Logistic Regression: a few facts.

Facts

- A regression model for categorical endpoints / responses (here: binary) used to analyze the impact (magnitude, direction, inference) of a predictor variables on the probability of occuring an event.

- As every regression, it predicts a numerical outcome : $E(Y|X=x)$

For the Bernoulli's distribution (or binomial with $k=1$) it's a probability of sucess (or log-odds, depending on formulation).

more precisely, it's a direct probability estimator

- Logistic regression itself does NOT produce labels. It's a direct probability estimator. To give classes, it needs additional post-fit activity:

applying a decision rule to the predicted probability.

This makes the logistic classifier.

(Further explained in detail)

Komal Dive

- LR is fit by the maximum likelihood (it's NOT just OLS with logit transformation)
- LR is a part of the Generalized Linear model (like linear, Poisson, gamma, beta, etc).

Logistic regression, invented and further developed in terms of the Generalized Linear model and Generalized Additive Model by Cox, Nelder, Hastie, Tibshirani and others, served exactly to address regression problems with categorical outcomes (via maximum likelihood Estimation)

Komal Dive

Generalized Linear models.

In Regression example, we had $y|x;\theta \sim N(\mu, \sigma^2)$ and in the classification one, $y|x;\theta \sim \text{Bernoulli}(\phi)$ for some appropriate definitions of μ and θ as function of x and θ.

Both of these ~~cases of a~~ methods are special cases of a broader family of models, called Generalized Linear models (GLMs)

Exponential Family.

To work our way up to GLMs, we will begin by defining exponential family distribution.

A class of distribution is said to be in the exponential family if it can be written in the form

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

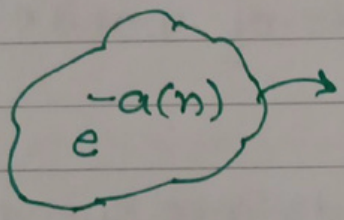
Here,

η = natural parameter, also called the canonical parameter of the distribution

Komal Dive

$T(y)$ = sufficient statistic

$a(\eta)$ = log-partition function



plays the role of a normalization constant, that makes sures the distribution $p(y;\eta)$ sums/integerates over y to 1.

A fixed choice of T, a and b defines a family of distributions that is parameterized by η, we get then different distributions within this family.

Bernoulli Distribution

The Bernoulli distribution with mean(ϕ), Bernoulli(ϕ), speacifies a distribution over $y \in \{0,1\}$, so that

$$\{p(y=1;\phi) = \phi \; ; \; p(y=0;\phi) = 1-\phi\}$$

Komal Dive

as we vary ϕ, we obtain Bernoulli distributions with different means.

Now, we will show that the class of Bernoulli distributions ones obtained by varying ϕ, is in Exponential family.

we write the Bernoulli distribution as

$$p(y;\phi) = \phi^y (1-\phi)^{1-y}$$
$$= \exp\left(y \cdot \log\phi + (1-y)\cdot\log(1-\phi)\right)$$
$$= \exp\left(\left(\log\left(\frac{\phi}{1-\phi}\right)\right) y + \log(1-\phi)\right)$$

Thus,

$$\text{natural parameter}(\eta) = \log\left(\frac{\phi}{1-\phi}\right)$$

$$T(y) = y$$
$$a(\eta) = -\log(1-\phi)$$
$$= \log(1+e^{\eta})$$
$$b(y) = 1$$

This shows that the Bernoulli distribution can be written in the form of Exponential family, using an appropriate choice of T, a and b.

Komal Dive

Similarly, we can also show for Gaussian distribution.

There're many other distributions that are members of the exponential family.

multinominal distribution, poisson, gamma, exponential distribution and many more are the members of exponential family.

Exponential Family $\neq$ Exponential distribution

Constructing GLMs

To drive a GLMs, we will make the following three assumptions about the conditional distribution of y given x and about the model:

1. $y|x;\theta \sim$ Exponential Family(η)

   i.e., given x and θ, the distribution of y follows some exponential family distribution, with parameter η.

Komal Dive

2. Given x, our goal is to predict the expected value of $T(y)$ given x.

Example: $T(y) = y$, so this means we would like the prediction $h(x)$ output by our hypothesis h to satisfy $h(x) \approx E[y|x]$

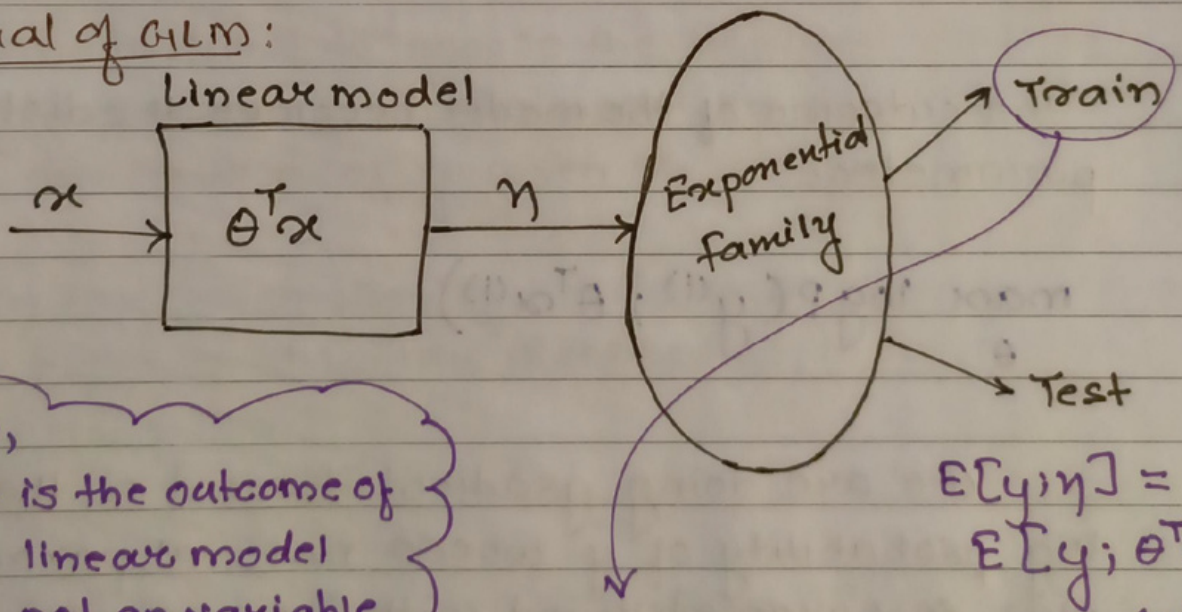
output: $E[y|x;\theta]$

$h_\theta(x) = E[y|x;\theta]$

3. The natural parameter η and the inputs x are related linearly: $\eta = \theta^T x$

(or, if η is vector-valued, then $\eta_i = \theta_i^T x$)

Visual of GLM:



Here,
η = is the outcome of linear model not an variable

$E[y;\eta] = E[y;\theta^T x] = h_\theta(x)$

$\max_\theta \log(p(y^{(i)};\theta^T x^{(i)}))$

(maximum likelihood)

Komal Dive

we are training θ to predict the parameter of the exponential family distribution whose mean is the prediction that we're gonna make for y.

How do we train this model?

So, in this model, the parameter that we are learning by doing gradient descent are $\theta^T x$ parameters.

So we are not learning any parameters in the exponential family, we are learning θ that's part of the model and not part of the distribution.

The outcome of the model becomes the distribution parameter.

$$\max_{\theta} \log P(y^{(i)}; \theta^T x^{(i)})$$

we are doing gradient descent on the log probability of y where natural parameter was re-parameterized with linear model and we are doing gradient descent by taking gradients on the theta.

Komal Dive

GLM training:

learning update rule:

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

This remains same for all distributions.

Logistic Regression

Here, we are interested in binary classification so, $y \in \{0, 1\}$

Given that y is binary-valued, it therefore seems natural to choose the Bernoulli family of distributions to the model the conditional distribution of y given x.

In our formulation of the Bernoulli distribution as an exponential family distribution, we had

$$\phi = \frac{1}{1 + e^{-\eta}}$$

if $y|x;\theta \sim \text{Bernoulli}(\phi)$, then $E[y|x;\theta] = \phi$

Komal Dive

$$h_\theta(x) = E[y|x;\theta]$$

$$= \phi$$

$$= \frac{1}{1+e^{-\eta}}$$

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

So, this gives us hypothesis functions of the form $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$

∵ once we assume that y conditioned on x is Bernoulli, it arises as a consequence of the defn of GLMs and exponential family distributions.

more terminology,

the function g giving the distribution's mean as a function of the natural parameter

$$g(\eta) = E[T(y);\eta]$$

is called the canonical response function.

It's inverse, g^{-1}, is called the canonical link function.

Komal Dive

Thus, the canonical response function for the Bernoulli is the logistic function.

when the $p(y\|x)$ are assumed to be.....	then you have....
Poisson	Poisson regression
Negative binomial	Negative binomial regression
Bernoulli	Logistic regression
Normal	Classical regression
Multinomial	Multinomial logistic regression
Beta	Beta regression
Lognormal	Lognormal regression
...	

The intent is simply to show that different assumed families of conditional distributions $p(y|x)$ give rise to different types of regression models.

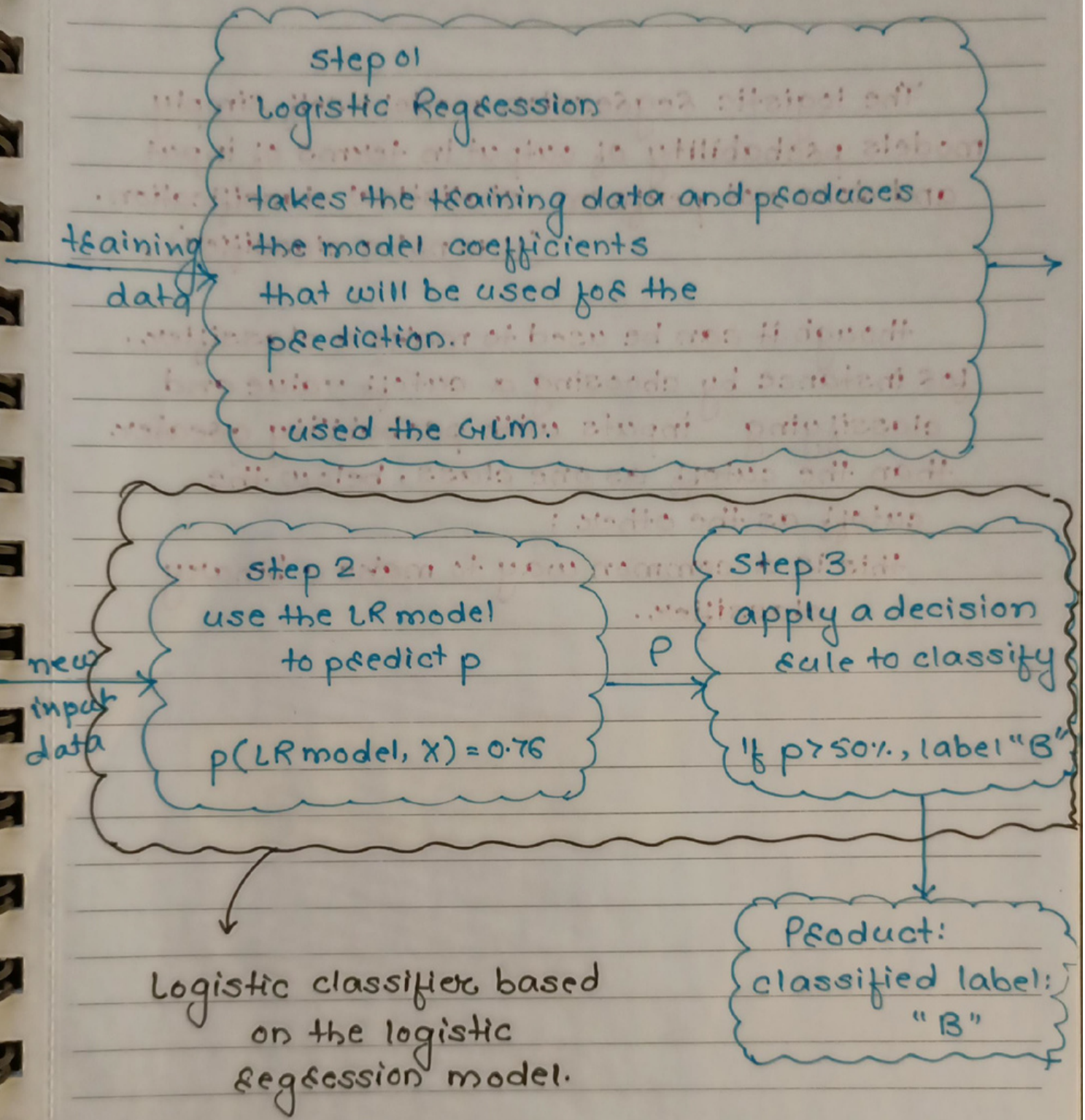
Komal Dive

To conclude,

Logistic regression is a special case of the generalized linear regression model (GLM).

In ML, it is used mainly for classification by taking the predicted probility and applying a decision rule to it.
This makes the logistic classifies.

Illustration →


Step 01
Logistic Regression
takes the training data and produces the model coefficients that will be used for the prediction.
used the GLM.
training data
Step 2
use the LR model to predict p
p(LR model, X) = 0.76
new input data
p
Step 3
apply a decision rule to classify
If p > 50%, label "B"
Logistic classifier based on the logistic regression model.
Product:
classified label: "B"

Komal Dive

The logistic regression model itself simply models probability of output in terms of input and does not perform statistical classification. (it is not a classifier),

though it can be used to make a classifier, for instance by choosing a cutoff value and classifying inputs with probability greater than the cutoff as one class, below the cutoff as the other;
this is a common way to make a binary classifier.

Logistic Regression $\neq$ logistic classifier

Komal Dive

How is it used for regression purposes?

It assesses how ~~many~~ a certain set of independent variables affects the odd-ratio of % of success (depending on formulation) through the assessment of the

- marginal and interaction effects:
  direction, magnitude, inference (p-value, confidence interval)

- simple affects (for categorical covariates):
  magnitude of different/ratio + inference (p, CI)

.. and, when the model is followed by ANOVA-like type analysis, here called the "analysis of deviance', it gives us also the main and their intersection effects,

eg. "does gender does gender and treatment affect the % of successes."
It's done either via likelihood Ratio tests of wald's joint analysis of the model coefficients.

The analysis employs the EM-means (estimated marginal means), which again - is a typical regression outcome.

Komal Dive