Research proposal

Machine Learning Project

Antonio Toral
Information Science

week 5

Research Proposal

- In a group of 3 students (2 also possible), same as for the project
- It concerns the *proposal*, not the *research* itself (yet)
- Use LaTeX, Word, LibreOffice or Openoffice
- 1-2 pages (700-1000 words)
- Submit PDF (not DOC!) via Nestor (see Course Documents)
- Deadline 21 March (end week 7). Feedback on 23 March.
  - If you submit by 18 March you'll get feedback on 19 March

Topics

The topic you choose should be in line with the lecture materials*:

- (text) classification (regression also possible)
- (document) clustering
- recommender systems

*If you're interested in another topic related to Machine Learning, that may be fine but you need to check with me beforehand

Requirements for the research proposal

Required components

- Problem statement, i.e. topic, research questions, objective
- Discussion of the relevant literature, i.e. the state-of-the-art in the topic
- Material/Methodology/Evaluation
- Bibliography

Topic

Choosing your topic

- Read the literature for the course
- Think about a possible dataset
  - Standard (public) datasets. E.g. Kaggle, github, ...
  - Gather it yourself. E.g. recommender data (purchases, Facebook/Twitter networks, search behaviour, ..). But do not spend too much on data collection: that's not the focus!

Some pointers

- Explore the research area (what is the state-of-the-art? Are there open questions? etc)
- Choose a feasible topic
- Delimit the topic (focus)
- Define and explain the key concepts
- (Explain your choice to your supervisor)

Literature Search

Google Scholar

- Scholar = scientific literature
- Search using keywords
- Search for a ‘classic’ paper and check which papers cite it (link ‘cited by’)

Some possible topics

With thanks to Gosse Bouma, Wilbert Heeringa, Lasha Abzianidze and Leonie Bosveld-de Smet

Sentiment analysis

Twitter Sentiment Analysis

`so so upset because school is coming :(`

`Yum Yum. Mom used to make Mango Jam for me every summer. :)`

- Is a tweet positive or negative?

Challenge

- Short texts (140 tokens $\approx$ 20 words)
- But vast amounts of data available
- Sentiment analysis of tweets: use tweets with emoticons as training data (distant supervision)

Classification of personality traits

- Predict the Big 5 (openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism) on the basis of application and motivation letter
- Challenge: content words are not the best predictors
- Data: 9,000 motivation letters and personality tests

On-line shopping reviews

Amazon product data

- Publicly available dataset (http://jmcauley.ucsd.edu/data/amazon/)
- Over 140 million reviews (score 1 to 5) for over 20 product categories
- Possible task: predict the score of a review
  - Features: bag of words of the review
  - For a given product category only
  - 1 to 5 or a more coarse grained classification, e.g. good (4,5), average (3), poor (1,2)
  - Note: the classes have an order: good $>$ average $>$ poor. Therefore regression better than classification.

Another dataset:
https://s3.amazonaws.com/amazon-reviews-pds/readme.html

Classify useful vs useless answers

- StackOverflow data: can be used for investigating knowledge sharing in social networks.
- Example of a classification task: predict whether an answer is well appreciated or not.

For the dataset, contact Leonie Bosveld-de Smet

Entailment

Task

Given two sentences, does one of them:

- Entail the other?
- Contradict the other?
- or is neutral

Example

- A: Kids in red shirts are playing in the leaves
- B: Children in red shirts are playing in the leaves
- A entails B

Links: http://alt.qcri.org/semeval2014/task1/index.php?id=data-and-tools,
https://nlp.stanford.edu/projects/snli/

Age on Twitter

Predict the age of a Twitter user

- Tweets from 2011
- Users with name ending in 19[6789][0-9]
- At least 1000 characters per user available

generation	60s	70s	80s	90s
# users	2,437	4,278	6,538	15,827

Challenge: unbalanced dataset

Age on Twitter

```
Correct: 2112 out of 2898 (72.88 percent accuracy)

- Confusion details, row is actual, column is predicted
  classname    0   1   2    3  :total
0   60.docs 116  92  29    5  :242  47.93%
1   70.docs 100 237  89    .  :426  55.63%
2   80.docs  49 133 407   63  :652  62.42%
3   90.docs  10  18 198 1352 :1578 85.68%
```

Genre classification

Determine genre of films

- Data: Wikipedia, Internet Movie Database, ...
- More than 1 genre per film?
- Definition of genre is not always clear
- Features: Script/subtitles, Actors, Director, ...

Wikipedia classification

Many categories

- Documents classified according to categories in Wikipedia (> 20.000 in nl.wikipedia)
- 'Trainingless' text categorisation: links in the text of Wikipedia articles, use the linked category/ies for classification

Clustering of vacancies

www.amedoo.org, jobs in the energy sector

If you search for a given word or skill, also related terms should show up. For example **Java Programmer** and **Object Oriented developer** or **Talent Aquisition** and **Human resource manager** and **recruiter**

Approach

- Cluster vacancy texts
- Evaluate manually whether related jobs end up in the same cluster
- Evaluate whether the examples above end up in the same cluster
- Search function: search with keyword K returns a set of documents D. In a 2nd step documents that are in the same cluster are also shown.

Clustering of tweets

Repeat this experiment with Dutch data?

- *probabilistic semantic similarity measurements for noisy short texts using wikipedia entities (Shirakawa et al., 2013)*
- Using tweets with only 1 hashtag as training data
- Remove hashtag
- Cluster the data
- Do the clusters correspond with the hashtags?

Clustering of people

Use large annotated corpora

- For all person names: collect contexts where they appear.
- Count words (grammatical relations) in contexts
- Cluster vectors of words
- Evaluate the results: clusters of footballers, politicians, artists?

Recommender

Twitter followers?

- Is it possible to obtain for user X his/her followers and which users he/she follows?
- Collect tweets from X (and from the users he/she follows too?)
- Recommend to X users he/she may want to follow

Some topics previously explored...

- The Big 5: Classification of personality traits
- Am I the Asshole? Moral classification on Reddit
- Twitter sentiment analysis
- Intent classification
- Spam filtering
- Satirical news classification
- Movie recomendation
- Netflix genre prediction
- IMDB score prediction based on tweets
- Titanic survival prediction
- ...

Some topics previously explored...

- Recommending movies based on their title, genre and ratings
- Predicting movie genres based on their title
- Auto Tagging Stack Overflow Questions
- Detecting sarcasm in Dutch tweets using a distantly supervised set
- Predicting wine prices by their description
- Political prediction of microblogs
- Prediction of appearance at medical appointments
- Classifying the genre of a song by their text (lyrics)
- Determining whether tweets are sent in the morning or the afternoon
- Predicting score of Amazon product reviews
- Evaluation of user-based vs. item-based filtering for an anime recommendation system.
- ...

Some topics previously explored...

- Building and Evaluating a Movie Recommender System Using Different Approaches
- Recommending Last.FM artists using different ways of Collaborative Filtering
- Clustering of new and old Dutch tweets
- Predicting author age in the CLiPS Stylometry Investigation Corpus
- Sentiment prediction for Amazon reviews with language-independent features
- Predicting political preference by tweets
- Automatic tagging of TED-talks based on transcripts
- Movie genre prediction based on user ratings
- ...