02.01_Daten_Iris

August 22, 2020

0.1 Daten werden in Tabellen organisiert.

- Jede Zeile entspricht einem Datensatz ("Sample")
- Jede Spalte bezieht sich auf eine Eigenschaft ("Feature")

Damit werden die Daten beschrieben durch eine Matrix `X` (die sog. *Design Matrix* oder auch *Features Matrix*) mit `n_samples` Zeilen und `n_features` Spalten. Diese wird häufig als Pandas `DataFrame` gehalten.

Neben den Features brauchen wir noch die sog. *Lables* oder *Targets* `y`, also das, was aus den Features erkannt werden soll. Dies ist ein Vektor, der `n_samples` Einträge hat.

[5]:
```
import seaborn as sns
iris = sns.load_dataset('iris')
iris.head()
```

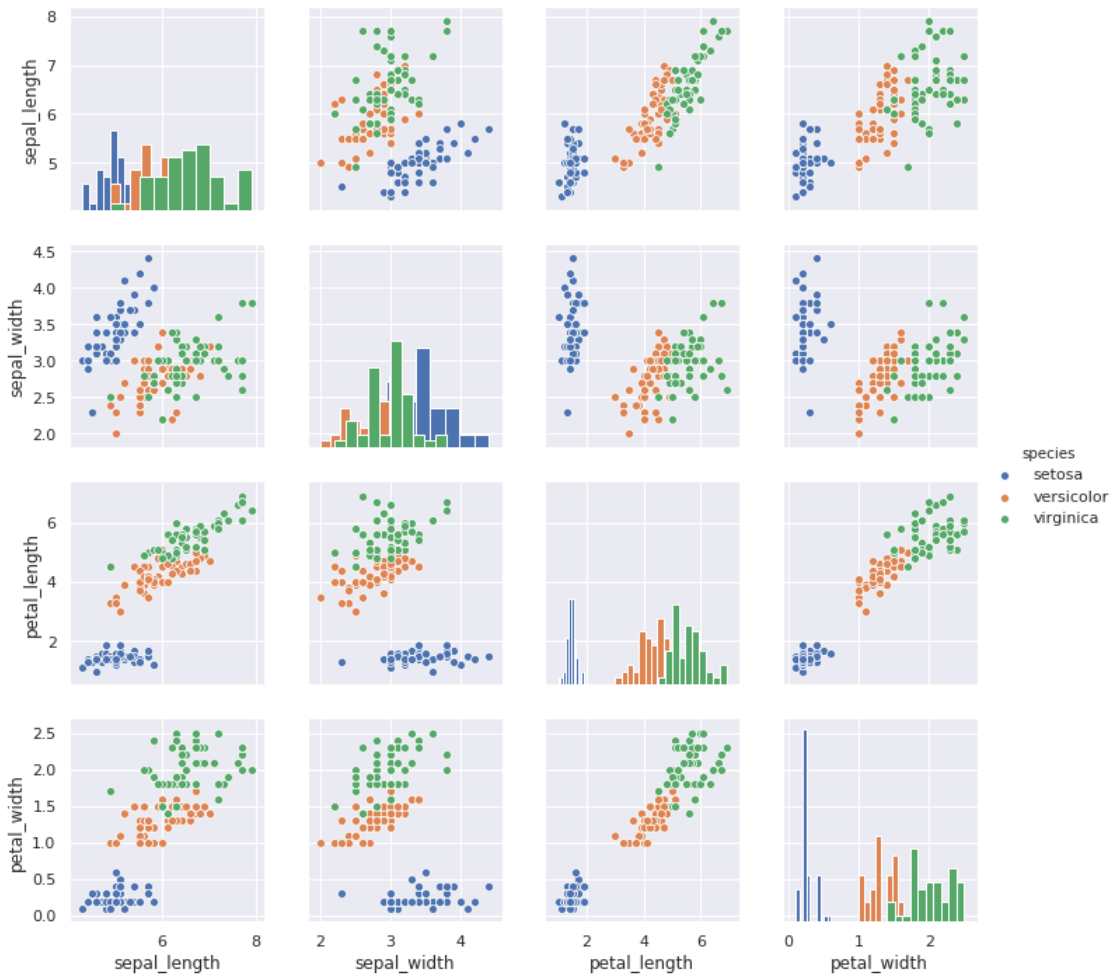
[5]:

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

0.2 Visualisierung der Daten

- Wie??? Problem: Vierdimensionale Features + Eindimensionale Labels…
- Mögliche Lösung: Plotte alles gegen alles

[6]:
```
%matplotlib inline
import seaborn as sns; sns.set()
sns.pairplot(iris, hue='species', diag_kind='hist', height=2.5);
```



Extrahiere aus dem `DataFrame` die Features Matrix `X`und den Labelsvektor `y`.

[3]:
```
X_iris = iris.drop('species', axis=1)
X_iris.shape
```

[3]: (150, 4)

[4]:
```
y_iris = iris['species']
y_iris.shape
```

[4]: (150,)

[0]:
```

```