8. ML Introduction Exercise (06.06 and 5 points)

Read the Pedro Domingos article mentioned in the ML Introduction exercise at the end thoroughly and answer all the questions there (5 points)

1. Read this paper by Pedro Domingos very carefully sentence by sentence! A Few Useful Things to Know about Machine Learning

- What is the meaning of ***generalization*** in the paper?

Die Generalisierung erfolgt aus den Beispielen im Trainingsset. Die Generalisierung ist das Ziel, des Machine Learnings. Generalisieren, ist das Lernen von Merkmalen, die für alle Beispiele einer Kategorie gelten. Z.B. erkennen einer Pflanze an ihren Attributen.

- What are the **many faces of overfitting**?

Wenn unser Wissen und die vorliegenden Beispieldaten nicht ausreichend sind um eine richtige Klassifikation zu bestimmen spricht man von „overfitting“. Es fehlt die Verankerung in der Realität. Bei den Trainingsdaten ist die Klassifikation dann 100% genau, bei den Testdaten allerdings eher 50%. Overfitting bedeutet übersetzt Überanpassung und bezeichnet eben die Überanpassung von Algorithmen bzw. deren Parameter an die beobachteten Daten. Im Machine Learning Kontext bedeutet das, der Algorithmus lernt im Prinzip den Datensatz „auswendig“, erkennt aber nicht das zugrundeliegende Muster oder System. Damit sind Prognosen, die der Algorithmus aus noch unbekannten Daten liefern soll, nicht sonderlich gut.

- Why do humans have problems in **higher dimensions**?

Unsere Intuitionen, die aus einer dreidimensionalen Welt stammen, gelten oft nicht in hochdimensionalen Raum. In hohen Dimensionen ist es schwierig zu verstehen, was vor sich geht. Das wiederum macht es schwierig, einen guten Klassifikator zu entwickeln.

- What is **feature engineering**?

Als Feature Engineering bezeichnet man die Vorbereitung von Daten für die Verarbeitung in Machine Learning Algorithmen.

- Why does **more data beats clever algorithms**?

Der schnellste Weg zum Erfolg sind Zyklen. Statt der Auswahl der besten Variationen haben Forscher herausgefunden, dass die Kombination vieler Varianten oft zu viel besseren Lösungen führt. Bei der einfachsten Technik, dem sogenannten Bagging, werden einfach zufällige Variationen der Trainingsmenge durch Resampling erzeugt und klassifiziert.

- What is **ensemble learning**?

Ensemble-Learning kombiniert mehrere Basismodelle um ein optimales Vorhersagemodell zu erstellen.

- What is **accuracy in data science**?

Die Genauigkeit oder Accuracy ist definiert als der Prozentsatz der richtigen Vorhersagen für die Testdaten. Sie kann einfach berechnet werden, indem die Anzahl der richtigen Vorhersagen durch die Anzahl der Gesamtvorhersagen geteilt wird.