Formularium Data Analytics – januari 2023

Normale verdeling en beschrijvende statistiek

Normale verdeling

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

μ = gemiddelde; σ = standaardafwijking

Z-score (standaardiseren)

$$Z = \frac{X - \mu}{\sigma}$$

Standaardafwijking (steekproef)

$$s_x = \sqrt{\frac{\sum_i n_i\,(x_i - \bar{x})^2}{n-1}}$$

Data en afstand

Euclidische afstand

$$d(X,Y) = \sqrt{\sum_{k=1}^{n}(x_k - y_k)^2}$$

Minkowski afstand

$$d_{ij} = \sqrt[r]{\sum_{k=1}^{n}\left|x_{ik} - x_{jk}\right|^r}$$

Classificeren - beslissingsbomen

Entropie(t)
$$= -\sum_{i=0}^{c-1} p(i|t)\log_2 p(i|t)$$

Gini(t)
$$= 1 - \sum_{i=0}^{c-1}[p(i|t)]^2$$

Informatiewinst
$$\Delta = I(\text{parent}) - \sum_{j=1}^{k}\frac{N(v_j)}{N} I(v_j)$$

Met $p(i|t)$ de fractie records die behoren tot klasse i in node t, $I(v_j)$ de onzuiverheid van een gegeven node v_j; *N* het totaal aantal nodes; *k* het aantal attribuutwaarden en $N(v_j)$ het aantal records van de child node v_j.

Classificeren - Bayesiaans

Bayes <u>posterior</u> kans

$$P(C|A_1A_2 \ldots A_i) = \frac{P(A_1A_2 \ldots A_i|C)P(C)}{P(A_1A_2 \ldots A_i)}$$

Naïef Bayes classificeren

We veronderstellen onafhankelijkheid tussen attributen A_i met gegeven klasse

$$P(A_1A_2 \ldots A_n|C) = P(A_1|C_j).P(A_2|C_j) \ldots P(A_3|C_j)$$

Nieuw punt wordt onder C_j geklassificeerd als $P(C_j)\prod P(A_i|C_j)$ maximaal is.

Confusion matrix

Recall of TPR

$$r = TPR = \frac{TP}{TP + FN}$$

FPR

$$FPR = \frac{FP}{TN + FP}$$

Precision

$$p = \frac{TP}{TP + FP}$$

F1

$$F_1 = \frac{2 * TP}{2 * TP + FP + FN}$$