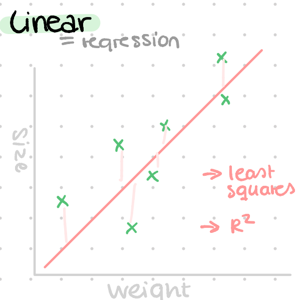
LINEAR REGRESSION

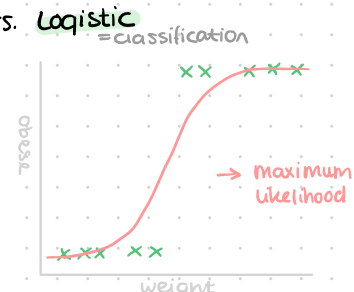
R^2 : determine if two variables are correlated
→ correlation explains strength of relationship
→ R^2 explains to what extend the variance of one variable explains the variance of the other (indep. & dependent variable)

⇒ $R^2 = 0.5$
half of the observed variation can be explained by the models inputs

→ adjusted R^2?

Linear vs. Logistic



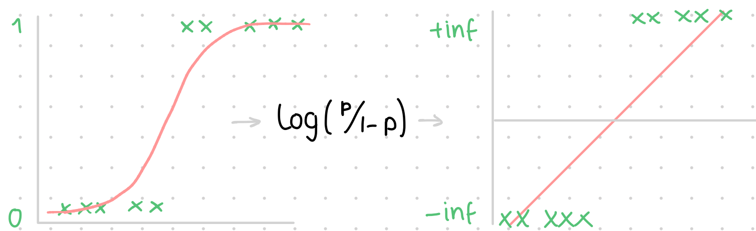


LOGISTIC REGRESSION

- Test if variables effect on prediction is significantly different from 0 using Wald's test.

- With logistic regression the y-axis is confined between 0 and 1 (probability)
  → Solution: y-axis is transformed from "probability of X" to the log(odds of X)". By doing so, the y-axis can go from -infinity to + infinity
  → transformation from s-curve to straight line

Probability (0,1) Log(odds) (-inf, + inf)



- For continuous variables:

	estimate	std. error	z-value (should be >2)	Pr|z| (should be < 0.05)
(intercept)	-3.47	2.364	-1.471	0.1414

→ z-value : estimated / std. error (Wald's test)
  → number of standard deviations away from zero
  → since the estimate in the example above is less than 2 std.dev's away from 0, we know it's not statistically significant.

→ Pr |z| : p-value
  → confirms that the coefficient is not significant as this value should be smaller than 0.05

- For discrete variables

POISSON DISTRIBUTION

- Discrete probability distribution

- A Poisson random variable is a count of the number of occurrences of an event in a given unit of time, distance area or volume

- Conditions:
  1. events are occurring independently
  2. the probability that an event occurs in a given length of time does not change through time (events occur randomly)

- probability mass function

$$P(X=x) = \frac{\lambda^x e^{-\lambda}}{x!} \quad \rightarrow \quad x! : 5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1$$

→ for $x = 0, 1, 2 \ldots\ldots$

→ mean $(\mu) = \lambda$

→ variance $(\sigma^2) = \lambda$

→ average = 2.3 decays p/s

what is the probability that in a 2 second period there are exactly 3 decays

$$P(X=3) = \frac{\lambda^x e^{-\lambda}}{x!} = \frac{4.6^3 e^{-4.6}}{3!} = 0.163$$

- relationship between binominal and poisson distributions

  → the binominal distribution tends toward the Poisson distribution as $n = \infty$, $p = 0$ and np stays constant

  → the Poisson distribution with $\lambda = np$ closely approximates the binominal distribution if n is large and p is small

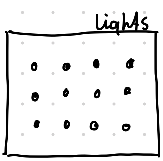
- Examples

  → the number of chocolate chips in a scoop of cookie dough
  ↳ possibly poisson

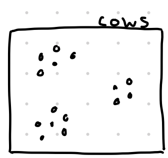
  → avg weight of customers arriving in a 10 min period
  ↳ definitly not Poisson

  → number of deaths from horse kicks in a year
  ↳ possibly poisson

yes
↳ scattered independently



no,
not scattered independently



no,
not scattered independently

- Example of students arriving at campus
  - → all individually and random → poisson
  - → all individually at fixed intervals → not poisson
  - → groups & individuals at random → not poisson

EIGENVECTOR
& eigenvalue

Eigenvectors are stability points of a matrix. This is a column vector that does not get changed when multiplied by the matrix.

$$\underbrace{A}_{\text{matrix}}\underbrace{v}_{\text{eigenvector}} = \overbrace{\lambda}^{\text{eigenvalue}} \underbrace{v}_{\text{eigenvector}}$$

Example:

matrix	eigenvector	eigenvalue
$\begin{bmatrix} -6 & 3 \\ 4 & 5 \end{bmatrix}$	$\begin{bmatrix} 1 \\ 4 \end{bmatrix}$	6

Av

$$\begin{bmatrix} -6 & 3 \\ 4 & 5 \end{bmatrix}\begin{bmatrix} 1 \\ 4 \end{bmatrix} = \begin{bmatrix} -6 \cdot 1 + 3 \cdot 4 \\ 4 \cdot 1 + 5 \cdot 4 \end{bmatrix} = \begin{bmatrix} 6 \\ 24 \end{bmatrix}$$

λv

$$6 \begin{bmatrix} 1 \\ 4 \end{bmatrix} = \begin{bmatrix} 6 \\ 24 \end{bmatrix}$$

→ Multiplying a matrix by a vector gives the same result as multiplying a scalar by that vector.

Why?

- When performing transformations on matrices, the eigenvector is the direction that doesn't change direction.

GOOGLE ALGORITHM

1. Each website gets equal rank.
   i.e. 8 "credits" give away.

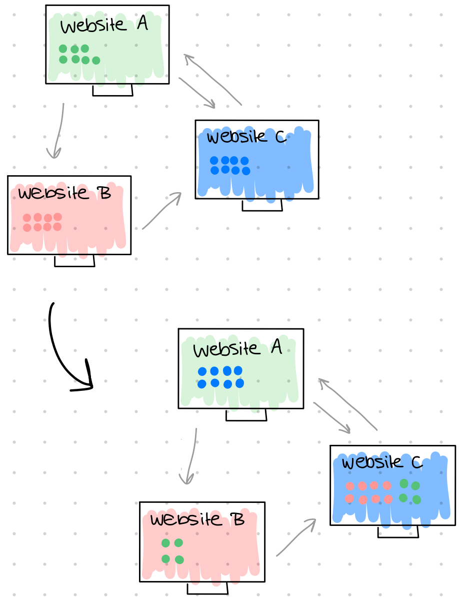
2. Distribute credits equally to websites they link to.
   i.e both B and C get 4 credits from A.

3. Proces needs to be repeated for optimal rank. i.e. A will be boosted by the fact that C is strong.

	1	2	3	4	5	6	7	8	9
A	8	8	12	8	10	10	9	10	9.5
B	8	4	4	6	4	5	5	4.5	5
C	8	12	8	10	10	9	10	9.5	9.5

→ problem: algorithm doesn't stabalize and is inefficient.

→ solution: matrices & eigenvectors



Matrix of redistribution:

$$\begin{array}{c|ccc} & A & B & C \\ \hline A & & & \\ B & & & \\ C & & & \end{array} \quad \begin{array}{c} \\ A \\ B \\ C \end{array}\begin{bmatrix} 0 & 0 & 1 \\ .5 & 0 & 0 \\ .5 & 1 & 0 \end{bmatrix} \qquad \text{eigenvector} \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix}$$

→ credits should be split with an 2 : 1 : 2 distribution.

Performance metrics search engine:

- Precision: relevance
- Recall: ranking