Grundlagen des Maschinellen Lernens

M. Sc. Jonas Schneider

Übung 3 - Klassifikation auf dem MNIST-Datensatz

Besprechung am Mi, 16/23.11.2016, 12:00 Uhr in der Übung

Aufgabe 3.1: MNIST

Der MNIST (Mixed National Institute of Standards and Technology database) Datensatz ist ein Standard-Benchmark, das im Bereich der Klassifikation seit vielen Jahren als Maßstab für den Vergleich verschiedener Methoden dient. Der Datensatz besteht insgesamt aus 70000 Datenpunkten. Jedes Datum besteht dabei aus einem grauwertcodierten Bild einer Ziffer als Eingabe und der korrekten Klasse, sprich der vom Schreiber beabsichtigten Ziffer, als Ausgabe. Die Eingabe liegt also linearisierter Vektor der Länge 784 der Pixel vor. Dadurch, dass der Datensatz alle arabischen Ziffern (0-9) enthält, liegt somit ein Multi-Klassen-Klassifikationsproblem mit 10 Klassen vor.
Ihre Aufgabe ist es, ein lernfähiges System zu entwerfen, das in der Lage ist, für einen gegebenen Input verlässlich festzustellen, welche Ziffer vorliegt. Nutzen Sie hierfür als Hypothesenmenge das Perzeptron und erzeugen Sie die Multi-Klassen-Klassifikation mit Hilfe der aus der Vorlesung bekannten Verfahren. Da der naive Eingangsraum $d = 784$ für ein erfolgreiches Training viel zu groß ist, muss dieser durch die Erzeugung geeigneter Feature reduziert werden. Welche Feature Sie wie dafür erzeugen, steht Ihnen völlig frei. Ebenso können Sie einen beliebigen Lernalgorithmus für das Perzeptron wählen (Bsp.: Perzeptron-Lernalgorithmus, Adaline, ...).
Für das Training Ihres lernfähigen Systems stehen Ihnen zu Beginn 40000 Trainingsdaten im Stud.IP zur Verfügung. Ihren fertigen Klassifikator, d.h. bestehend aus Ihrer Feature-Extraction und dem Perzeptron-Gewichtsvektor, senden Sie bis Dienstag, den 15.11., 12:00 an den Veranstaltungsleiter. Ihr Algorithmus wird dann auf 20000 Testdaten ausgewertet. Bereiten Sie für den Mittwochstermin eine kurze Präsentation (ca. 10 Minuten) vor, die Ihre Feature-Extraction, Ihren Lernalgorithmus und Ihre Ergebnisse vorstellt. Im Anschluss an diese Präsentationen, wird der Score, den Ihr Ansatz auf den Testdaten erzielt hat, bekannt gegeben. Nach der Übung erhalten Sie die Testdaten und Sie haben erneut eine Woche Zeit, Ihr lernfähiges System zu verbessern. Schicken Sie wieder bis Dienstags, den 22.11. Ihren Klassifikator an den Veranstaltungsleiter und bereiten Sie sich darauf vor, Mittwochs über Ihre Änderungen/Erweiterungen Ihres Ansatzes zu berichten. An diesem Termin wird dann ebenfalls der finale Score auf den letzten 10000 Trainingsdaten bekannt gegeben.
Es ergibt sich also folgender Zeitplan:

15.11. 12:00	Einsenden Ihres ersten Klassifikators
16.11.	Präsentation Ihres Ansatzes und Bekanntgabe des 1. Scores
22.11. 12:00	Einsenden Ihres überarbeiteten Klassifikators
23.11.	Präsentation Ihrer Überarbeitung und des 2. Scores

P.S.: Die bisher beste Klassifikationsleistung auf dem Datensatz liegt bei einer Fehlerrate von 0.23%
P.P.S: Für das Einlesen und den Umgang mit dem Datensatz finden Sie ein Beispielprogramm in Stud.IP.