Machine Learning with Text
Leitfaden für die Projektarbeit

Nils Witt

June 12, 2017

1 Einleitung

Die Prüfungsleistung für das Modul "Machine Learning with Text" wird in Form eines Projektes erbracht. Dazu erhält jeder Student eine eigene Forschungsaufgabe. Diese besteht aus diesem Leitfaden, einem Datensatz und einer Aufgabenstellung. Dieser Leitfaden erklärt die Notenvergabe, beschreibt die zu erarbeitenden Dokumente und skizziert den Ablauf des Projektes.

2 Projektablauf

In diesem Kapitel wird der grobe Projektablauf stichwortartig skizziert. Die hier beschriebene Vorgehensweise dient als Orientierungshilfe. Sie sollte also nicht zwangsweise befolgt werden. Wenn es gute Gründe für einen anderen Projektablauf gibt, sollte dieser verfolgt werden:

1. Ausgabe der Forschungsaufgabe
2. Literaturrecherche zur gestellten Aufgabe. Dabei sollten u.a. Antworten zu folgende Fragen gesucht werden: Habe ich das genaue Ziel des Projektes verstanden? Ist mein Datensatz geeignet für diese Art von Aufgabe? Wie haben andere diese Aufgabe bewältigt? Was ist die einfachste Lösung zu meiner Aufgabe? etc..
3. Sichtung des Datensatzes. In welchem Format liegen die Daten vor? Wie kann ich die Daten lesen? Welche Vorverarbeitungsschritte sollte ich durchführen (vielleicht gibt es auffällige Fehlermuster, die leicht beseitigt sind)? Gibt es Möglichkeiten weitere Daten zu sammeln, um bessere Ergebnisse zu erhalten.
4. Entwickeln einer ersten einfachen Arbeitshypothese. Die könnte im Falle einer Sentiment Analysis z.B. lauten:

   > Es gibt Wörter, die hauptsächlich verwendet werden, um negative Stimmungen auszudrücken (z.B. miserable, langweilig, schlecht) und Wörter, die verwendet werden, um positive Stimmungen auszudrücken (lecker, schnell, gemütlich). Außerdem gibt es

noch neutrale Wörter (z.B. Telefon, Wetter, Baum). Wie kann ich die jeweiligen Wörter identifizieren? Nachdem ich das geschafft habe, kann ich die positiv assoziierten und die negativ assoziierten Wörter zählen. Abhängig davon, welche der Summen größer ist, habe ich entweder ein vorwiegend positives Statement oder ein vorwiegend negatives Statement in dem Text.

5. Evaluieren der Hypothese: Wie kann überprüft werden, ob die Arbeitshypothese stimmt? Im vorherigen Beispiel könnte z.B. ein Datensatz gesucht und gefunden werden, der neben den Texten auch von Menschen generierte Labels (Text x hat eine positive Stimmung, Text y hat eine negative Stimmung) enthält. Dieser Datensatz kann verwendet werden, um zu bestimmen, wie gut die These hält.

6. Implementieren der erdachten Verfahrens. Wie repräsentieren ich den Text? Gibt es eine geeignet Visualisierung, damit ich Einblicke in meine Daten erhalte?

7. Sollte die Hypothese nicht bestätigt worden sein, muss eine neue, angepasste Hypothese formuliert und geprüft werden.

8. Wird die These bestätigt, sollten Möglichkeiten erwogen werden die These zu verfeinern. Im Beispiel könnten Wörter z.B: eine Gewichtung erhalten: "Das Essen hat gut geschmeckt" erhält dabei ein geringeres positives Gewicht als "Das Essen hat sehr gut geschmeckt".

9. Der Einfluss der Machine Learning Maschinerie muss überprüft werden. Was passiert wenn ich statt einer simplen Bag-of Words-Repräsentation tf-idf verwende? Gibt es für meine derzeitige Aufgabe einen besseren Classifier als den, den ich derzeit verwende?

10. Das Verfeinern/Anpassen der Hypothese und die Änderungen an der Maschine Learning Maschinerie können natürlich beliebig oft wiederholt werden. Vermutlich erkennt man nach einigen Iterationen Muster in seinen Versuchen. Diese sollten genutzt werden um Schlussfolgerungen zu ziehen (z.B: Verfahren x ist immer besser als Verfahren y. Diese liegt daran, dass...).

Noch einige Hinweise:

- Zu jedem dieser Punkte sollten stets Notizen gemacht werden und ggf. Code geschrieben werden. Jupyter Notebooks erlauben es Notizen und Code nah beinhaltender zu notieren. Das sollte genutzt werden. So entsteht auf natürliche Weise eine "Labortagebuch", welches schon weitestgehend dem Projektbericht, welcher am Ende abzugeben ist, entspricht. So beginnt man nicht den Projektbericht vor einem weißen Blatt Papier (was sehr demotivierend ist).

- Möglichkeiten zum Visualisieren von Daten sollten unbedingt genutzt werden! Oftmals sind Visualisierungen die Grundlage für neue Ideen. Menschen denken sehr visuell, daher formen Visualisierungen das menschliche Denken sehr.

- Wahrscheinlich werden einige praktische Probleme auftreten: z.B: brauchen einige Machine Learning Modelle sehr viel Arbeitsspeicher oder Rechenzeit. Dann muss entweder die Größe des Datensatzes reduziert werden oder es müssen einfachere Modelle verwendet werden. Sollten derartige Entscheidungen gefällt werden müssen, sollte diese unbedingt niedergeschrieben werden.

- Es gibt keine Beschränkung der Hilfsmittel. Alle Forschungsaufgaben basieren auf altbekannten Forschungsfragen, die bereits umfänglich studiert wurden. Daher gibt es reichlich Literatur dazu zu lesen. Diese darf und soll verwendet werden. Hierbei ist auf korrektes Zitieren zu achten. Auch das Reproduzieren anderer Arbeiten ist gestattet, das Kopieren von anderen Arbeiten ist jedoch nicht gestattet!

3 Notenvergabe

Die Forschungsaufgabe dient den Studenten als Ausgangspunkt zur selbstständigen Durchführung eines Forschungsprojektes. Das heißt insbesondere, dass die Forschungsaufgabe die Studenten nicht in ihrer Neugier und Forschungslust einschränkt. Ergebnisse, die über die ursprüngliche Aufgabe hinausgehen werden überaus positiv bewertet, diese werden jedoch keineswegs erwartet. Dabei sollte jedoch das vorgegebene Forschungsziel nicht vernachläßigt werden!
Ob das vorgegebene Forschungsziel erreicht wurde, ist nicht relevant für die Bewertung der Arbeit. Entscheidend ist, dass gezeigt wird, dass dem Forschungsziel akribisch nachgegangen wurde und das verstanden wurde, warum es das Ziel erreicht oder nicht erreicht wurde. Dabei sollten im Falle eines ausbleibenden Erfolgs Ideen für Alternative Forschungsfragen erarbeitet und ausprobiert werden. Die Minimalanforderungen, um das Modul zu bestehen sehen wie folgt aus:

- Der Student hat im Projektbericht und während der Präsentation gezeigt, dass er die gestellte Aufgabe verstanden hat.

- Der Student hat das Forschungsziel erreicht (oder nicht) und versteht warum er es erreicht hat (oder nicht).

- Die Arbeit zeigt, dass der grundlegende Ablauf eines Machine Learning Projektes verstanden wurde (Textvorverarbeitung, Textrepräsentationen, Featureauswahl, Verwendung von supervised und unsupervised Modellen, Visualisierungen, Modelvalidierung etc.).

- Der Projektbericht ist strukturiert (Text und Code) und gibt die Arbeitsweise geeignet wieder.

- Der Code im Projektbericht ist ausführbar (wirft keine bzw. wenige Exceptions).

- Die Präsentation ist strukturiert und verständlich.

- Während der Präsentation kann der Student seine Arbeit berzeugend erläutern.

Die beste Note wird vergeben, wenn zusätzlich zu den Minimalanforderungen noch ein Teil der folgenden Anforderungen erfüllt wird:

- Der Student hat Verfahren, Bibliotheken oder Methoden in geeigneter Weise verwendet, die nicht Teil der Vorlesung waren und hat deren Funktionsweise in seinem Projektbericht korrekt beschrieben. Z.B: ROP Curves, Standard Errors und Confidence Intervals, Crossvalidation, Precision and Recall, F1-Measure, Neuronale Netzwerke, Scikit-Learn's `Pipeline`-Klasse, automatische Hyperparameter Optimierung, Hierarchisches Clustering, Bokeh, Seaborn, etc..
- Der geschriebene Code ist sehr leserlich und effizient.
- Komplexe Codestrukturen sind kommentiert, sodass der Zweck des Codes leicht verständlich wird.
- Es werden Visualisierungen präsentiert, die einen komplexen Sachverhalt leicht verstädlich machen. Dazu sind insbesondere interaktive Visualisierungen geeignet.
- Die Präsentation ist sehr klar strukturiert, unterhaltsam, vermittelt die wichtigsten Erkenntnisse der Arbeit, verzichtet auf unnötige Details und ist auch für nicht-Experten verstädlich.
- Texte und Codes sind so kurz wie möglich, angesichts der enthaltenen Informationen:

  > "Perfection is finally attained not when there is no longer anything to add but when there is no longer anything to take away."
  > - Antoine de Saint-Exupery

- Die zur Überprüfung der Hypothesen verwendeten Verfahren werden korrekt verwendet.
- Andere, hier nicht explizit genannte Eigenschaften der Arbeit, die diese als ausgewönlich kennzeichnen.

4 Abgaberelevante Dokumente

Zum Nachweis der geleisteten Arbeit wird ein Projektbericht erstellt. Dieser soll sowohl ausführbaren Code als auch schriftliche Erläuterungen enthalten. Idealerweise wird ein Jupyter Notebook verwendet. Der Projektbericht umfasst das gesamte Projekt und nicht bloß Teilaspekte. Der Projektbericht führt durch das Projekt, enthält Arbeitshypothesen und deren Nachweise bzw. Widerlegungen (in Form von ausführbarem Code). Es wird sowohl über Misserfolge als auch über Erfolge berichtet.
Weiterhin wird eine 15 minütige Präsentation erarbeitet, die die wichtigsten Ergebnisse der Arbeit vorstellt, gefolgt von einer 15 mintigen Diskussion über den Inhalt der Prsentation. Dabei soll davon ausgegangen werden, dass das Publikum den Projektbericht nicht kennt. Ziel der Präsentation soll es sein, die eigene Arbeit so interessant darzustellen, dass das Publikum dazu angeregt wird, den Projektbericht zu lesen. Die Präsentation sollte mindestens folgende Punkte beinhalten:

- Erläuterung des Projektziels
- Vorstellung des Datensatzes
- Arbeitshypothesen
- Idee zur Überprüfung der Arbeitshypothesen
- Aufgetretene Probleme und Umgang mit den Problemen
- Endgültige Lösung
- Zusammenfassung

5 Rückfragen

Es liegt in der Natur von Foschungsarbeiten das es Situationen gibt, in denen man ein Problem nicht lösen kann (oder glaubt es nicht lösen zu können). In solchen Situationen sollte Hilfe von außerhalb gesucht werden. Dabei soll folgende Reihenfolge eingehalten werden:

1. **Internet:** So ziemlich jedes Problem, das im Rahmen der Forschungsprojekte auftreten wird, hat schon irgendjemand gehabt und darüber im Internet geschrieben. Dieser Beitrag soll gesucht werden! Suchmaschinen sind dabei natürlich das wichtigste Werkzeug. Aber auch Communities wie Stackoverflow, Google Groups und Reddit können sehr hilfreich sein. Sollte das Problem tatsächlich neuartig sein (was sehr unwahrscheinlich ist), kann die Frage auch direkt an eine der gerade erwähnten Communities gerichtet werden.

2. **Kommilitonen:** Mitstudierende werden wahrscheinlich ähnliche Probleme haben. Sich darüber auszutauschen kann sehr hilfreich sein.

3. **Dozent:** Sollten die ersten beiden Schritte zu keiner Lösung führen, kann natürlich der Dozent gefragt werden. Dazu sollte eine E-Mail an `nils@tmfw.de` geschickt werden.

In jedem Falle gilt folgendes: Wenn jemandem (einer Community, einem Kommilitonen oder dem Dozenten) ein Problem angetragen wird, sind gewisse Regeln zu beachten. Diese sind am bei Stackoverflow zusammengefasst:

`https://stackoverflow.com/help/how-to-ask`

Werden diese Regeln eingehalten, steigt die Chance eine hilfreiche Antwort zu erhalten. Gelegentlich stößt man beim ausformulieren der Frage und der damit verbundenen Recherche auf die Lösung. Weiterhin gilt, dass rechtzeitiges Fragen notwendig ist. Grundlegende Fragen sollten nicht zwei Tage vor Abgabe der Arbeit gestellt werden!
Explit ausgeschloßen von diesem Verfahren sind formale Fragen oder Fragen die zur Klarstellung der Aufgabenstellung dienen.