Variables

Data Mining

Edgar Roman-Rangel.
edgar.roman@itam.mx

Computer Science Department.
Instituto Tecnológico Autónomo de México, ITAM.

Varying magnitudes

Often, variables have different order of magnitude among them. This characteristic might affect calculations.

- Decisions might bias towards variables of higher magnitude.
- Distances might lost meaning.
- Pure uniform distribution could appear.

It is common to rely on scaling and normalization techniques.

Scaling

Column-wise:

Min-Max scaler

$$x_n = \frac{x_n - \min(x_n)}{\max(x_n) - \min(x_n)}.$$

Standard scaler

$$x_n = \frac{x_n - \mu_n}{\sigma_n}.$$

Robust scaler

$$x_n = \frac{x_n - P50_n}{P75_n - P25_n}.$$

Normalization

Row-wise:

L1-norm

$$x_n = \frac{x_n}{\sum_m |x_m|}.$$

Max-norm

$$x_n = \frac{x_n}{\max_m(x_m)}.$$

L2-norm

$$x_n = \frac{x_n}{\sqrt{\sum_m (x_m)^2}}.$$

Softmax-norm

$$x_n = \frac{e^{x_n}}{\sum_m e^{x_m}}.$$

Q&A

Thank you!

`edgar.roman@itam.mx`