Montag, 16.10.2023

Wahrscheinlichkeitsverteilung (Probability Distribution)

f: Ereignis ⟶ "Häufig"
Häufigkeit

- diskret
- kontinuierlich

Galton Board / Galton Brett ~ Binomial Distribution
~ Bernoulli Verteilung

Log-normal distribution

Bayes..

Daten x, Ziel y

Bedingte Wahrscheinlichkeit ~ Conditional probability

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

A given B — A oder B, d durch B

$$P(A|B) * P(B) = P(A \cap B)$$

$$P(A|B) * P(B) = P(B|A) * P(B)$$ Durchschnitt

naive Bayes ⟶ Bayes dengan asumsi data independen
↳ hanya cek korelasi data & independent features

Kategorial verteilung

Dienstag 17.10.2023

Encoding → Hot
Ordinal

Bayes - Klassifikator für nominale Merkmale (Kapitel 4.3.2) Seite 88 } Bsp. Fischkiste

$$P(i,x) = \frac{\prod_{u=1}^{p_{x1}} P(x^{(u)} | i) \cdot P(i)}{\sum_{j=1}^{M} P(j) \prod_{u=1}^{n} P(x^{(u)} | j)}$$

Mittwoch, 18.10.2023

Linear Regression

fit in sklearn

$$y = \text{intercept\_} + \text{coef\_} * x$$

$$y = b + mx$$

Newton Gauss Verfahren

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

Nacheinander ohne Pipeline

Donnerstag, 19.10.2023

Skalierung / Scaling

Linear Regression:

- ~~Skalieren~~ Nicht Skalieren → wenn nur das predict interessiert
- Skalieren → wenn die Wichtigkeit der Spalten interessiert

Lasso, Ridge

↳ Skalieren, weil es sonst nicht konvergiert.

Bayes

↳ nicht skalieren

$O(n)$ → Big O ← Aufwand von Algorithmus

Vector n Komponent → addieren: ~~Schleife~~ n ~ (O_n)

Matrix n Zahlen, 2 Spalten, Summen der Spalten

~~For~~ +n ---- $O(n)$

Matrix n Zeilen ~~f Spalten~~ f Spalten

$O(n*f)$

Bayes: ft $O(n*f)$

predict (of)

Linear Regression: $O(n*f^2)$

Nächste Nachbarn (kNN)

- Label
- Eigenschaften
  - Gewichtung der Eigenschaften
  - Gruppierung nach Ähnlichkeit
    - totale Übereinstimmung
    - relative kurzer Abstand

KD Tree

Ball Tree

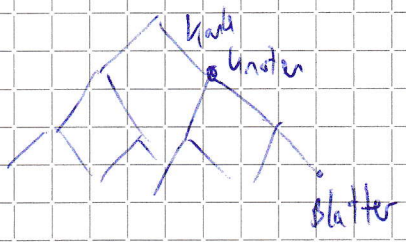
Freitag, 20.10.2023

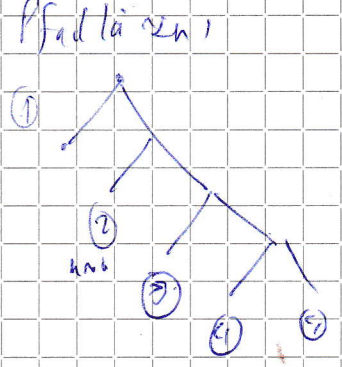
~~Decision Tree~~ kNN (nearest neighb.)

Distanz: - Euclidean (L2)
- Haversine
- Manhattan, cityblock (L1)
- Cosine

Decision Tree:

$O(\log(n))$





- Anzahl Blätter: wenig ist besser
  ↳ Anzahl der Regeln
- Pfadlänge: kurz ist besser
  ↳ Länge der ~~Regel~~ längsten Regeln (worst case)
- Pfadlängensumme: weniger ist hi.
  ↳ durchschnittl. Regellänge

Entropy