MAT 555 Additional Material

Regression Models belong to Supervised Learning Class of Algorithms

what is regr.

numerical input $\longrightarrow$ $\boxed{\text{model}_\beta}$ $\longrightarrow$ numerical

*Assume the relationship input ~ output is linear

$(x^{(i)}, y^{(i)})$ $\quad x^{(i)} \in \cancel{\mathbb{R}^n}\ \mathbb{R}^n \quad y^{(i)} \in \mathbb{R}$

$$\boxed{y^{(i)} \approx \beta \cdot x^{(i)}}$$

model parameter

$$\boxed{y^{(i)} \approx \beta \cdot x^{(i)} + \alpha}$$

intercept

What do we learn from the Data?
we determine β!

How β is determined

$$TSE_\beta = \sum_{i=1}^{N} |y^{(i)} - \beta \cdot x^{(i)}|^2$$

we choose β s.t. $\boxed{TSE_\beta \text{ is minimized}}$

$$\beta_0 = \underbrace{\operatorname{argmin}_\beta TSE_\beta}$$

$TSE: \mathbb{R}^n \to \mathbb{R}$

β

TSE is a function in β
if $D = \{(x^{(i)}, y^{(i)}) \in \mathbb{R}^n \times \mathbb{R}\}$ is fixed

---

There are two main Reasons why we do ML

1. make predictions!

   (i) $\longrightarrow$ $\boxed{M}$ $\longrightarrow$ (P)

2. To understand the data particularly to understand the functional relationship between input and output!

Remember

Data $\rightsquigarrow$ we decide on the model type $\rightsquigarrow$ we determine the best model param.

$\boxed{\text{structured Num. Data}} \rightsquigarrow \boxed{\text{Regression}} \rightsquigarrow \boxed{\text{the model param. } \beta}$

Today the β (components of β) will tell us the functional linear relationship between the features and the output
predictors ... regressor

Require an analysis of the model and the relationship between predictors & regressor

$\beta = \langle 1, 2, 3 \rangle$
$f_1 \quad f_2 \quad f_3$

$f_1 < f_2 < f_3$

f_1	f_2	f_3	y

Doesn't mean f_3 more important.
One has to look at the distribution of each of these columns.

proj on the third feat

for example if $\text{Data} \xrightarrow{\pi_3} \mathbb{R}$

$[0, 0.1]$

as opposed to $\text{Data} \xrightarrow{\pi_2} \mathbb{R}$

$[0, 10000]$

Categorical variable Encodings

Categorical: A, numerical: B

A	B
a_1	b_1
a_2	b_2
$\vdots$	$\vdots$
a_n	b_n

Mean

I want to encode A using values in B

(1) Split the data set B using the labels in a.

(2) $\text{Bag}_a \ \forall a \in A$ then

represent each $a \in A$ using mean(Bag_a)

Freq (1) is the same

(2) instead of mean use the # of elements

A	B
a	10
b	11
a	12
b	13

mean

$A = \{a, b\}$

$Bag_a = \{10, 12\} \longrightarrow a \longleftrightarrow 11$

$Bag_b = \{11, 13\} \longrightarrow b \longleftrightarrow 12$

Freq No need for B

A | a | b | a | b |

A | a | a | a | b |

$a \longleftrightarrow 2$
$b \longleftrightarrow 2$

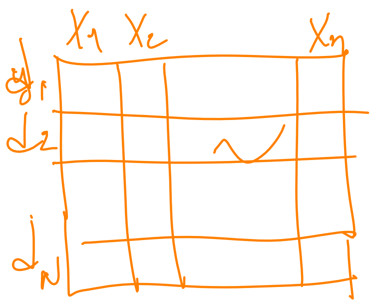
$a \longleftrightarrow 3$
$b \longleftrightarrow 1$

represented by the same number.

Regression with Categorical Variables

Regression must use numerical vectors as input

numerical input $\longrightarrow$ $\boxed{model_\beta}$ $\longrightarrow$ numerical output



If any of X_i is categorical we must convert/encode it as a numerical variable

Logistic Regression

Setup Input $\in \underbrace{\mathbb{R}^n}$ $\times$ $\boxed{\text{Categorical}}$

Output $\in \underbrace{\{0, 1\}}_{\text{Binary.}}$

In other words I want to split my data set into <u>two</u> disjoint subsets

$$\text{Data} \begin{cases} \to 0 & \text{Data}_0 \\ \to 1 & \text{Data}_1 \end{cases}$$

$\boxed{\text{Idea}}$ Create a linear regression model

so that

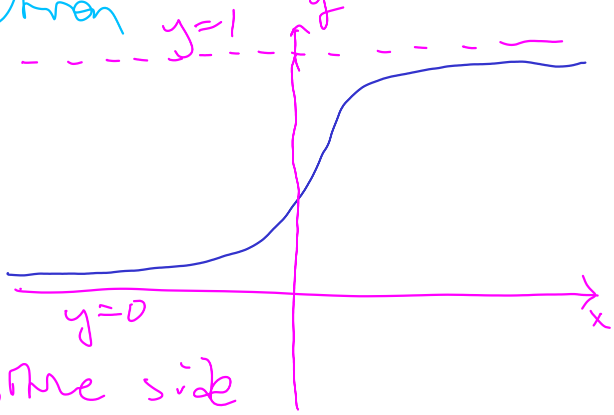
$\text{model}_\beta(x) > 0 \longrightarrow \text{Data}_1$ $\boxed{\text{model}_\beta}$

$\text{model}_\beta(x) < 0 \longrightarrow \text{Data}_0$

Decision on whether $x \in \text{Data}_1$ or $x \in \text{Data}_0$ is done via the sign of model_β output.

sigmoid, logistic function

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



why this function

① it rapidly increases to 1 as soon as x passes to the positive side

② it rapidly decreases to 0 as soon as x passes to negative region model $\quad \beta \cdot X + \alpha \quad$ numerical values in $\mathbb{R}$

$$\rightsquigarrow \sigma(\beta \cdot X + \alpha)$$

input $\longrightarrow$ model $\longrightarrow$ output $\longrightarrow$ $\sigma(x)$

this number here is a number btw 0 and 1

The interpretation

Recall ① if the output is close to 0 the input belongs to Data_0

② if the output is close to 1 the input belongs to Data_1

read $\sigma(\beta \cdot x^{(i)} + \alpha)$ as the $\boxed{\text{probability}}$ that $x^{(i)}$ belongs to Data_1!