Montag, 09.10.2023

KLR-335 Machine Learning

Jupyter Notebook
Pycharm ↘

Dienstag, 10.10.2023

Machine Learning ← still slide 1

Projektarbeit ...

Box Turtles / OpenAI Hide & Seek ← youtube
Emergent Tool Use from Multi-Agent Interaction

Vorliegende Texte
Textdaten
}
Neu daten ... neu trainieren

Test Menge { Training Menge } Pareto
20% 80%

Test Menge Trainingsmenge

numpy, pandas, matplotlib
Scipy, Scikit-learn

Mittwoch, 11.10.2023

Test_set, Train_set

- Stratified Sampling

train_test_split(...

Stratify = housing_geom["income_cat", ... "median_price_cat"]

...

↑

has to be category, not continue.

Correlation!

Aussagekraft

Donnerstag, 12.10.2023

Antisipasi data yang hilang → mengandung None atau NaN

↳ scikit learn

↳ from sklearn.impute import SimpleImputer

imputer = SimpleImputer(strategy="median")

imputer.fit(housing_num, y=None)

pandas dataframe

sparse matrix in numpy → Compressed Sparse Row format (CSR)

Freitag, 13.10.2023

Pipeline

↳ hard to debug

from sklearn.pipeline import Pipeline

has to have

fit & transform ← class/fungsi in Pipeline

sklearn metrics

(A) fit & transform ← Transform...

(*) fit & predicts ← model & estimator

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

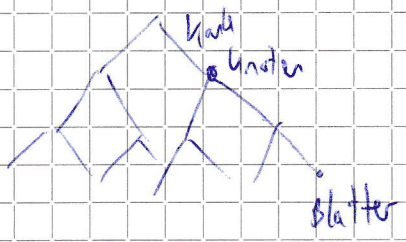
Freitag, 20.10.2023

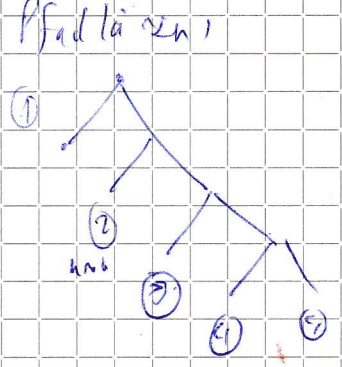
~~Decision Tree~~ kNN (nearest neighb.)

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

Montag, 23.10.2023

Decision Tree

↳ minimizing standard deviation ← continuous data

↳ famous of overfitting.

Cross validation

~~grid~~ Decision Tree Classifier

Criterion:

- gini → Gini impurity
- log_loss } Shannon Information Gain
- entropy

Splitter:

- best
- random

Warum bestrafen? ← Erklären / Begründe

↳ Imputer } Process...?

↳ ordinal ...

↳ scaler ...

↳ Algorithmus ...

SVM = Support Vector Machine

↳ takes time in Sklearn.

from sklearn.datasets import make_blobs ← new version

Steigungsfaktor = gradient

Achsenabschnitt = (with) pitch symbol

Randbereich = margin

Dienstag, 24. 10. 2023

SVM = Support Vector Machine

↳ StandardScaler → für SVM ← so scaling may be needed

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$$

$$\gamma = \frac{1}{2\sigma^2} \quad \Leftarrow \text{für GridSearch}$$

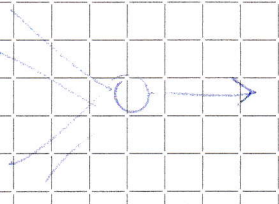
$$k(x, x') = \exp\left(-\gamma \|x - x'\|^2\right)$$

↑ Gamma in GridSearch

γ = Gamma in SVC in sklearn

```
gs = GridSearchCV(svc, parameter)
gs.fit(X, y)
best = gs.best_estimator_
best.predict(Xneu)
```

Neuronal Networks: → Kapitel 7 Früchte



$$e = \hat{y} - y$$

desired output

$$y_i = \sum_j w_{ij} x_j$$

$$\frac{\partial e}{\partial y_i} = -1 \quad \& \quad \frac{\partial y_i}{\partial w_{ij}} = x_j$$

$$\frac{\partial e}{\partial w} = x_j$$

$$\frac{\partial e^2}{\partial w} = \frac{\partial e^2}{\partial e} \cdot = 2e\, x_j$$

Mittwoch, 25.10.2023

Projekt

~~Score~~ mean

- cv - score ← cross validation
- gs - score ← grid search score

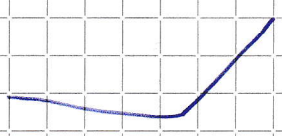
```
def __init__(self, skalier = True, scaler = MinMaxScaler()):
    self.sc = scaler
    self.skalier = skalier
def fit(self, X, y)
    if skalier == self.skalier:
        X_skal = self.sc.fit_transform(X)
    else
        X_skal = X
```

predict(X)

↳ scale or not scale...

Activation functions

RELU



Batch Learning & Incremental Learning

Voting Classifier ⟶ for project ?!?! } vs. GridSearch

↳ find best Classifier...

Random Forest

↳ has many trees

Bagging:

↳ BaggingClassifier & BaggingRegressor

Bootstrap ← in RandomForest

oob_score ← out-of-bag ← only if Bootstrap is true to RandomForestClassifier

Bagging vs Boosting

Donnerstag, 26.10.2023

Dimensionality Reduction } Feature Engineering
- feature selection
- PCA

Compact Estimator

PCA (Hauptkomponentanalyse)
└ Kovarianz-Matrix

Zentriert: ← mean will be zero
$(x - \bar{x})$

x = x - x.mean()

np.linalg.eigh(M) ← eigen hermitian
↓ eigen vector
eig ← eigen

Hermitian Matrix ~ symmetrisch diagonale Spiegel
diagonal mirror

pca.components_ ← eigen vectors
pca.singular_values ← eigen values
└ sorted from large to small

pca.explained_variance ← sorting pca.components based on this
pca.noise_variance
pca.explained_variance_ratio

TSNE ~ Manifold

Projekt:

↳ Ziel von Projekt ① — Richtgröße / was ist minimiert/maximiert

Recall = gefunden in Varianz / alle echt

Precision = richtig in Varianz / alle Vorhersagen

Bericht → PDF

↳ Word

Freitag, 27.10.2023

Clustering

- DBSCAN
- Hierarchical Clustering
- K-Mean

DBSCAN?

- Kernpunkt
- Randpunkt
- Rauschpunkt

K ist random

when is cluster good?

KD Bäume..

Montag, 30.10.2023

Red Wine & White → CSV Datei

output = kategorisch.

- Bayesian? Score? →
- kNN?
- Neural network? X
- PCA?

Duplicates?

Ausreißer

df1.loc[df1.duplicated(), :]

Was für Sorte (metrics)

↳ precision

↳ recall

↳ classification_report.

↳ Confusion matrix

nur Weiss Wein ↔

↳ outlier out ...

Dienstag, 31.10.2023

3, 4 schlecht

5, 6 OK

7, 8, 9 gut

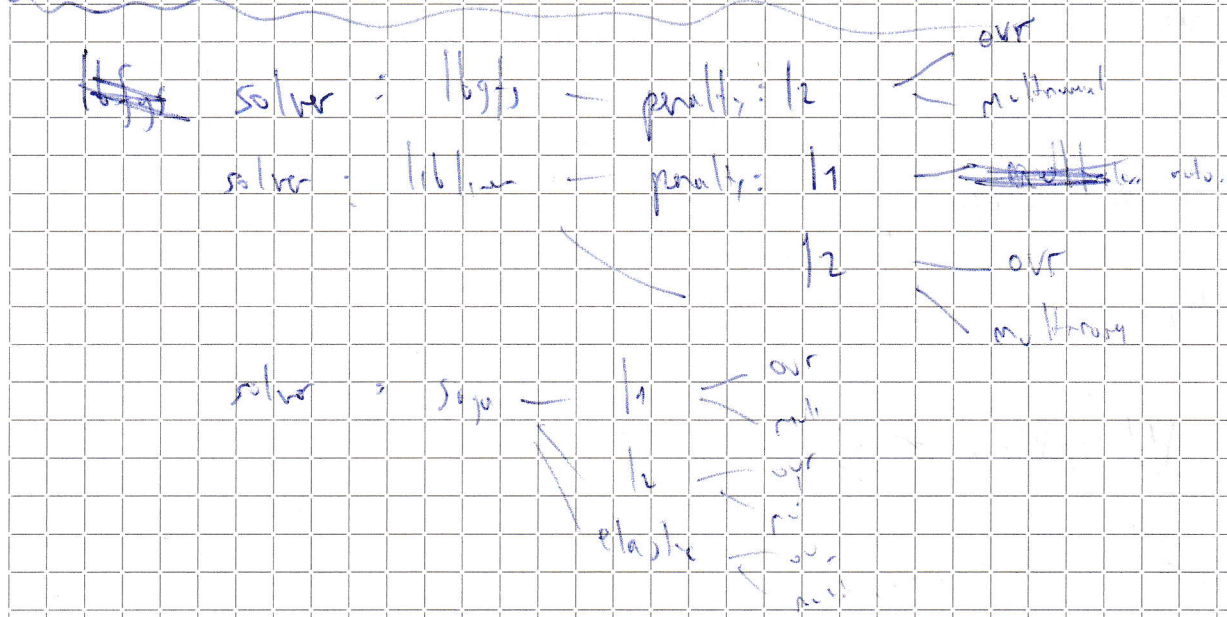
SVM ok

PCA + Logistische Regression

Precision 80% Gute Vorhersage

Oder

Recall



~~Mittwoch~~, 02.11.2023
Donnerstag

Klassen
- cross_val_score(...) ①

2 Klassen ?

Abgabe: Freitag ~~mittag~~

Ausreißer behalten ???

pickle Daten laden

Gaussian Naive Bayes ???

Log Reg.