Aufgabe3

November 29, 2018

1 Aufgabe 3 - Datenaufbereitung

1.1 Teilaufgabe a)

Nicht-numerische Datentypen, wie zum Beispiel Strings, sollten vor der Analyse wenn möglich in numerische Datentypen (Zahlen) umgewandelt werden, um mit ihnen besser arbeiten zu können. Dafür sind Regeln zu finden, nach denen zum Beispiel jedem Buchstaben eindeutig eine Zahl zugeordnet wird. Falls es nicht zwingend erforderlich ist, es keine Vereinfachungen im Code oder eine bessere Laufzeit zur Folge hat, kann man theoretisch auch mit den nicht-numerischen Datentypen weiterarbeiten.

1.2 Teilaufgabe b)

Es kann hilfreich sein, Attribute zu normieren. Wenn sich Attribute in ihren GröSSenordnungen stark unterscheiden, können viele Classifier versagen, da Abstände oft euklidisch berechnet werden. Dann wird einem Attribut mit groSSen Werten mehr Bedeutung zugemessen als einem Attribut mit kleinen Werten, obwohl die GröSSenordnungen der Attribute im Allgemeinen nichts über ihre Wichtigkeit aussagt. Normierte Attribute tragen dann in etwa gleich zur Abstandsmessung bei.

1.3 Teilaufgabe c)

Wie ist mit Lücken oder Infs und Nans in den Daten zu verfahren? Zuerst: Man sollte alles auf Nan konvertieren, um auch die Lücken und fehlenden Daten zu standardisieren, falls man Sie untersuchen möchte. Die brutale und einfachste Möglichkeit ist es, die fehlenden/fehlerhaften bzw. Attribute vollständig aus den Daten zu entfernen. Das ist sinnvoll, wenn der Anteil an Nans gering ist und man abschätzen kann, dass man keine Information verliert, wenn man diese Daten wegschneidet (auch fehlende Daten sind eine Information!). Es ist auch möglich, die Nans zu analysieren. Insbesondere kann es sinnvoll sein, Korrelationen in Bezug auf fehlende Daten zu analysieren; man kann zum Beispiel die Frage untersuchen, ob bei einem Sample von Menschen aus aller Welt das Fehlen eines Geburtsdatums mit einer Herkunftsregion korreliert. Ein interessantes Package dafür ist https://github.com/ResidentMario/missingno.

1.4 Teilaufgabe d)

Beim Zusammenführen von Datensätzen ist zu beachten, dass diese den gleichen Datentyp haben sollten. Insbesondere bei der Verwendung von Numpy Arrays ist das sogar eine Notwendigkeit. Des Weiteren kann es sinnvoll sein, die Daten zu labeln und das Label gleich mit ihm kombinierten

Array zu speichern, um später seine Analysen kontrollieren zu können. AuSSerdem sollte man auch nachdenken; die Zusammenführung und gemeinsame Untersuchung von Datensätzen, die eine verschiedene Anzahl an Attributen aufweisen, ist wohl nur selten sinnvoll.

1.5 Teilaufgabe e)

Vor dem Trainieren des Klassifizierers sind unwichtige, auch redundant genannte Informationen bzw. Attribute zu entfernen. Dies wird als Dimensionsreduktion bezeichnet. Dabei gibt es Methoden, die die Daten unverändert lassen und Attribute wegschneiden. Dies kann von Hand bzw. "mit dem Auge" geschehen. Es existieren auch spezielle Verfahren zur "Feature Selection", zum Beispiel kann man mit Entscheidungsbäumen Schnitte finden, die den grösSten Informationsgewinn liefern und dadurch unter Anderem Attribute binär machen oder andere, die wenig Informationsgewinn liefern, weglassen. Methoden, die die Daten transformieren, fallen unter "Feature Selection". Dabei werden die Daten in einen Unterraum projeziert und es wird versucht, redundante Informationen wegzulassen. Dies kann zum Beispiel durch das Maximieren der Intervarianz zwischen Klassen und das Minimieren der Intravarianz innerhalb der Klassen geschehen. Dieses Prinzip verwendet die Lineare (Fisher-)Diskriminanzanalyse (LDA). Ein anderes Verfahren ist zur Feature Selection ist die bereits bekannte Hauptkomponentenanalyse (PCA). Möchte man die Dimension von Daten aus einer Simulation reduzieren (ohne direkten Bezug zur Realität), ist das Beachten der theoretischen Hintergründe hinter der Simulation sehr wichtig. So kann man zum Beispiel bei Molekulardynamiksimulationen (MD-Simulationen) die Daten vorverarbeiten, indem zum Beispiel bestimmte Abstände auf 0 transformiert werden. Dadurch erhält man dünnbesetzte Matrizen, die vielseitige Vorteile bei der Berechnung liefern. Möchte man mit Simulationen ein Trainingsset erstellen, um dann gemessene Daten zu klassifizieren, ist es sehr wichtig, systematische Fehler klein zu halten. Damit sind Abweichungen der Simulation von der physikalischen Realität gemeint. Erfasst man mit der Simulation wichtige Details nicht, so kann es vorkommen, dass scheinbar redundante Attribute weggeschnitten werden, die jedoch in Wahrheit wichtig sind oder scheinbar sehr wichtige Attribute sind in Wahrheit eher unwichtig. Daher ist die kritische Betrachtung der Simulation wichtig, da man nur mit gemessenen Daten oft nicht die redundanten Informationen erkennen kann, weil die Klassenzugehörigkeit der gemessenen Ereignisse grundsätzlich unbekannt ist (Sie sind nicht gelabelt).