LOGISTIC REGRESSION

Questions

1) **Why do we need it?**
2) **Assumptions?**
3) **How does it work?**
4) **How does it make predictions?**

OVERVIEW

A classification algorithm which predicts output as 2 numbers for given input X.

Some notation
X being for entire discussion a matrix of all input with dimension m*n (m= number of input),
(n=number of features)
X(i) denoting ith test input (dim=1*n)
y(i) denoting output for ith test input(dim=m*1)

Like any machine learning algorithm the model requires :-

a) h(X): A hypothesis/function which finally gives output for a required input of matrix X.
b) J(Θ): A cost function to evaluate the error
c) Gradient descent to update the parameters

WHY DO WE NEED IT?

It's one of the approaches for binary classification as linear regression can give output till any range so we need to squeeze the data to predict 0 or 1.

Interpretability: Logistic regression provides interpretable results, as it estimates the probability of an instance belonging to a particular class. The coefficients associated with each feature can be interpreted as the influence of that feature on the likelihood of the positive class. This can be valuable in understanding the impact of different variables on the outcome.
Linear decision boundary: Logistic regression assumes a linear relationship between the input features and the log-odds of the positive class. This linear decision boundary makes logistic regression computationally efficient and allows for straightforward interpretation. However, as mentioned earlier, you can introduce non-linearity by using non-linear transformations or by combining logistic regression with other techniques.
Robustness to noise: Logistic regression is known to be robust to noise and outliers, as it estimates probabilities rather than relying solely on point predictions. Outliers and noise can affect the estimated coefficients, but the overall model performance in terms of predicting probabilities may still be reliable.

ASSUMPTIONS

1)Logistic model came from considering that the best possible interpretation of binary data is Bernoulli's equation. We are assuming that at each point x(i) in the data set y(i) is given by the Bernoulli probability curve.So in the GLM model we take is the Bernoulli equation.

2)Independence of observations: Logistic regression assumes that the observations are independent of each other. This means that the probability of an instance belonging to a particular class should be independent of the other instances.

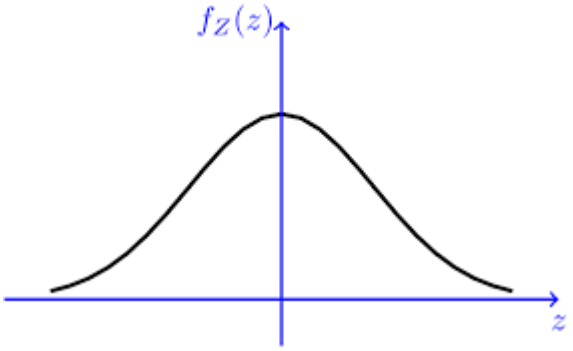
SOME CONCEPTS BEFORE GOING INTO LOGISTIC REGRESSION

Probabilistic interpretation

Key Assumption

a) y(i)=(Θ.T)*(X(i))+e(i) ,where e(i) is the noise of random error we have in data.
b) p(e(i))=N(0,6^2), 6: the bandwidth of error we are choosing
c) all the error are independent to one another

Normal distribution



$$p(\epsilon^{(i)}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\epsilon^{(i)})^2}{2\sigma^2}\right).$$

Substituting the e(i) from (a) we get

$$p(y^{(i)}|x^{(i)};\theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y^{(i)}-\theta^T x^{(i)})^2}{2\sigma^2}\right).$$

The above probability signifies the probability of y(i) given x(i) and parameterized Θ.
Now we define the likelihood function.

$$L(\theta) = L(\theta; X, \vec{y}) = p(\vec{y}|X;\theta).$$

The meaning of above being the likelihood of a given parameter where data is fixed.

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)};\theta) \\ &= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y^{(i)}-\theta^T x^{(i)})^2}{2\sigma^2}\right). \end{aligned}$$

Hence try to make data better fit and we increase this likelihood in general.
But increasing log(L(Θ) would be easier in general so we define

$$\ell(\theta) = \log L(\theta)$$

And as ‘6’ is constant for example we can expand the term to

$$m\log\frac{1}{\sqrt{2\pi}\sigma} - \frac{1}{\sigma^2}\cdot\frac{1}{2}\sum_{i=1}^{m}(y^{(i)}-\theta^T x^{(i)})^2$$

Here we see in order to maximise log(L(Θ) we try to minimise the second term which in turn is similar to linear regression.

Generalised Linear Models

To build a model we use a GML to be a h(x) it has some key assumption in itself:

1) (y | x; θ) ~ ExponentialFamily(η). I.e., given x and θ, the distribution of y follows some exponential family distribution, with parameter η.
2) The natural parameter η and the inputs x are related linearly: η = θ T x. (Or, if η is vector-valued, then η(j) = θ .T (j) x)
3) h(x) = E[y|x]

So when we choose bernoulli model we get

$$\begin{aligned} h_\theta(x) &= E[y|x;\theta] \\ &= \phi \\ &= 1/(1+e^{-\eta}) \\ &= 1/(1+e^{-\theta^T x}) \end{aligned}$$

The above derivation came from directly comparing the Bernoulli equation with the standard Exponential Family equation.

$$f_Z(z) = \begin{cases} p^z(1-p)^{1-z}, & z \in \{0,1\} \\ 0, & altramente \end{cases}$$

Bernoulli

$$p(y;\eta) = b(y)\exp(\eta^T T(y) - a(\eta))$$

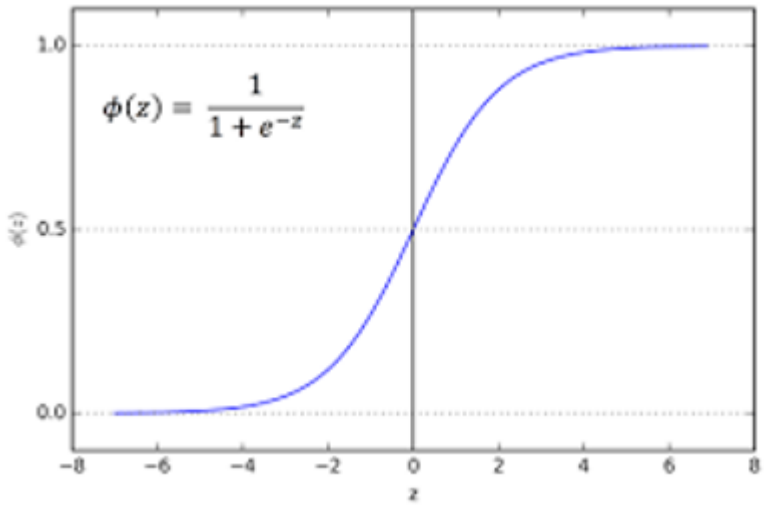
Exponential family

Now as we have got our h(x) we can work on training it.

Logistic Regression

We define

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + e^{-\theta^T x}},$$



Now the Probability equation will be

$$p(y \mid x; \theta) = (h_\theta(x))^y \, (1 - h_\theta(x))^{1-y}$$

Now to define a likelihood parameter over all test case we say

$$L(\theta) \quad = \quad p(\vec{y} \mid X; \theta)$$

Which boils down to

$$= \prod_{i=1}^{m} \left(h_\theta(x^{(i)})\right)^{y^{(i)}} \left(1 - h_\theta(x^{(i)})\right)^{1-y^{(i)}}$$

In our model we want to maximise this L(Θ) i.e log(L(Θ) which we did earlier.

COST FUNCTION

As it turns out the cost function is simply the likelihood function

$$\begin{aligned}\ell(\theta) &= \log L(\theta) \\ &= \sum_{i=1}^{m} y^{(i)} \log h(x^{(i)}) + (1 - y^{(i)}) \log(1 - h(x^{(i)}))\end{aligned}$$

In the future we simply need to maximise this function.
Calculation
1)Derivative of logistic regression

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

$$\begin{aligned}g'(z) &= \frac{d}{dz} \frac{1}{1 + e^{-z}} \\ &= \frac{1}{(1 + e^{-z})^2} (e^{-z}) \\ &= \frac{1}{(1 + e^{-z})} \cdot \left(1 - \frac{1}{(1 + e^{-z})}\right) \\ &= g(z)(1 - g(z)).\end{aligned}$$

We need to keep this in mind while doing calculations later.
2) Derivative of l(Θ) wrt to Θj for each jth coefficient.

$$\begin{aligned}\frac{\partial}{\partial \theta_j} \ell(\theta) &= \left(y \frac{1}{g(\theta^T x)} - (1 - y) \frac{1}{1 - g(\theta^T x)}\right) \frac{\partial}{\partial \theta_j} g(\theta^T x) \\ &= \left(y \frac{1}{g(\theta^T x)} - (1 - y) \frac{1}{1 - g(\theta^T x)}\right) g(\theta^T x)(1 - g(\theta^T x) \frac{\partial}{\partial \theta_j} \theta^T x \\ &= \left(y(1 - g(\theta^T x)) - (1 - y) g(\theta^T x)\right) x_j \\ &= (y - h_\theta(x))\, x_j\end{aligned}$$

Here we used simple forms like derivative of log(x) and chain rule and above calculation.

GRADIENT DESCENT

We update the parameter in various all written in vector form

$$\theta := \theta + \alpha \nabla_\theta \ell(\theta)$$

Which boils down to using our above derivation

Repeat {

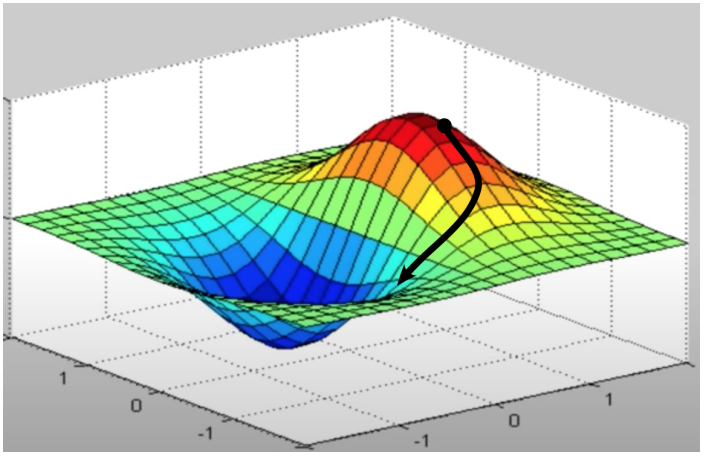
Repeat for each j{

$$\theta_j := \theta_j + \alpha \left(y^{(i)} - h_\theta(x^{(i)}) \right) x_j^{(i)}$$

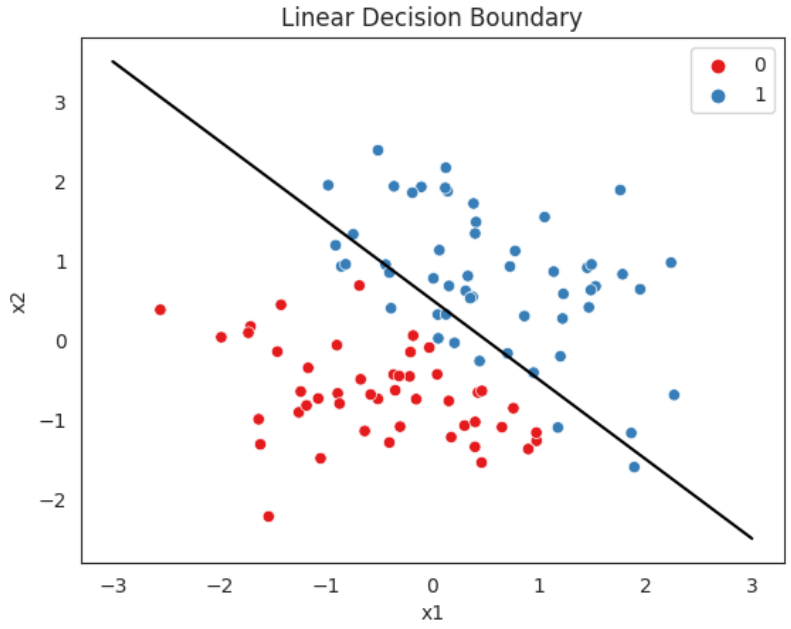
}

}

Following is the example of gradient descent in logistic regression for two parameters(j=2) .



Final result after running model we get function like this



Prediction

When h(x)>0.5 we say it is 1 else 0 for h(x)>0.5 we need $(\Theta.T).X < 0$ for for h(x)>0.5 and so point line away from origin side of line(line with normal vector (Θ) and vica-versa and so we make prediction.

Thank you