Data Analysis and Machine Learning

Continuous Assessment 1

Author: Jim Coen

Wednesday, November 25, 2020

Contents

1 Definitions

AI
When humans consider the operation of a computer program as exhibiting intelligence, then that program has Artificial Intelligence.

Machine Learning
This is a type of Inductive Learning which 'learns' the prediction function. It can then make generalizations about new data. The prediction function is a probability estimate.

Deep Learning
In deep learning, the computer program is structured so as to mimic the human brain. This neural network is a replication of the interconnection between neurons via dendrites.

Parametric and Nonparametric models
A parametric learning model simplifies the prediction function to a known form and uses a fixed set of parameters to characterize the data. It learns the coefficients of the function from the training data. A non-parametric model learns the form of the prediction function from training data. They are a balance between constructing a function that best fits the training data and the ability to generalize.

Supervised Learning and its components
In supervised learning, each observation is type-coded with a 'label'. For instance, 'Spam', 'Not Spam' for email data. Each observation has the same set of characteristic 'features'. The algorithm learns the mapping from features to labels and so can make predictions about new observations. The predicted output is a probability estimation. Representations include decision trees, rules and instances. Evaluation using absolute error is robust to outliers while squared error is not. Gradient descent is one optimization technique for numeric outcomes.

Unsupervised Learning and its components
Data does not contain desired outputs or labels. The algorithm must interpret the inherent structure in the data. It finds clusters of similar feature vectors. It is also used in dimension reduction in order to transform the feature vectors into a more concise form. Representations can be clusters or principal components. Similarity measures evaluate true clusters that do not occur by

chance. Optimization seeks to minimise the variance of feature vectors from the cluster centre.

Common Types of Errors
Errors can occur 'in-sample' on training data or 'out-of-sample' on new data. Over-fitting is a problem that arises when the model is tuned to the training data. Tuning on the test set gives an incorrect estimate of the accuracy on out-of-sample data. Error can be due to bias (simplifying assumptions about prediction function) and to variance (sensitivity of the prediction function to specific sets of training data).

2 Case Study: Online approval rating for a political party

> "It takes many good deeds to build a good reputation, and only one bad one to lose it."
> Benjamin Franklin

2.1 Background

Fianna Fáil is in danger of loosing its distinctiveness as a political party. Maintaining and increasing voter approval is vital. At the same, there is major upheaval in Irish society due to the Wuhan virus and Brexit. In order to manage these threats, the public needs to 'buy in' to policy decisions. In addition, the behaviour of high-ranking party members can incite adverse reaction from the public. The party needs to be aware of issues and personalities in order to deal with crises pro-actively. Users and bloggers discuss all these topics in social media posts.

In addition, the reputation of the party leader has a high correlation with overall party image[1]. Consequently, this is the most important dimension of party reputation. Sub-dimensions are leadership ability, decision-making and sentiment towards the person themselves.

[1]Davies, G., & Mian, T. (2010). *The reputation of the party leader and of the party being led.* European Journal of Marketing, 44(3–4), 331–350. https://doi.org/10.1108/03090561011020453

Table 1: Classifying Political Party Reputation

Party Criteria	Reputation Dimension
Identity	Recognition
Policy	Approval
Personality	Sentiment
Leader	Competence

Focus groups have been employed to inform election campaigns and to assess election outcomes[2]. However, there is time lag between an issue arising and assessing public opinion. On the contrary, an online application can provide 'up-to-the-second' insight into public perception of Fianna Fáil and its policy announcements.

2.2 Business Outcome

The desired outcome is improved awareness by Fianna Fáil of its perception among the public. Limiting the spread of the Wuhan virus demonstrates the importance of reputation awareness. It could inform Fianna Fáil of what level of restrictions that the public are willing to adopt. By assessing the level of approval, it could provide insight into areas that need greater explanation.

On the contrary, divergence between reputation and reality is a risk to the operation of large organisations[3].

Granularity in reputation dimension would enable Fianna Fáil to focus on specific aspects party activity, as shown in table 1. It would monitor public response to policy statements, political actions and personal behaviour.

[2]Breitenfelder, U. et al (2004) *The Use of Focus Groups in the Process of Political Research and Consultancy Qualitative Social Research* [Online]. Available at: https://www.qualitative-research.net/index.php/fqs/article/view/591

[3]Eccles, R. et al (2007) *Reputation and Its Risks* Harvard Business Review, February 2007. Available at: https://hbr.org/2007/02/reputation-and-its-risks

2.3 Evaluating business success

Business success occurs when one or all of the reputation dimensions improve. What a reputation management system can do is to make the party more aware of its reputation among voters. In that sense, success comes when the reputation system reflects accurately the general attitude towards the party.

Fianna Fáil has an extensive organisation with party members throughout the whole country. It is especially important to keep these individuals engaged with and promoting party policy. It would be possible to create an internet forum for these members and evaluate party reputation among this cohort. In particular, a reputation management system could detect shifts in opinion. Human judgement has a role to play here. Certain members are aware of the general attitude towards the party. They could confirm whether the reputation system reflects opinion accurately.

Having proved itself with party members, the reputation system could be rolled out for the public.

2.4 Strategic Plan for Machine Learning Adoption

1. It is necessary to educate Fianna Fáil executives of successful Machine Learning (ML) projects.

2. They need to realise the viability of a reputation management system by pointing to the ML model created for the banking sector[4].

3. By presenting the framework of how such a system could work for Fianna Fáil, it would give executives greater confidence in the project. Explaining the concept of supervised learning would give greater assurance.

As in the banking reputation system, training data could be created from historic social media posts. An initial system could classify reputation as

[4]Rantanen, A., Salminen, J., Ginter, F., & Jansen, B. J. (2019). *Classifying online corporate reputation with machine learning: a study in the banking domain.* Internet Research, 30(1), 45–66. Available at: https://doi.org/10.1108/INTR-07-2018-0318https://www.emerald.com/insight/content/doi/10.1108/INTR-07-2018-0318/full/html#sec005

simply 'positive' or 'negative'. The confusion matrix and ROC curve could evaluate such a binary classifier. Later developments would give a ranking based on probability estimates.

It is important to have benchmark data to evaluate the accuracy of an ML algorithm. Fortunately, Kaggle contains the Political Social Media Posts dataset[5].

2.4.1 Current overall state of machine learning readiness.

One researcher has experience of using PySpark over a three month period in 2015 and the same in 2016. However, a system based upon Pandas dataframes and Python libraries would be more feasible. Further training is necessary before this can be achieved.

2.5 Risk Management

Development of the training dataset is critical. There are several concerns:
1. Limited domain knowledge may lead to inaccuracies in the training set.

2. The index or statistic to rank reputation needs to be a true reflection of opinion expressed in a social media post.

3. The reputation system needs to reflect broad public opinion and not just that of a few activists who write a high proportion of posts.

4. Creating the training data may take considerable work effort and time. For Banking Reputation system, there were one researcher and two research assistants manually coding the data.

5. The US presidential election has given rise to the phenomena of the silent Trump voter. These are people who have been castigated by public figures as 'deplorable' and even violently assaulted by antifa. Since antifa is already organising in Ireland, non-response bias may become a problem.

[5] *Political Social Media Posts* Figure Eight and Kaggle. (2016). Available at: https://www.kaggle.com/crowdflower/political-social-media-posts

6. In the political sphere, there will be posts that are highly emotive and even inflammatory.

7. Social media companies are biased towards certain viewpoints and censors alternative opinion[6].

3 Review of SenTube: A Corpus for Sentiment Analysis

3.1 Summary

The authors created a dataset for use as a benchmark in text categorisation and sentiment analysis of social media texts. The application domain is comments of products (tablets and automobiles) on YouTube. It goes beyond per-document sentiment labeling. It distinguishes between comments on the product itself and those concerning the video. Researchers manually annotated comments according to content and sentiment polarity, as shown in table 2.

The paper highlights problems with data from Twitter streams. Twitter does not provide the text fragments themselves, only IDs to those tweets. Researchers who wish to create an annotated dataset of Twitter comments must download the individual tweets. This leads to reproducibility problems since tweets are often deleted later on.

A means of incorporating user profile is important. Those who create videos tend to be experts on a topic and thus, are social "influencers". They are the counterpart of the activist in the political arena.

An interesting finding from the statistical distribution of comment sentiment is that single opinion labels per document are insufficient. Comments can relate to more than one item (both video and product) and have distinct sentiment polarity for each.

[6]Epstein, R. (2018). *Google and Big Tech bias hurts democracy, not just conservatives.* USA Today. Available at: https://eu.usatoday.com/story/opinion/2018/09/13/google-big-tech-bias-hurts-democracy-not-just-conservatives-column/1265020002/

Table 2: Annotating comments with labels

Classification	Sentiment Polarity	Information Content
Product-related	positive	0 - 3 stars
Video-related	negative	
Spam		
off-topic		
non-English		

3.2 Annotation Guidelines

Besides the benchmark itself, the main contribution of Sentube project is the set of guidelines to promote accuracy and consistency in annotations. Annotators have a fixed set of options for selecting labels. These annotations deal only with identity (whether the text is video-related or product-related). Apart from sentiment polarity, labels do not have qualitative dimensions (attractiveness, convenience, durability).

In order to ensure standardisation in labeling, the authors assessed agreement among four annotators working on a sample set of one hundred comments. They measure the level of agreement using Krippendorff's α coefficient. This varies from 0 (perfect disagreement) to 1 (perfect agreement). An α level greater that 0.8 is recommended[7]. The Sentube annotators achieved α levels greater than 0.75 on comment content. However, when dealing with sentiment polarity, α was in the range of 0.6 - 0.7. Authors report that they created "a gold-standard sample, adjusted by the four annotators". They do not mention how disagreement in labeling was dealt with.

At length, the task was given to a single annotator, not a member of the group above. She inspected 208 videos and waded through almost 36 thousand comments!

[7]Krippendorff, K. (2004). *Content analysis: An introduction to its methodology.* Thousand Oaks, California: Sage.

3.3 Machine Learning Techniques

1. Labeling
   The article is primarily concerned with creating a benchmark dataset and how to assign labels appropriate for sentiment analysis. This is in order to facilitate Supervised Learning. The authors note the high variance in social media posts. However, the use of annotation guidelines introduces bias, thus reducing variance.

2. Baseline model
   In addition to a benchmark dataset, it is necessary to report on the accuracy obtained from a baseline model. The authors created a "bag-of-words" model to process the Sentube dataset. This can serve as a standard for independent researchers, working in sentiment analysis, to evaluate the accuracy of their ML algorithm.

3. Multi-Class Classification
   The authors created a multi-class classifier. This means that it can handle both video and text and it classifies along two dimensions, sentiment polarity and comment type.

4. Support Vector Machine
   The Machine Learning algorithm is a Support Vector Machine that produces a linear separation between types of comments including spam. It also sets the linear boundary between positive and negative sentiments.

3.4 Advantages and Disadvantages

3.4.1 Pro

1. Since the source data comes from YouTube, researchers emphasise the stability of the Sentube dataset in comparison to Twitter posts.

2. YouTube comments tend to follow along a particular topic or thread. This is useful for natural language analysis of discussions and conversations.

3. Despite the promise of automation implicit in the term 'Machine Learning', there is a high degree of human input in creating a labeled dataset. Such input is particularly necessary in sentiment analysis. A person can recognise nuances that a machine cannot. The Sentube project is a template for annotating data in this field.

4. Well-formulated annotation guidelines are necessary in application domains not conducive to automatic label assignment.

3.4.2 Con

1. Twitter, Facebook and YouTube are commercial organisations. They may favour companies that pay product placement fees.

2. Sentiment analysis is likely to require finer granularity than polarity alone.

3. Doesn't take account of secondary input form 'likes' / 'dislikes'