a

Lecture 14

- Let's see a psychology research done by Simons & Chabris (1999) at Harvard.

Look at the video and count the passes by white shirts.

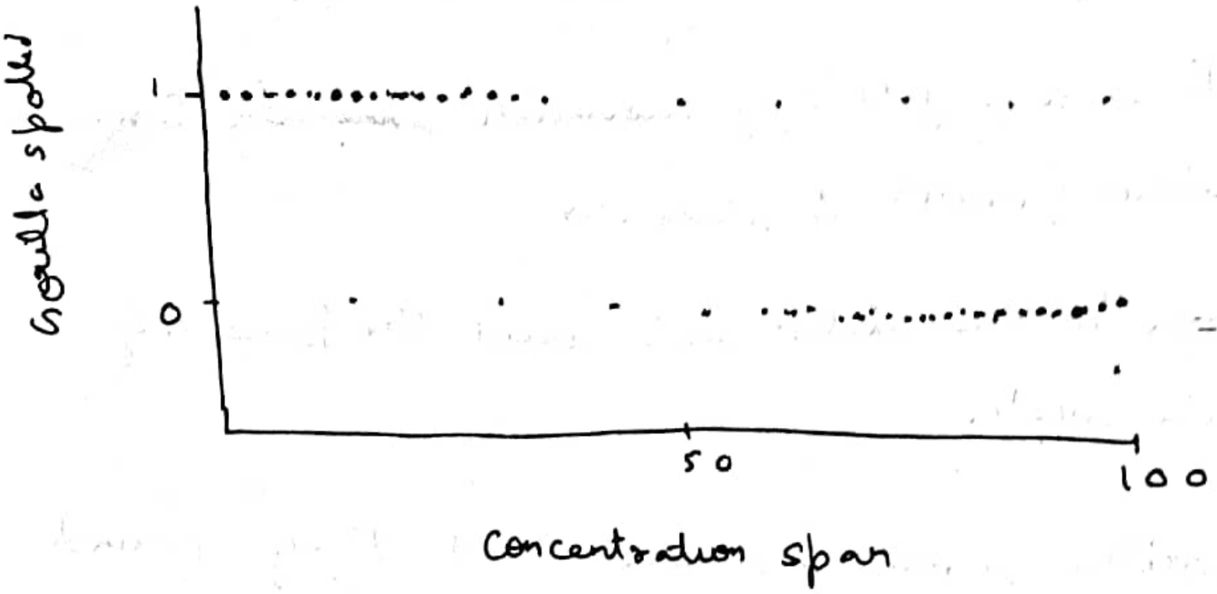
A gorilla comes in between and they found out that the harder the task; more likely people might miss to see the gorilla.

→ Only 50% of his subjects spotted the gorilla.

One have out put

gorilla seen → binary
gorilla not seen ↗

Independent variables
↳ concentration span
→ difficulty of task
→ time of day



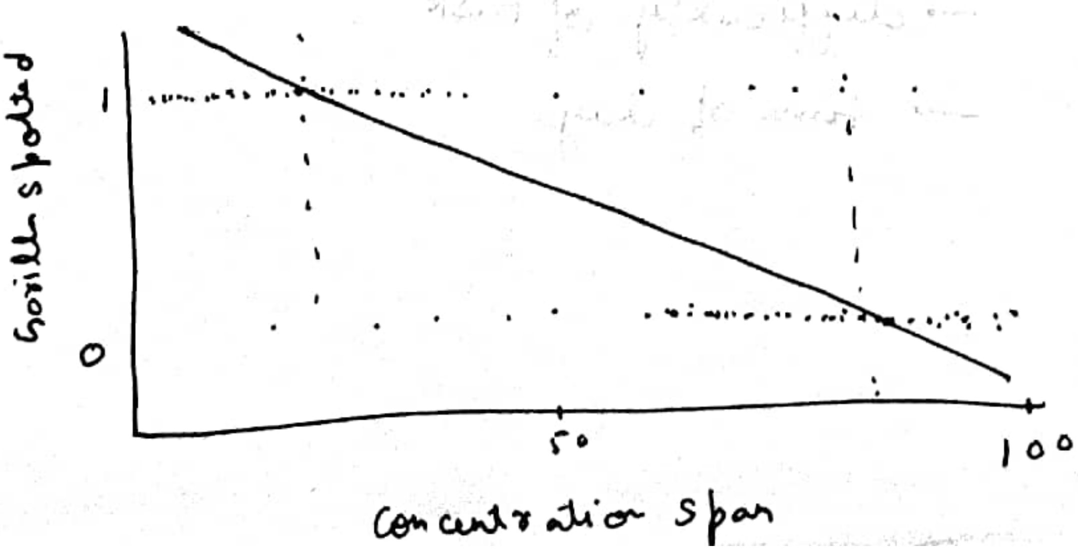
More low concentration people spot gorilla

More high concentration people don't spot the gorilla.

→ Now we want to know is the probability whether a person will see gorilla or not for any Value of concentration span.

→ Last time we used simple linear regression (SLR)

→ SLR may predict values that are below zero or above 1.



L4-~~2~~c If we see people get >1 and <0 with the linear regression line:

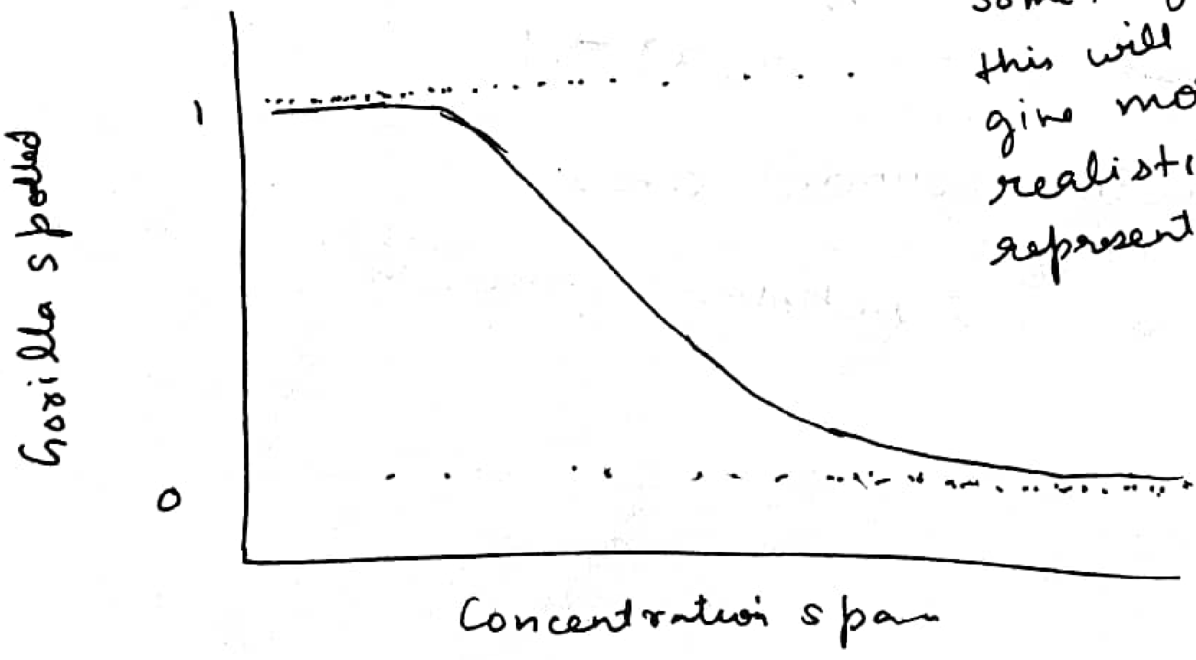
→ As for linear regression we assumed that the population distribution was normally distributed ~~among~~ around the mean, for each value of the X variable.

While in this case due to binary response, distribution around mean is going to be a bit different.

Instead of linear regression; one uses.

Logistic regression.

→ It is widely used.



Something like this will give more realistic representation

In Binary, we have only 2 classification
↳ One need to think in term of probability

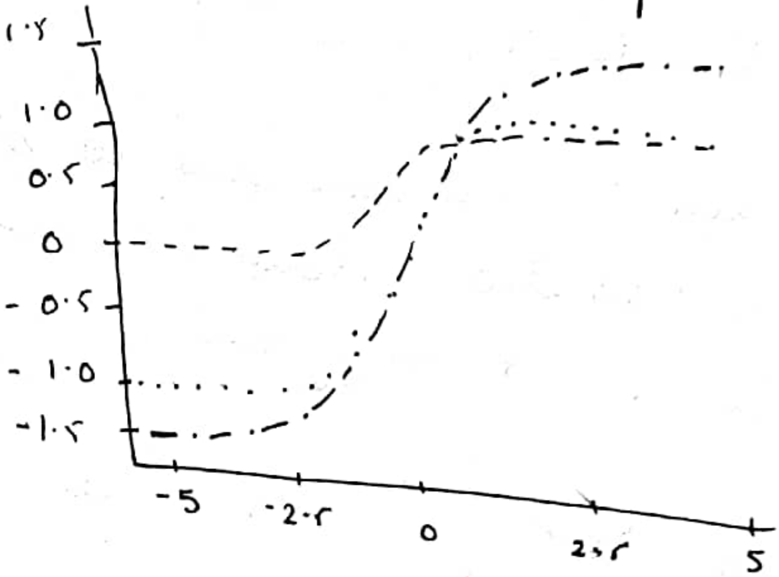
Sigmoid Functions

A sigmoid function is a mathematical function which has a characteristic S-shaped curve.

→ Logistic function ----

→ hyperbolictangent ....

→ Arctangent —.—.—.



$$S(x) = \frac{1}{1+e^{-x}}$$

$x \to -\infty \implies S(x) \Rightarrow 0$

$x \to \infty \implies S(x) \Rightarrow 1$

function bounded 0 to 1

It is a well behaved function.

Lecture 14

Logistic regression

Logistic regression is a supervised machine learning algorithm mainly used for classification tasks where the goal is to predict the probability that an instance of belonging to a given class or not.

→ It is a kind of statistical algorithm. that analyse the relationship between a set of independent variables and the dependent binary variables.

→ It is a powerful tool for decision - making

Some examples

① Predicting if a shopping activity in an ecommerce website is fraudulent or not.

② whether a tumor is benign or malignant.

③ If a customer will buy a product or not.

④ If the satellite will be successful or not.

Logistic regression is basically a supervised classification algorithm.

We have some input features, X

Some Target output variables, Y → can take only discrete values

Just like linear regression assumes that the data follows a linear function

→ Logistic regression models the data using the Sigmoid function

→ Logistic regression becomes a classification technique only when a decision threshold is brought into the picture.

Based on the no. of categories. it is classifies as:

→ Binomial → target variable can have only 2 possible types "0" or "1"

only two categories

Multinomial → Target variables can have 3 or more possible types which are not ordered (i.e. have no quantitative significance) e.g: "Leaf A" or "Leaf B" or "Leaf C"

Let's review Probability basics (definition for now)

① Sample space (Ω)

→ It is the set of all possible outcomes of the experiment

② Event space A

→ Event space is the space of all the possible results of the experiment. The event space is obtained by considering the collection of subsets of Ω.

In case of discrete prob. distribution $P(\Omega)$

$\Omega = \{H, T\}$

$P(H) = \frac{1}{2}$

$P(T) = \frac{1}{2}$

③ Probability

We associate $P(A)$ which measures the probabil that the event will occur. The no. $P(A)$ is called probability of A.

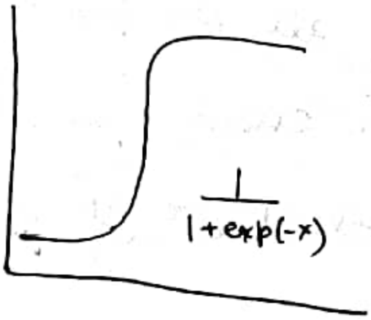
$$P(A) \in [0,1]$$

$$\sum_i P(A_i) = 1$$

$$\Omega = \{H, T\}$$

$$P(H) = \frac{1}{2} \qquad P(T) = \frac{1}{2}$$

$$P(H) + P(T) = 1$$



Sigmoid function has values very close to either 0 or 1 across most of its domain. This makes it suitable for application in classification methods

1a Logistic Regression

instead of $Y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \dots + \alpha_p x_p = \alpha^T x$.

one need to model the probability such that Y is equal to class 1, for given X

$$p(X) = P(Y=1, \text{for given } X)$$

$$p(X) = \frac{1}{1+e^{-\alpha^T x}} = \frac{e^{\alpha^T x}}{1+e^{\alpha^T x}}$$

using sigmoid

$$1 - p(X) = 1 - \frac{e^{\alpha^T x}}{1+e^{\alpha^T x}} = \frac{1}{1+e^{\alpha^T x}}$$

$$\frac{p(x)}{1-p(x)} = e^{\alpha^T x}$$

$$\Rightarrow \log\left(\frac{p(x)}{1-p(x)}\right) = \alpha^T x = \alpha_0 + \alpha_1 x_1 + \dots + \alpha_p x_p$$

↓

logit of $p(x)$ [log-odds]

Aim of logistic regression is to determine the best value of $\alpha_0, \alpha_1, \dots \alpha_p$ such that $p(x)$ is closer to all actual response (y)

→ Given samples $\{x_i, y_i\} \in \mathbb{R}^p \times \{0,1\}$; $i = 1, 2, \dots, n$

$$\log\left(\frac{p(x_i)}{1-p(x_i)}\right) = \alpha^T x_i \quad ; \; i = 1, 2, \dots, n$$

one need to estimate $\{\alpha_0, \alpha_1, \dots, \alpha_p\} = \hat{\alpha}$

L14.

One uses the technique of the maximum likelihood estimation

$$L(\alpha) = \underbrace{\prod_{i,y_i=1} p(x_i)}_{\text{for class 1}} \cdot \underbrace{\prod_{i,y_i=0} (1-p(x_i))}_{\text{for class 2}}$$

$$= \prod_{i=1}^{n} (p(x_i))^{y_i} (1-p(x_i))^{1-y_i}$$

$$\max_{\alpha} L(\alpha)$$

$$l(\alpha) = \log L(\alpha) = \sum_{i=1}^{n} y_i \log(p(x_i)) + (1-y_i)\log(1-p(x_i))$$

$$= \sum_{i=1}^{n} y_i \left[\log p(x_i) - \log(1-p(x_i))\right] + \log(1-p(x_i))$$

$$= \sum_{i=1}^{n} \boxed{y_i \log\left(\frac{p(x_i)}{1-p(x_i)}\right)} + \log(1-p(x_i))$$

$$= \sum_{i=1}^{n} y_i(\alpha^T x_i) - \log(1-p(x_i))$$

Numerical optimization → Gradient descent.

a | instead of using y as linear combination of different features; one uses sigmoid - convert to probabilities.

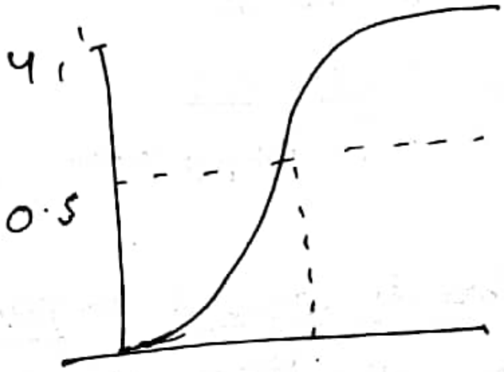
$x_i^* = (x_{1i}^*, x_{i2}^*, \ldots x_{ip}^*)$

$y_i = 0$ or 1

$\Rightarrow \alpha_0 + \alpha_1 x_{i1}^* + \alpha_2 x_{i2}^* + \ldots \alpha_p x_{ip}^*$

$\rightarrow m$

$$Y = \frac{1}{1+e^{-m}}$$



$x_i^* \in$ class-0 if $Y < 0.5$

$x_i^* \in$ class-1 otherwise

Threshold value

Logistic regression returns a probability. The returned probability should be converted to a binary value.

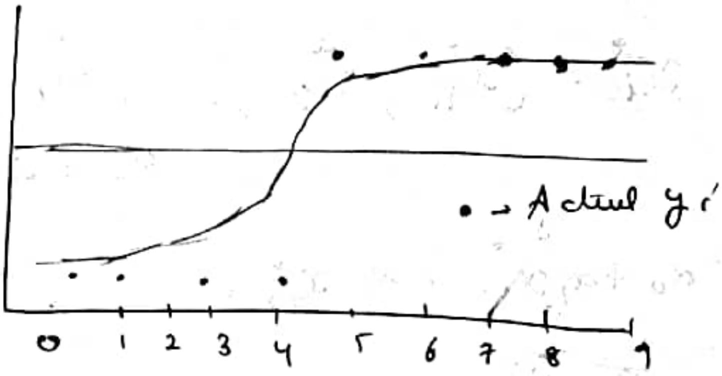
The decision ~~of~~ for converting a predicted probability into a class label is made with the help of a parameter called "threshold".

This is called tuning hyperparameter, which can govern the binary classification.

L14

Single Variate logistic regression

Most straight forward logistic regression is when there is only one independent variable, t.



$$\text{logit}(x_1) = \alpha_0 + \alpha_1 x$$

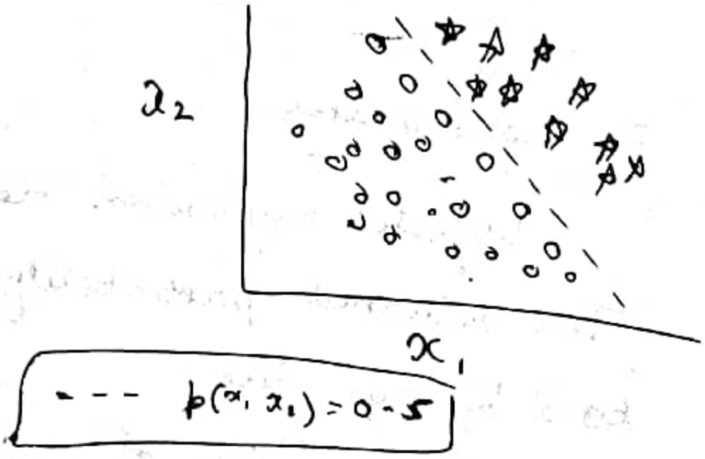
$$p(x) = \frac{1}{1 + \exp(-(\alpha_0 + \alpha_1 x))}$$

Multi Variate logistic regression

has more than one input variables

$$\text{logit}(x_1, x_2) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2$$

$$P(x.x_1) = \frac{1}{1 + \exp(-(\alpha_0 + \alpha_1 x_1 + \alpha_2 x_2))}$$



5a

Classification metrics

One of the key concept in classification performance or metric is Confusion matrix (C.M.)

C.M. → is a tabular visualization of the model predictions nersus the actual labels.

Each row of C.M. represents instances in an actual class.

Each colum represents instances is predicted class

Confusion mtx

Suppose 1100 samples are tested and found 100 +ve & 1000 -ve

		Predicted class	
		Positive	Negative
Actual class	Positive	80	20
	Negative	50	950

Our classification Algorithm use some data and predict 80 as +ve out of 100 and 950 as negative out of 1000.

True Positive = 80

True Negative = 950

Fake Positive = 50

Fake Negative = 20.

Classification Accuracy

$$= \frac{\text{No. of correct prediction}}{\text{Total No. of samples}}$$

$$= \frac{80 + 950}{1100} \approx 93.5\%$$

There are many cases in which classification accuracy is not a good indicator of ones model performance

→ Scenario when class distribution is imbalanced (one class is more frequent)

$$= \frac{2 + 998}{1100} \sim 90.9\%$$

Precision

↳ gives the fraction of correctly identified as positive out of all predicted as positive

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{80}{80 + 50} \sim 0.62$$

useful when false positive is high

L14-6a

Recall / Sensitivity

gives fraction one correctly identified as the out of all positives.

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{80}{80+20} \sim 0.8$$

useful when cost of false ~~positive~~ negative is high

F1 score

F1 score is a measure that combines precision + recall.

Harmonic Mean b/w precision + recall

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$