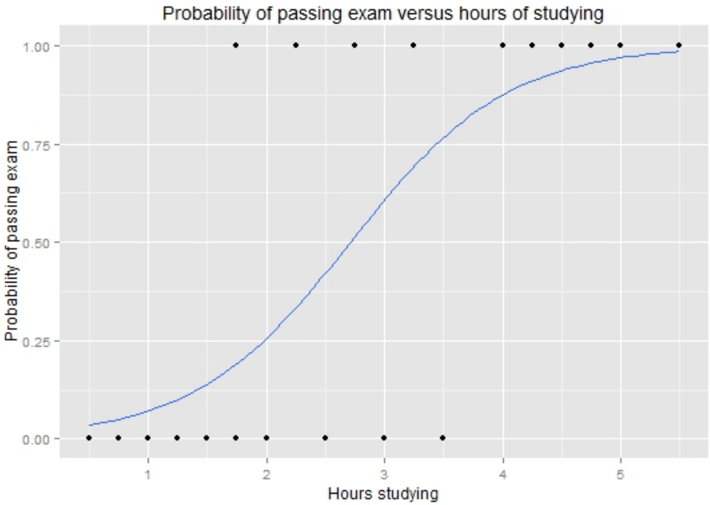
Source: `https://en.wikipedia.org/wiki/Logistic_regression`

Probability of passing an exam versus hours of study

Hours	Pass	Hours	Pass
.5	0	2.75	1
.75	0	3	0
1	0	3.25	1
1.25	0	3.5	0
1.5	0	4	1
1.75	0	4.25	1
1.75	1	4.5	1
2	0	4.75	1
2.25	1	5	1
2.5	0	5.5	1

Source: `https://machinelearningcoban.com/2017/01/27/logisticregression/`

Probability of passing an exam versus hours of study

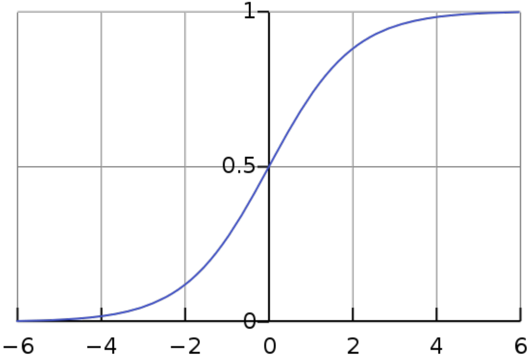


Source: `https://machinelearningcoban.com/2017/01/27/logisticregression/`

Contents

- Use a function $\Phi(\mathbf{w}^T\mathbf{x})$
- As this is a probability, we want $0 \leq \Phi(\mathbf{w}^T\mathbf{x}) \leq 1$
- Sigmoid function (Logistic function)

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

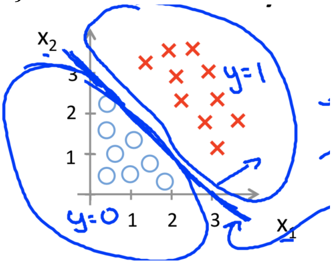


Source: `https://en.wikipedia.org/wiki/Logistic_regression`

Suppose we predict $y = 1$ if $P\{y = 1\} \geq 0.5$.

$$\sigma(z) \geq 0 \iff \mathbf{w}^T\mathbf{x} \geq 0$$

Predict $y = 0$ if $P\{y = 1\} < 0.5 \iff \mathbf{w}^T\mathbf{x} < 0$



Source: Andrew Ng – Machine Learning (Coursera)

Contents

Recall that the training set is $\left(\left(\mathbf{x}^{(1)}, y^{(1)}\right), \left(\mathbf{x}^{(2)}, y^{(2)}\right), \ldots \left(\mathbf{x}^{(N)}, y^{(N)}\right)\right)$.

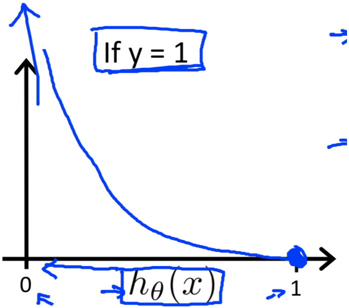
where $\mathbf{x}^{(i)}$ is given by $\mathbf{x}^{(i)} = \begin{bmatrix} x_0^{(i)} \\ x_1^{(i)} \\ \ldots \\ x_D^{(i)} \end{bmatrix}$

$x_0^{(i)} = 1,\ y^{(i)} \in \{0, 1\}$

Loss for each training example

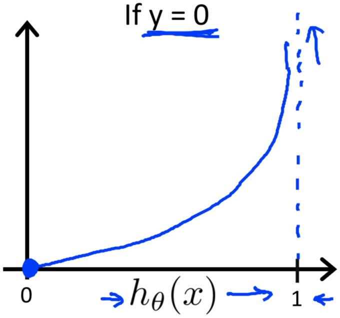
$$
\begin{aligned}
L(\hat{y}, y) &= \begin{cases} -log(\Phi(\mathbf{w}^T\mathbf{x})) & if \quad y = 1 \\ -log(1 - \Phi(\mathbf{w}^T\mathbf{x})) & if \quad y = 0 \end{cases} \\
&= -ylog(\Phi(\mathbf{w}^T\mathbf{x})) - (1 - y)log(1 - \Phi(\mathbf{w}^T\mathbf{x}))
\end{aligned}
$$

Loss for each training example: $y = 1$



$L = 0$ when $\Phi(\mathbf{w}^T\mathbf{x}) = 1$
$L \to \infty$ as $\Phi(\mathbf{w}^T\mathbf{x}) \to 0$
Source: Andrew Ng – Machine Learning (Coursera)

Loss for each training example: $y = 0$



$L = 0$ when $\Phi(\mathbf{w}^T\mathbf{x}) = 0$
$L \to \infty$ as $\Phi(\mathbf{w}^T\mathbf{x}) \to 1$
Source: Andrew Ng – Machine Learning (Coursera)

Loss function

$$
\begin{aligned}
L(\mathbf{w}) &= \frac{1}{N}\sum_{i=1}^{N} -y^{(i)} log(\Phi(\mathbf{w}^T\mathbf{x}^{(i)})) - (1-y^{(i)}) log(1-\Phi(\mathbf{w}^T\mathbf{x}^{(i)})) \\
&= -\frac{1}{N}\sum_{i=1}^{N} y^{(i)} log(\Phi(\mathbf{w}^T\mathbf{x}^{(i)})) + (1-y^{(i)}) log(1-\Phi(\mathbf{w}^T\mathbf{x}^{(i)}))
\end{aligned}
$$

We find the $\hat{\mathbf{w}}$ such that

$$\hat{\mathbf{w}} = argmin_{\mathbf{w}} L(\mathbf{w}) \tag{1}$$

Once we have $\hat{\mathbf{w}}$, the prediction for a new $\mathbf{x}$ is

$$P\{\hat{y} = 1\} = \Phi(\mathbf{w}^T\mathbf{x}) \tag{2}$$

Contents

Initialize $\mathbf{w} = [0, ..., 0]$;
Repeat $\mathbf{w} \leftarrow \mathbf{w} - \eta \nabla_{\mathbf{w}} L(\mathbf{w})$

- η: step size
- $\nabla_{\mathbf{w}} L(\mathbf{w})$: gradient

The update for each w_j:

$$w_j = w_j - \eta \sum_{i=1}^{N} \left(\Phi(\mathbf{w}^T \mathbf{x}^{(i)}) - y^{(i)} \right) x_j^{(i)} \tag{3}$$

Contents

References

[1] Bishop, C. M. (2013). Pattern Recognition and Machine Learning. Journal of Chemical Information and Modeling (Vol. 53).
[2] Wikipedia – Logistic Regression – https://en.wikipedia.org/wiki/Logistic_regression
[3] Vu Huu Tiep – Machine Learning Co Ban – https://machinelearningcoban.com/2017/01/27/logisticregression/
[4] Andrew Ng – Machine Learning (Coursera) – https://www.coursera.org/learn/machine-learning