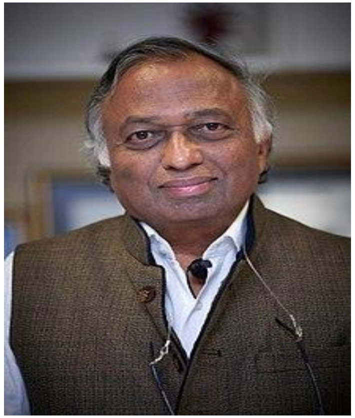
Introduction

Can you recognize these pictures?



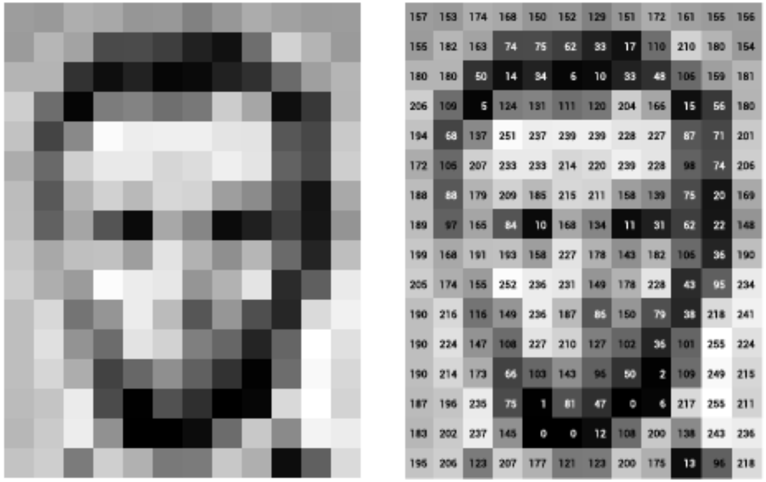
Origin of Machine Learning

- …Lies in early efforts of understanding intelligence

- Intelligence??
  - Ability to comprehend
  - Understand and profit from experience

- Capability to acquire and apply knowledge

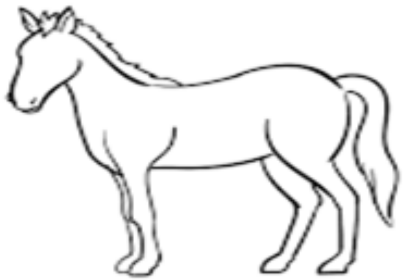
Descriptors/Feature Vectors

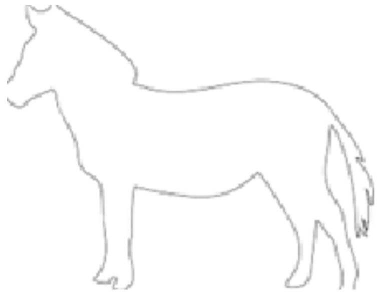


157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	33	17	110	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	15	56	180
194	68	137	251	237	239	239	228	227	87	71	201
172	105	207	233	233	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	165	84	10	168	134	11	31	62	22	148
199	168	191	193	158	227	178	143	182	106	36	190
205	174	155	252	236	231	149	178	228	43	95	234
190	216	116	149	236	187	86	150	79	38	218	241
190	224	147	108	227	210	127	102	36	101	255	224
190	214	173	66	103	143	96	50	2	109	249	215
187	196	235	75	1	81	47	0	6	217	255	211
183	202	237	145	0	0	12	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

Descriptors/Feature Vectors

Descriptors/Feature Vectors: Shape





Feature Vector: Region

170	238	85	255	221	0
68	136	17	170	119	68
221	0	238	136	0	255
119	255	85	170	136	238
238	17	221	68	119	255
85	170	119	221	17	136

Variations: Viewing Angle

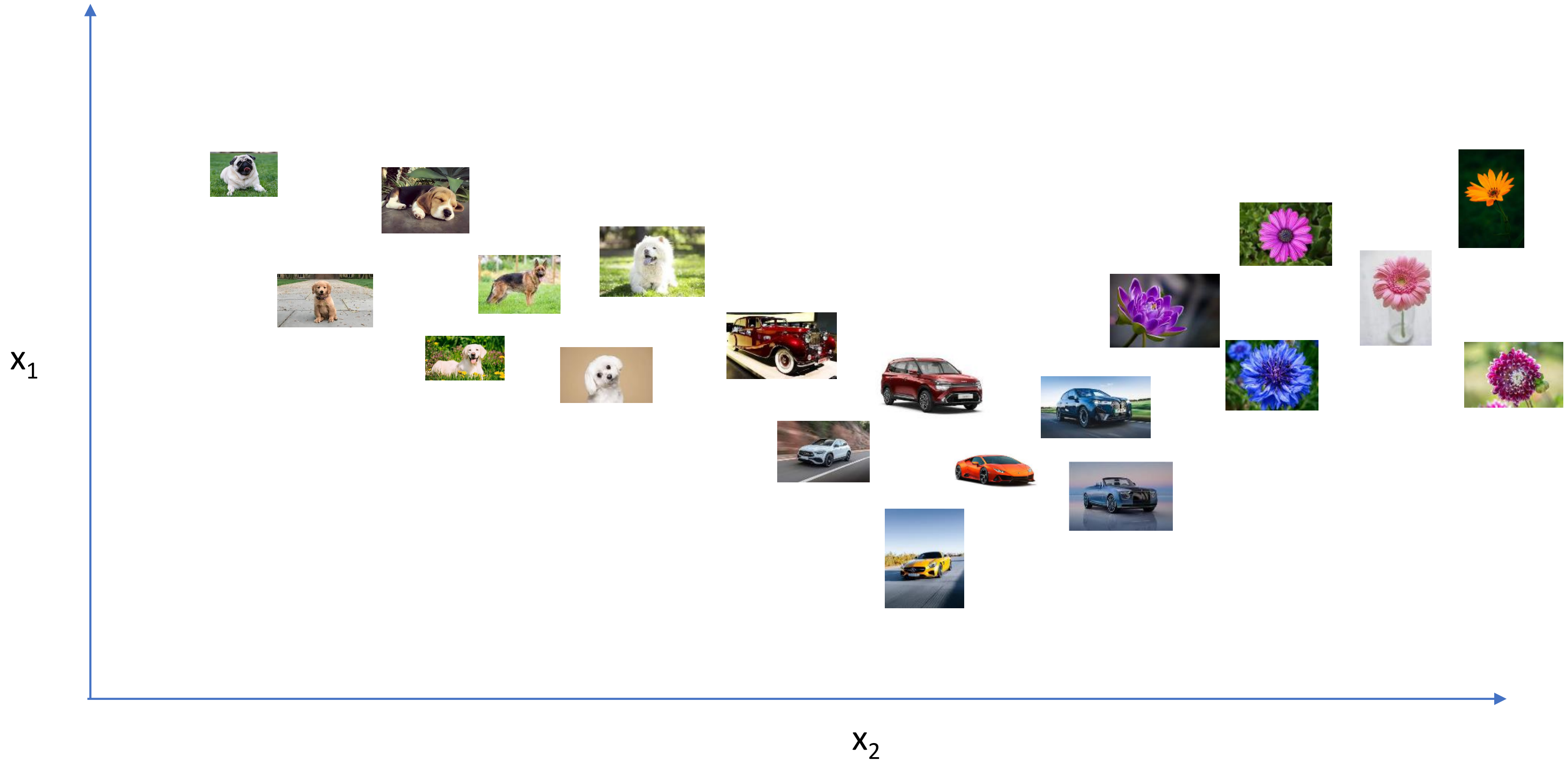
Variations: Pose

Variations: Illumination

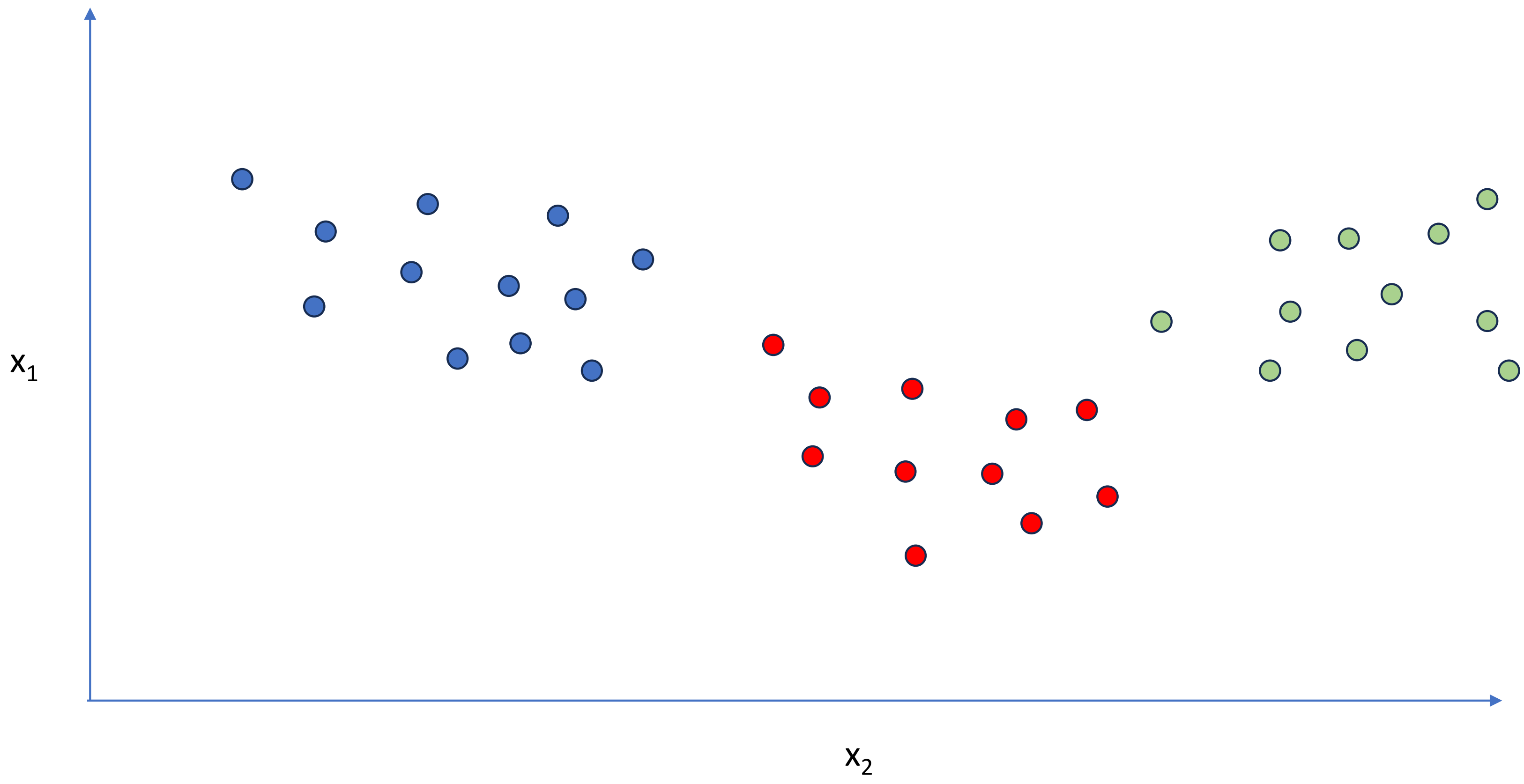
Variations: Intraclass

Variations: Distortions, Occlusions

Distribution of Vectors



Distribution

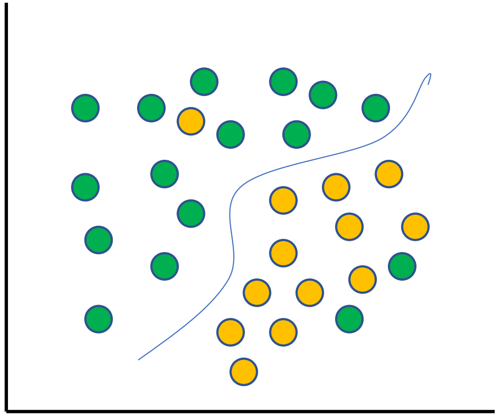


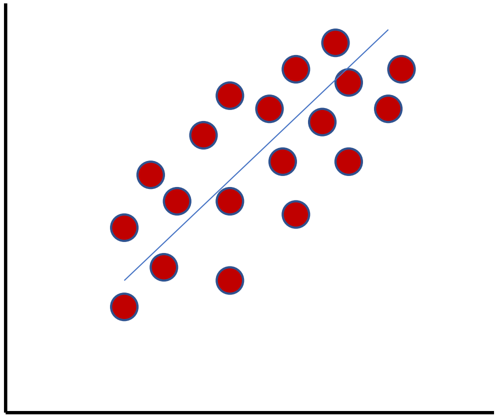
Supervised Learning

- Classification - output variable can be categorized
  - Used to predict category of data
  - Spam detection, sentiment analysis, face recognition

- Regression - output variable is a real value (continuous output)
  - Used to predict numerical values based on previous data observations
  - Some familiar regression algorithms - linear regression, logistic regression, polynomial regression

Classification and Regression

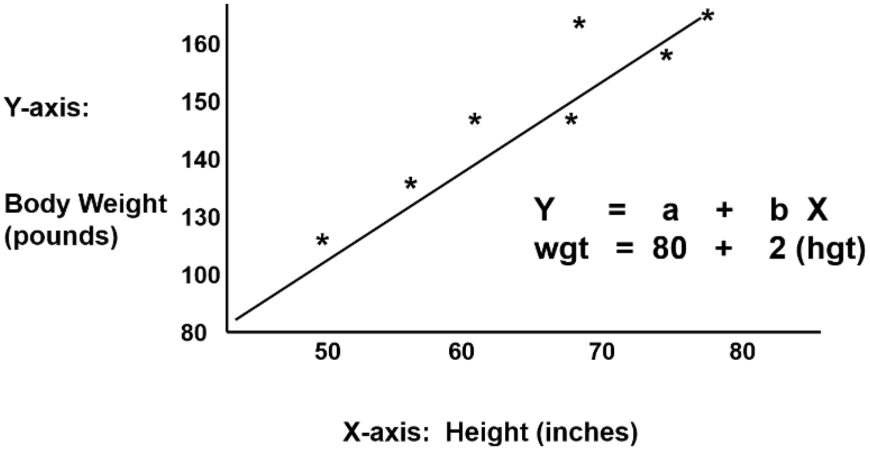




Regression

Linear Regression

- Used to model relationship between two variables by fitting a linear equation to observed data
  - One variable is an explanatory (independent) variable
  - Other is a dependent variable
  - Ex., relate weights of individuals to their heights using a linear regression model



Linear Regression

- Before attempting to fit a linear model to observed data
  - Determine whether or not there is a relationship between variables of interest
  - Not necessary one variable *causes* other
  - A scatterplot can be a helpful tool in determining strength of relationship between two variables
  - If no association between proposed explanatory and dependent variables, fitting a model probably will not provide a useful model

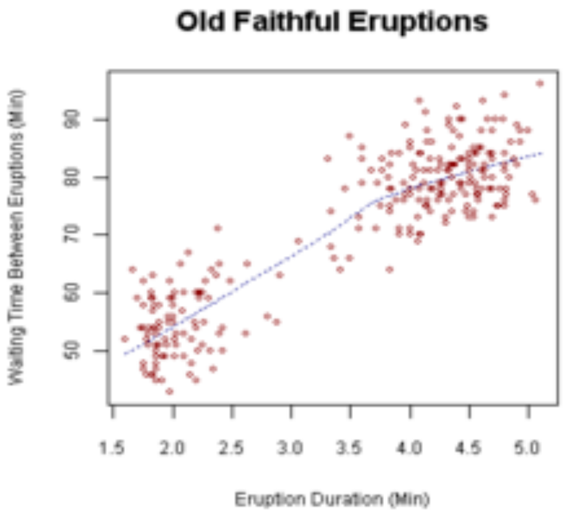
Linear Regression

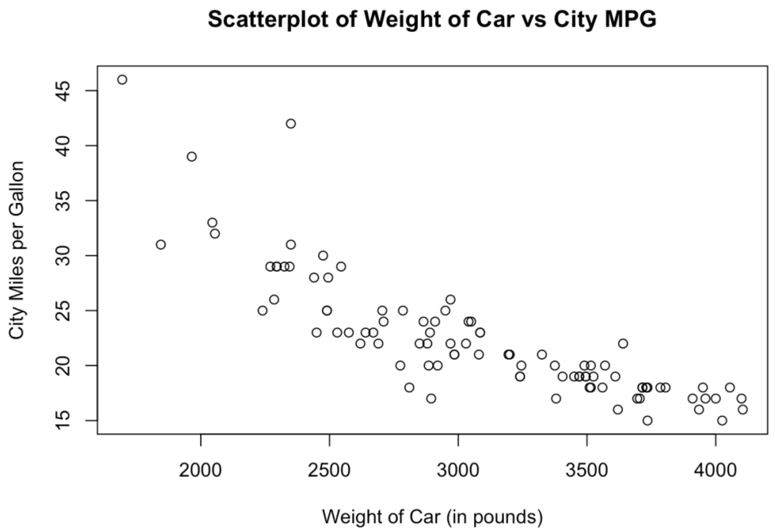
- **Correlation coefficient:** A valuable numerical measure of association between two variables
  - Value between -1 and 1
  - Positive correlation - increasing values in one variable correspond to increasing values in other variable
  - Negative correlation - increasing values is one variable correspond to decreasing values in the other variable
  - Correlation value close to 0 - no association between the variables

Correlation Coefficient Formula

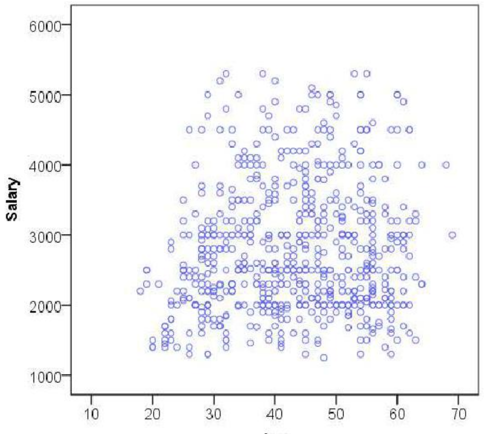
$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

Linear Regression





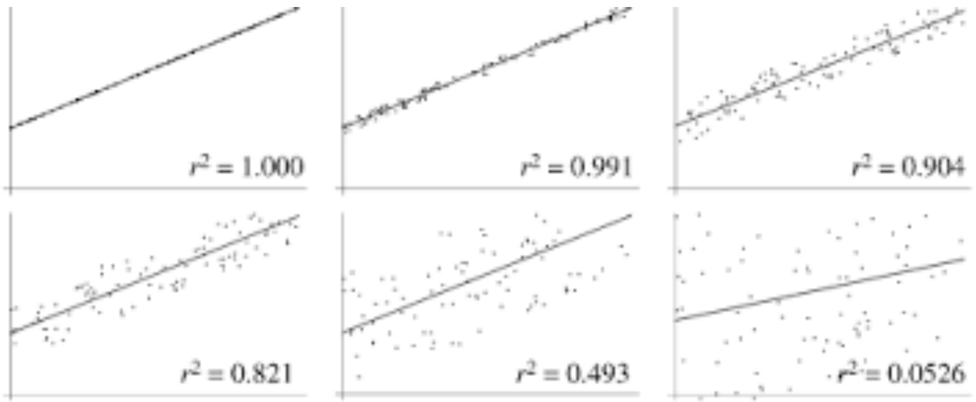
https://www.youtube.com/watch?v=_w2mIqR8VNg



https://bookdown.org/dli/rguide/scatterplots-and-best-fit-lines-single-set.html

Linear Regression

Strength of Association	Coefficient, r Positive	Negative
Small	.1 to .3	-0.1 to -0.3
Medium	.3 to .5	-0.3 to -0.5
Large	.5 to 1.0	-0.5 to 1.0



Linear Regression

- A regression line is obtained which will give minimum error
  - This linear equation used for any new data
  - Ex. given no. of hours studied by a student, model should predict their mark with minimum error

$$y' = w_0 + w_1 * x$$

- Values w_0 and w_1 must be chosen to minimize error

Linear Regression

- Assumption of linearity means that expected value of target (ex. price) can be expressed as a weighted sum of features (ex. area and age):

$$price = w_0 + w_1 * area + w_2 * age \qquad (1)$$

  - w_1 and w_2 are called *weights*, w_0 is called a *bias* (or *offset* or *intercept*)
  - Weights determine influence of each feature on prediction
  - Bias determines value of estimate when all features are zero - allows to express all linear functions of features (versus restricting to lines that pass through the origin)

- Eqn. 1 is an *affine transformation* of input features
  - a *linear transformation* of features via weighted sum, combined with a *translation* via added bias

- Given a dataset, goal is to:
  - Choose weights and bias that, on average, make our model's predictions fit the true prices observed in the data as closely as possible

Linear Regression

SSE → sum of squared errors

$Error = \sum (y_i' - y_i)^2$

- If **sum of squared error (SSE)** is taken as a metric of loss/error, goal is to obtain a line that best reduces error

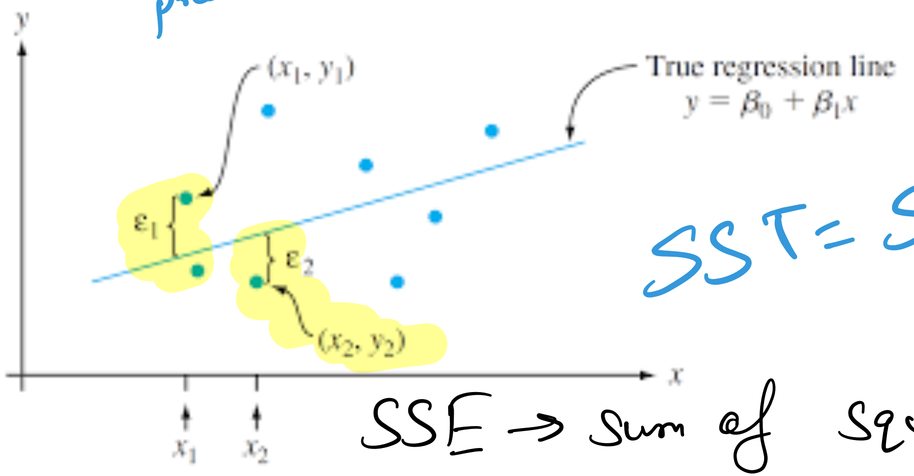
Loss function = Error = $\sum (y'_i - y_i)^2$ for $i = 1 \ldots m$

→ true label

↑ predicted

$SST = \sum (y_i - \bar{y})^2$

$SSR = \sum (y_i' - \bar{y})^2$



$SST = SSR + SSE$

SSE → sum of squared errors

$= \sum (y_i - \hat{y}_i)^2$ → predicted value.

Linear Regression

SST → Sum of squares total

$= \Sigma(y_i - \bar{y})^2$

→ Actual observation

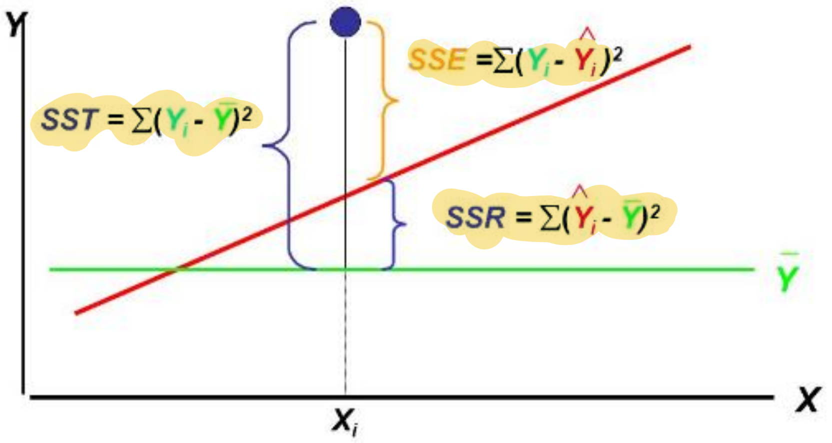
SST = Sum of squares total → Observation - mean

Sum of squared difference between observed dependent variable and its mean

SSR → Sum of squared regression

SSR = Sum of squares regression

Sum of squared differences between predicted value and mean of dependent variable



$R^2 = 1 - \frac{SSE}{SST}$

$1 - \frac{SSE}{SST}$

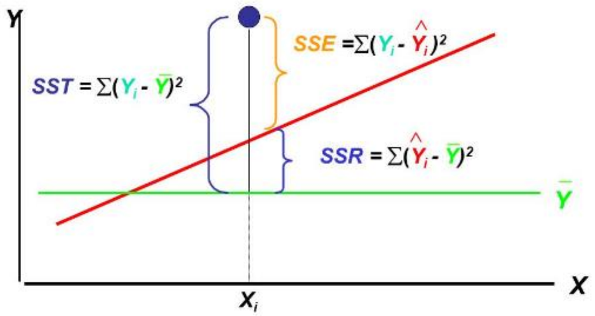
Linear Regression

$R^2 = SSR/SST$

SST = SSE + SSR

Coefficient of determination = R² = **SSR/SST** = **1 – SSE/SST**

- Proportion of variation in dependent variable that is predictable from the independent variable(s)
- Explains variability of data
- R^2 varies between 0 and 1
- If $R^2 = 1$, all data points fall perfectly on regression line; predictor *x* accounts for *all* of the variation in *y*!
- If $R^2 = 0$, estimated regression line is perfectly horizontal; predictor *x* accounts for *none* of the variation in *y*!



https://vitalflux.com/40-regression-analysis-interview-questions-for-data-scientists/

Example

Weight	Height
140	60
155	62
159	67
179	70
192	71
200	72
212	75

Example

x Weight	y Height	X*Y	X^2	Y^2
140	60	8400	19600	3600
155	62	9610	24025	3844
159	67	10653	25281	4489
179	70	12530	32041	4900
192	71	13632	36864	5041
200	72	14400	40000	5184
212	75	15900	44944	5625
1237	**477**	**85125**	**222755**	**32683**

$w_0 = [(\sum Y)(\sum X^2) - (\sum X)(\sum XY)] / [n(\sum X^2) - (\sum X)^2]$

$= [(477)(222755) - (1237)(85125)] / [7(222755 - (1237)^2]$

$= 32.783$

$w_1 = [n(\sum XY) - (\sum X)(\sum Y)] / [n(\sum X^2) - (\sum X)^2]$

$= [7(85125) - (1237)(477)] / [7(222755 - (1237)^2]$

$= 0.2001$

mean

$SST = \sum (y_i - \bar{y})^2$

$y' = 32.783 + 0.2001 * x$

Observed Value

- Compute y' for each data point
- Compute SST, SSR, SSE, R^2

$SSR = \sum (\hat{y} - \bar{y})^2$

$SSE = \sum (y_i - \hat{y})^2$

Linear Regression

- When our inputs consist of *d* features, our prediction y'

  $$y' = w_1 * x_1 + w_2 * x_2 + \ldots w_d * x_d + w_0$$

- Collecting all features into a vector $\mathbf{x} \in R^d$ and all weights into a vector $\mathbf{w} \in R^d$:

  $$y' = \mathbf{w}^T \mathbf{x} + w_0$$

  - Vector **x** corresponds to features of a single example

- Often convenient to refer to features of our entire dataset of *m* examples via matrix **X**
  - **X** contains one row for every example and one column for every feature
  - Predictions $\boldsymbol{y'} = \mathbf{Xw} + w_0$

Linear Regression

- How to update w_0 and w_1 to get best fit line?
    - Update w_0 and w_1 values to reach minimum error

- **Loss Function (*L*):** between y_i and y_i' for sample i

  $L = (y'_i - y_i)^2$

- **Cost Function (*J*):** between y and y' for all m samples

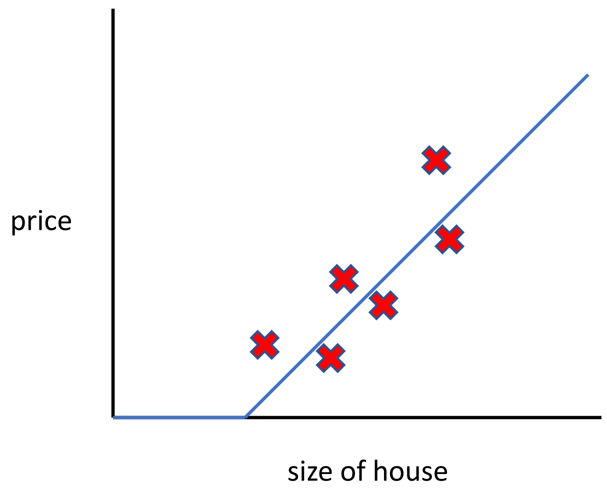
  $J = [(1/2m) * \sum (y'_i - y_i)^2]$ for $i = 1 \ldots m$

  minimize (J)

- How to minimize cost?
    - Gradient Descent – more on this later

Linear Regression - Neural network?

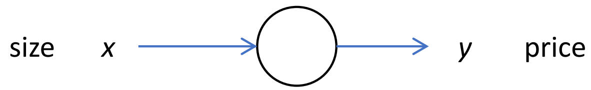
- Ex. Problem: Given the size of a house, predict its price



Linear Regression - Neural network?

- ReLU: Rectified Linear Unit

- A neuron can implement the function ReLU
  - Simplest activation function - linear activation

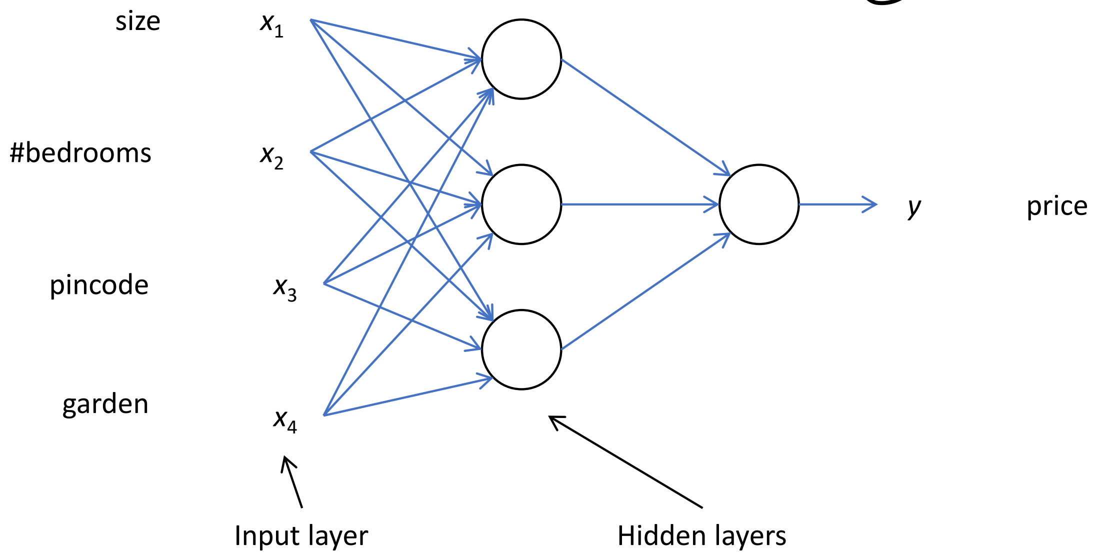


  - Nonlinear activation functions :
    - Allow nodes to learn more complex structures in data
    - Two widely used nonlinear activation functions are **sigmoid (logistic)** and **hyperbolic tangent (tanh)** activation functions

Multiple Regression - Neural network?

- Price - can be affected by other features such as number of bedrooms, pin code, garden area etc.

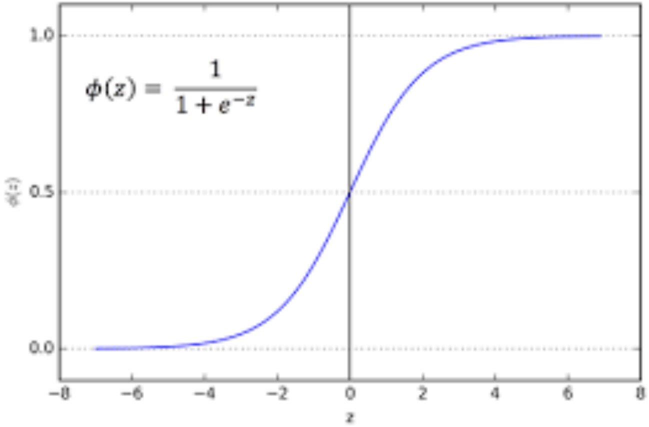
2 Layer Neural network



Logistic Regression → Probability of an event

- Estimates the probability of an event
  - Used for predicting categorical dependent variable using a given set of independent variables
  - Gives probabilistic values which lie between 0 and 1
    - If z is large, $\sigma(z) = 1$
    - If z is large negative number, $\sigma(z) = 0$
    - If $z = 0$, $\sigma(z) = 0.5$
  - Can be used for solving classification problems

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

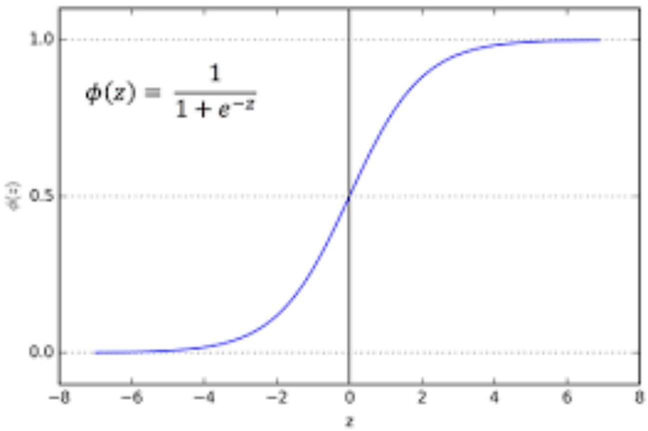


https://www.researchgate.net/figure/Sigmoid-activation-function-for-binary-classification-in-neural-networks_fig3_352647580

$$\sigma'(z) = \sigma(1-\sigma)$$

Logistic Regression

- To predict class - a threshold can be set
  - Obtained estimated probability is classified into classes
  - Ex. if predicted_value ≥ 0.5, then classify email as spam else as not spam



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

Logistic Regression Loss Function

- MSE loss function $(1/2)*(y' - y)^2$ becomes non-convex while learning parameters in logistic regression
  - Global optima is not reached
- **Bernoulli trial:** In binary classification problems:
  - Have to predict value only for one class
  - Probability of negative class can be easily derived from it

$P(y = 1|x) = y'$ $\longrightarrow$

$P(y = 0|x) = 1 - P(y = 1|x) = 1 - y'$

$\Rightarrow P(y|x) = (y')^y (1-y')^{1-y}$

In summary:

$$P(y|x) = y' \qquad \text{if } y = 1$$
$$= (1 - y') \qquad \text{if } y = 0$$

now for loss take log on both sides

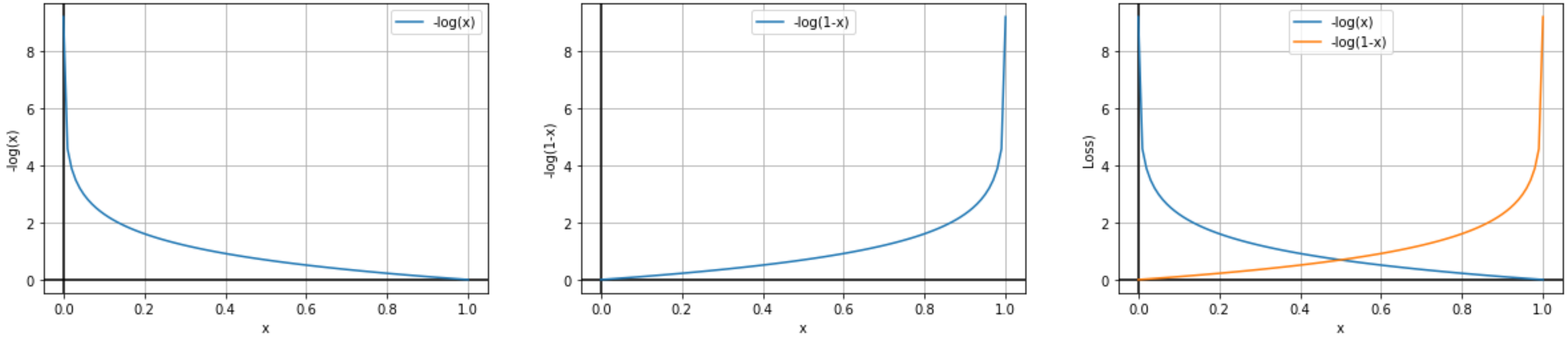
Logistic Regression Loss Function

- $P(y|x) = y'^{y} * (1 - y')^{1-y}$ ..**Bernoulli trial**
  - If $y = 1$, $P(y|x) = y'$
  - If $y = 0$, $P(y|x) = 1 - y'$
- Applying natural log:

  $\log(P(y|x)) = y*\log y' + (1 - y)*\log(1 - y')$ ...Sum of log of probabilities
- To minimize the function:

  $-\log(P(y|x)) = -y*\log y' - (1 - y)*\log(1 - y')$ ...Binary cross-entropy loss

Logistic Regression Loss Function

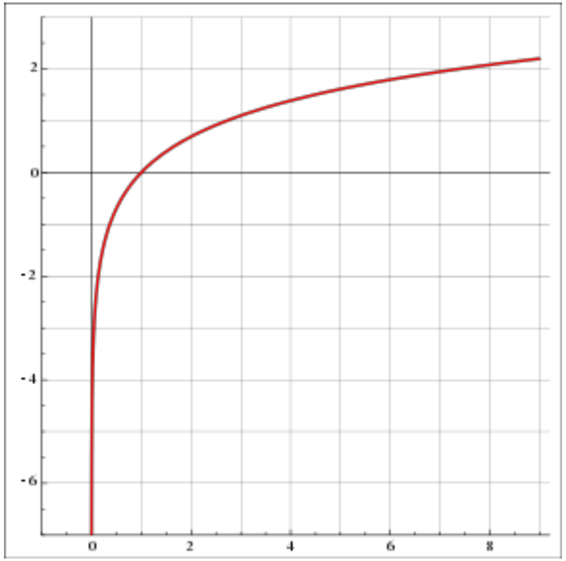


Logistic Regression Loss Function

- **Binary Cross-Entropy Loss function** = $L\ (y, y') = -(y \log y' + (1-y) \log (1-y'))$
  - **True class** y is 0/1; **Predicted** y' is probability between 0 and 1
  - y' is probability for class 1; $(1-y')$ is probability for class 0
  - Log of probabilities for numbers between 0 and 1 is a negative number
- Desirable:
  - $L = 0$ if $y = y'$
  - L should be very high for misclassification
  - $L > 0$

Logistic Regression

- **Loss function** = $L\ (y, y) = -(y \log y' + (1\text{-}y) \log (1\text{-}y'))$
- Correct classification:
  - If $y = 0$ and $y' \sim 0$, $L = -\log(1\text{-}y') \sim 0$
  - If $y = 1$ and $y' \sim 1$, $L = -\log y' \sim 0$
- Misclassification:
  - If $y = 0$ and $y' \sim 1$, $L = -\log(1\text{-}y') = -\log(0) \sim$ large number
  - If $y = 1$ and $y' \sim 0$, $L = -\log y' \sim$ large number



- **Cost function** = $\boldsymbol{J} = (1/m) \sum L\ (y'_i, y_i) \qquad \forall\ i = 1 \ldots m$

The notation $P(1|x) = y$ in the context of logistic regression represents the probability that a binary outcome (usually denoted as 1 for success or the positive class) occurs given a specific set of predictor variables x, and this probability is estimated to be y.

In other words:

- $P(1|x)$ stands for the probability of the binary outcome being 1 (success) given the values of the predictor variables x.
- "x" represents the vector of predictor variables (features) associated with an observation or data point.
- "y" is the estimated probability that the outcome is 1 (success) based on the logistic regression model.

Source → Chat GPT