Vorbereitung für das Seminar „Einführung ins Maschinelle Lernen: Hype oder Hybris“

Marvin Kastner

2. Februar 2020

In diesem Dokument werden die Schritte aufgezählt, die von den Seminarteilnehmern selbständig *vor* Beginn des Seminars auf dem eigenen Laptop durchgeführt werden müssen. Dies ist notwendig, weil zur Einrichtung der Arbeitsumgebung mehrere Gigabyte Daten heruntergeladen und verarbeitet werden. Falls Sie während des Seminars keinen eigenen leistungsfähigen Laptop zur Verfügung haben sollten, sprechen Sie bitte frühzeitig die Seminarleitung an.

Im Folgenden wird aufgezählt, welche Software und welche Dateien benötigt werden. Hier gibt es nur manchmal eine zeitliche Abhängigkeit, schauen Sie also gerne schon etwas voraus und parallelisieren Sie Aufgaben da, wo dies den Prozess beschleunigt.

Installation der IDE

Im Bereich Maschinelles Lernen und Data Science spielen Jupyter Notebooks eine immer größere Rolle[1]. Hierbei bezeichnen Jupyter Notebooks gleichzeitig das Medienformat als auch die Standard-IDE. Diese kann Jupyter Notebooks öffnen, den enthaltenen Code ausführen, die Visualisierungen nah am Code platziert anzeigen und das Ergebnis auch wieder speichern. Bei der Standard-IDE steht die Seminarleitung bei technischen Problemen umfassend zur Seite.

Anaconda ist eine Python-Distribution, die die IDE *Jupyter Notebook* bereits mit ausliefert. Unter `https://www.anaconda.com/distribution/#download-section` wählen Sie bitte die Version für Python 3 aus. Dies sollte ungefähr wie in Abbildung 1 aussehen.

Falls Sie auf Ihrem PC zwei Accounts, einen Administrator-Account und einen Account fürs alltägliche Arbeiten, verwenden, seien Sie bitte vorsichtig. Die Installation von Anaconda erfordert keine Administratorenrechte. Eine Installation mit Administratorenrechten kann u. U. dazu führen, dass Anaconda nur für den Administrator installiert ist bzw. dass es zu Problemen mit der Rechteverwaltung kommt. In der Vergangenheit hat ein erneutes Installieren von Anaconda dieses Problem dann *nicht* beheben können.

Folgen Sie den Installationsschritten und konsultieren Sie im Fehlerfall offizielle Quelle (wie z. B. `https://docs.anaconda.com/anaconda/install/`) oder Foren.

[1] Jeffrey M. Perkel. Why jupyter is data scientists' computational notebook of choice. URL `https://www.nature.com/articles/d41586-018-07196-1`; and Or Hiltch. Jupyter notebook is the cancer of ML engineering. URL `https://medium.com/@_orcaman/jupyter-notebook-is-the-cancer-of-ml-engineering-70b98685ee71`

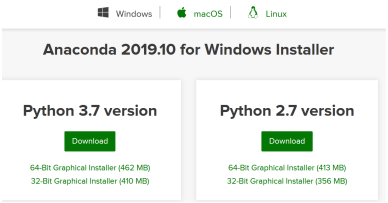

Abbildung 1: Der Download-Bereich von anaconda.

Bezug der Seminar-Materialien

Das Seminar besteht aus einer Zusammenstellung von u. a. Vorträgen, Diskussionen, gemeinsamem Programmieren und Gruppenarbeit. Unter `https://github.com/1kastner/machine-learning-hype-or-hybris` können Sie die Jupyter Notebooks (diesmal ist das Dateiformat gemeint) herunterladen. Die Plattform github erlaubt es, dies auf verschiedene Wege zu tun.

Klonen mit git

Für die Versionsverwaltung ist git quasi der Standard und wird immer häufiger auch außerhalb der Software-Entwicklung eingesetzt. Deswegen lohnt es sich für (fast) jeden, sich Fähigkeiten mit diesem Tool anzueignen. Das Original-Tool ist ein Kommandozeilentool, welches über `https://git-scm.com/` heruntergeladen werden kann. Wer lieber Visualisierungen mag und klicken können möchte, kann sich eine von vielen GUI Clients[2] aussuchen. Hier sollte natürlich neben dem Betriebssystem auch die Lizenz beachtet werden, falls man den Client später noch mal auf der Arbeit einsetzen möchte.

Installieren Sie sich ein Tool Ihrer Wahl, um dann die aktuelle Version des git-Repositorys `https://github.com/1kastner/machine-learning-hype-or-hybris` klonen zu können. Bis zum Start des Seminars werden u. U. die Materialien noch überarbeitet oder erweitert. Aktualisieren Sie also bitte regelmäßig Ihre vorliegende Version.

[2] Eine Liste ist auf `https://git-scm.com/download/gui/win` zu finden.

Download des ZIP-Ordners

Falls Ihnen git unbekannt ist und Sie keine Zeit dafür haben, sich mit git auseinanderzusetzen, gibt es auch die Möglichkeit, den Inhalt als ZIP-Ordner herunterzuladen. Klicken Sie dafür auf den Button, wie er in Abbildung 2 zu sehen ist. Falls Lernmaterialien später noch angepasst werden, ist das Zusammenführen der eigenen und der neuen Version allerdings schwieriger.

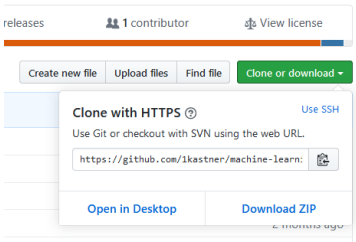

Abbildung 2: Ein github-Repository bietet verschiedene Möglichkeiten zum Bezug der Inhalte an.

Installation der Bibliotheken

Die Jupyter Notebooks, die Sie soeben heruntergeladen haben, benötigen jeweils spezielle Bibliotheken. Einige werden bereits bei der Installation mit Anaconda bereitgestellt, allerdings stehen für das Seminar noch Downloads von weiteren Bibliotheken aus. Auf der obersten Ebene des git-Repositorys befindet sich die Datei `environment.yml`, in der alle benötigten Bibliotheken aufgelistet werden. Diese sollten Sie nun bereits auf dem PC in einem lokalen Ordner Ihrer Wahl

vorliegen haben. Das Format stammt von Anaconda und erlaubt es, die Abhängigkeiten von Bibliotheken automatisch aufzulösen. Damit man an einem PC in verschiedenen Projekten unterschiedliche Versionen einer gleichen Bibliothek haben kann, strukturiert Anaconda Bibliotheken standardmäßig in Umgebungen (eng. *environments*.) Für das Seminar erstellen wir im Folgenden die Umgebung `ml-hype-or-hybris` basierend auf der gegebenen `environment.yml`. Dies kann auf zwei unterschiedlichen Wegen geschehen.

conda

Mit der Installation von Anaconda ist bereits das Kommandozeilentool `conda` vorhanden. Unter Windows wird dies je nach Auswahl während der Installation nicht in die Pfad-Variable mit aufgenommen. Dann steht Ihnen immer noch die *Anaconda Prompt* zur Verfügung. Weitere Informationen sind auf `https://docs.anaconda.com/anaconda/install/verify-install/#conda` zu finden. Eine Übersicht über das Kommandozeilentool ist auf `https://docs.conda.io/projects/conda/en/latest/user-guide/tasks/manage-environments.html#creating-an-environment-from-an-environment-yml-file` einsehbar. Dort wird erläutert, wie eine existierende `environment.yml` eingelesen werden kann. Folgen Sie der Anleitung und erstellen Sie die passende Umgebung.

Anaconda Navigator

Mit Anaconda Navigator kann die gleiche Aufgabe über eine graphische Oberfläche erledigt werden. In der Abbildung 3 sehen Sie unten den Button „Import“. Klicken Sie hierauf und behalten Sie den voreingestellten Datentyp *Conda environment file* bei. Wählen Sie die `environment.yml` aus und schließen den Vorgang ab.

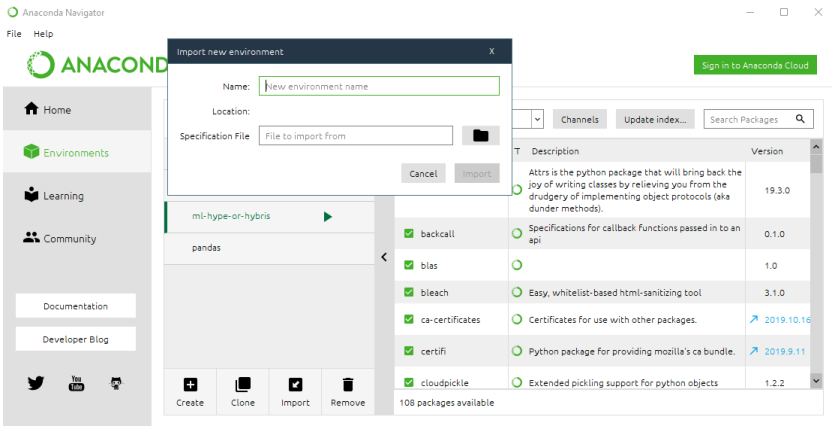

Abbildung 3: Der Anaconda Navigator erlaubt das Importieren von `environment.yml`-Dateien.

Download von Daten

Manche der Jupyter Notebooks verwenden Daten, die bereits von den Bibliotheken bereitgestellt werden. Im Seminar wollen wir aber auch über die Standard-Datensätze hinausgehen.

Verkehrsschilder

Auf `http://benchmark.ini.rub.de/index.php?section=gtsrb&subsection=dataset#Downloads` finden Sie den Link zum Download-Archiv. Laden Sie bitte die Datei `GTSRB_Final_Training_Images.zip` herunter und entpacken Sie das ZIP-Archiv.

Literatur

Or Hiltch. Jupyter notebook is the cancer of ML engineering. URL `https://medium.com/@_orcaman/jupyter-notebook-is-the-cancer-of-ml-engineering-70b98685ee71`.

Jeffrey M. Perkel. Why jupyter is data scientists' computational notebook of choice. URL `https://www.nature.com/articles/d41586-018-07196-1`.