DataCraft

3.5 KI & Machine Learning

DataCraft

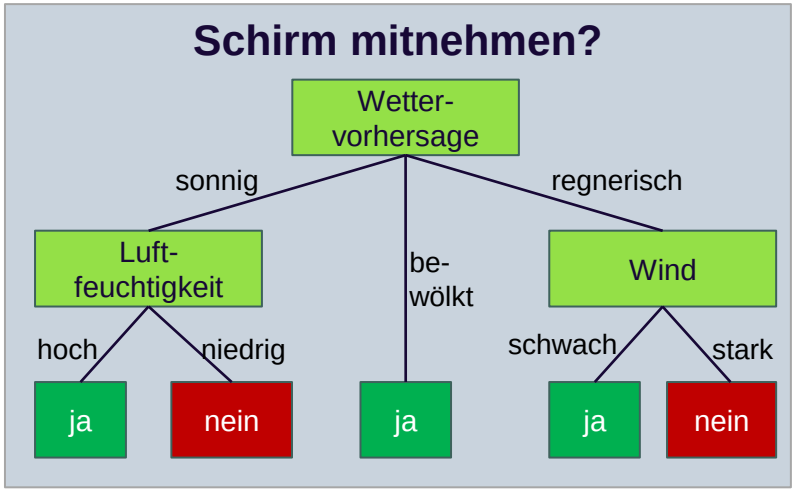
4

Entscheidungsbäume & Random Forest

- Ein Entscheidungsbaum besteht aus Knoten, die hierarchisch angeordnet sind. Der oberste Knoten wird Wurzel (root) genannt. Die Knoten, bei denen es nicht mehr weitergeht, werden Blätter (leafs) genannt.

- Jeder Knoten entspricht einer Entscheidung anhand der gegebenen Daten. Bei einem Blatt steht dann zum Beispiel die Klasse fest.

- Entscheidungsbäume können sowohl für Klassifikation als auch Regression verwenden werden.



- Entscheidungsbäume gehören zur Klasse des überwachten Lernens, da anhand eines Datensatzes mit Lösungen der Baum aufgebaut wird.

- Entscheidungsbäume neigen zum Overfitting. Es gibt verschiedene Methoden, dem entgegenzuwirken, z.B. Pruning (Beschneiden der Äste). Ein einfaches Verfahren ist, einen Knoten durch die größere Klasse zu ersetzen und dann zu schauen, ob sich damit die Güte des gesamten Baums ändert.

- **CART** steht für „Classification and regression trees“ und ist der Oberbegriff für eine ganze Reihe von Algorithmen, der bekannteste ist **C4.5**.

- Random Forest ist ein Ensemble-Algorithmus, d.h. es werden mehrere Modelle erzeugt (hier Entscheidungsbäume) und dann die Ergebnisse der einzelnen Modelle miteinander verglichen. Meistens wird einfach das Ergebnis gewählt, das am häufigsten genannt wird

- Ein wichtiger Vorteil ist, dass der Algorithmus gut parallelisierbar ist, da jeder Baum unabhängig von den anderen erstellt werden kann

- Scikit-Learn hat das Untermodul tree mit mehreren Entscheidungsbaum-Algorithmen. Deren Klassifikationsalgorithmen gehören zur Klasse DecisionTreeClassifier.
- Mit der Funktion plot_tree() lässt sich der Baum visualisieren
- Der RandomForest-Algorithmus ist im Untermodul sklearn.ensemble, die Klassifikationsvariante heisst RandomForestClassifier()