Data Science

Vorlesung 10

Prof. Dr. Christian Bockermann

Hochschule Bochum

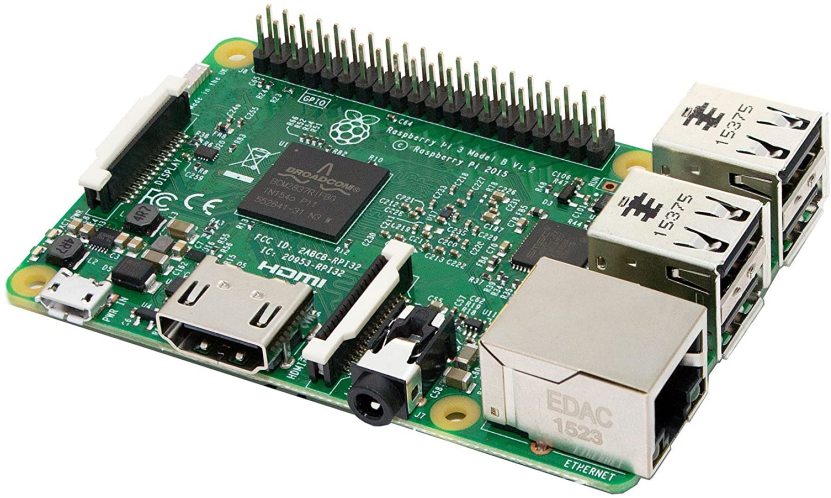
Wintersemester 2020/2021

Hörspiele für Kinder

Verschiedene Bastelprojekte

- Python-Tools für RFID-Karten
- Open-Source MP3-Player Software
- Phoniebox: `http://phoniebox.de/`

Raspberry Pi



Wie ist das Hör-Verhalten der Kinder?

- Welche Titel/Themen sind beliebt?
- Gibt es Zielgruppen mit bestimmten Vorlieben?

Wie ist das Hör-Verhalten der Kinder?

- Welche Titel/Themen sind beliebt?
- Gibt es Zielgruppen mit bestimmten Vorlieben?

Andere machen das auch:

Spotify

Music

SOUNDCLOUD

amazon music

DEEZER

Fragerunde

Warum errechnet man den Trainingsfehler als 1 - accuracy_score?

```
from sklearn.metrics import accuracy_score

y_pred = m.predict(X_train)

train_error = 1 - accuracy_score(y_train, y_pred)
```

Die *Confusion Matrix* (Konfusionsmatrix)

		Vorhersage ŷ		
	“Wahrheit” y	**Klasse Pos**	**Klasse Neg**	
	Klasse Pos	True Pos (TP)	False Neg (FN)	TP / (TP + FN)
	Klasse Neg	False Pos (FP)	True Neg (TN)	TN / (FN + TN)
		TP / (TP + FP)	TN / (TN + FP)	

Definition: accuracy

Die accuracy ist definiert als

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Hinweis:

$$accuracy = 1 - relativeError$$

Definition: precision

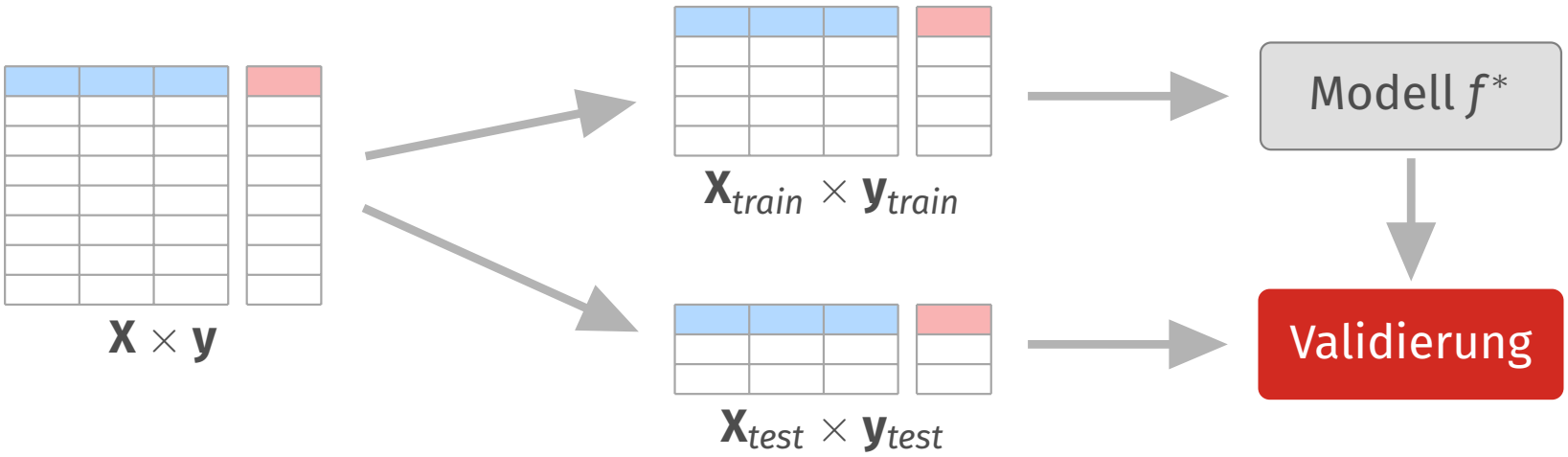
Die Kennzahl precision ist definiert als

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

$$
\begin{aligned}
& accuracy = 1 - relativeError \\
\Leftrightarrow \quad & accuracy + relativeError = 1 \\
\Leftrightarrow \quad & relativeError = 1 - accuracy
\end{aligned}
$$

Wie berechnet man den Test-Fehler?

Ansatz: Aufteilung in Trainings- und Test-Daten

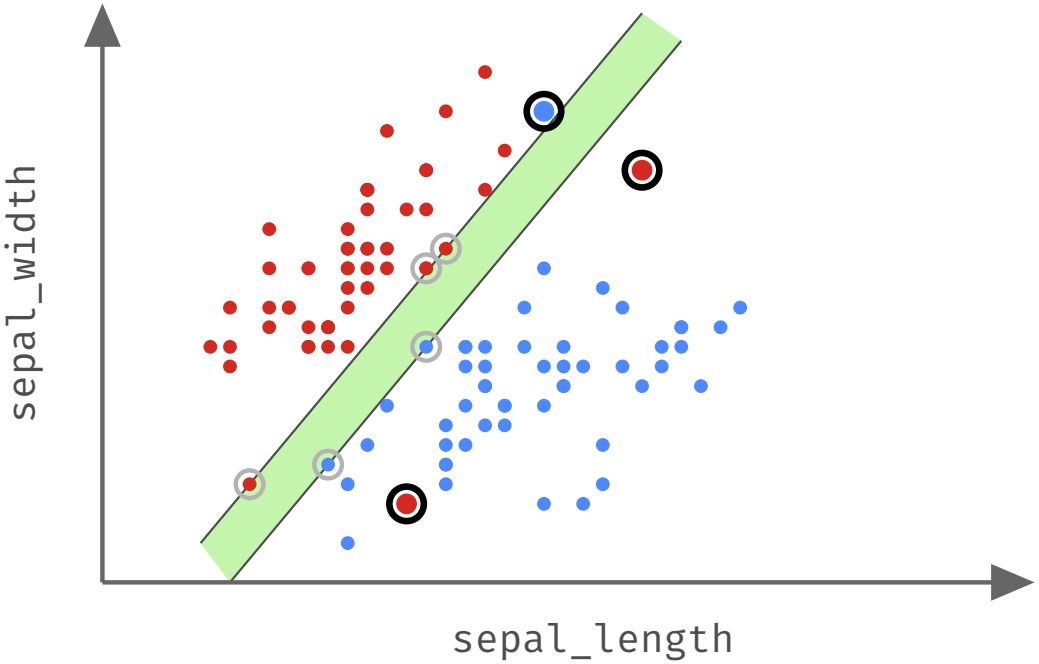


- Nutze *unabhängige* Test-Daten um f^* zu validieren!
- Oft 80% Trainingsdaten, 20% zum Testen (auch 70/30)

Muss ich wissen, wie der `train_test_split` den Split genau vornimmt?
Oder reicht es, wenn ich den Parameter `train_size` kenne?

Was macht der C Parameter bei der SVM?
Wie bestimmt man ihn?

Lineare SVM geht von linearer Separierbarkeit aus



- Was ist, wenn es Ausnahmen in den Daten gibt?
- Dann sind einige Bedingungen verletzt $\Rightarrow$ keine Lösung

Wie kommen wir trotzdem zu einer Lösung?

Ich habe noch Probleme, lambda Funktionen zu verstehen.

```
# Beispiel aus Uebungsblatt 8

df['Volumen'] = df.apply(lambda r: r['Hoehe'] * r['
                                 Breite'] * r['Tiefe'],
                                 axis=1)


df['Volumen'] = df['Hoehe'] * df['Breite'] * df['
                                 Tiefe']
```

Ich habe noch Probleme, lambda Funktionen zu verstehen.

```
xs = [('A', 2), ('Z', 0), ('B', 4), ('G', 7)]

sorted( xs, key = lambda x : x[1] )
```

Ohne `lambda` Funktion:

```
def zahl(tup):
  return tup[1]

sorted(xs, key=zahl)
```

Blatt 6 - warum ist die Antwort von `ist_prim` ein `Bool` Typ?

Blatt 6 - warum ist die Antwort von `ist_prim` ein `Bool` Typ?

```
def teiler(x):
    return [t for t in range(1,x) if x % t == 0]

def ist_prim(p):
    return len(teiler(p)) == 1
```

Was macht `len(s.values)`?

Was macht `len(s.values)`?

`len` ist eine Funktion zu Berechnung der Länge von Sequenzen:

```
# Angenommen s ist ein Series Objekt:

s = pd.Series([1,2,3,4,5])

len(s.values)   # ergibt 5 => 5 Elemente in der
                            Series

sum(s.values)   # ergbit 15 => 1 + 2 + 3 + 4 + 5
```

Was macht `len(s.values)`?

`len` ist eine Funktion zu Berechnung der Länge von Sequenzen:

```
# Angenommen s ist ein Series Objekt:

s = pd.Series([1,2,3,4,5])

len(s.values)    # ergibt 5 => 5 Elemente in der
                              Series

sum(s.values)    # ergbit 15 => 1 + 2 + 3 + 4 + 5
```

Diese Dinge können Sie sehr gut im Jupyter Notebook ausprobieren!!