

Artificial Intelligence and Machine Learning: Alpha Generation Using Text Mining Techniques on Financial Statements

Rob Cornish, Investment Manager

FactSet Investment Symposium, 11 June 2009

GAM

Contents

GAM

1. Unstructured information, Classification & Machine Learning
2. Literature review
3. Case example
4. Conclusion
5. Q&A

Structured and unstructured information

GAM

- Quantitative investment strategies mostly employ structured information e.g., forecast P/E, stock price return, directors' share dealings available in databases.
- But what about unstructured information? E.g., broker research reports, newspaper reports, media TV coverage, company regulatory trading statements
- Estimated that c.80% of information held within an organization is unstructured (Merrill Lynch 1998 / www.autonomy.com)

GAM



What is the problem we are trying to solve?

GAM

- Problem is one of classification/categorization
- E.g., Take a random email from your inbox. How do you classify it?
 1. Spam
 2. Non-Spam

How is this classification problem solved? (Part 1)

GAM

- Knowledge Engineering
- E.g., “Expert Systems” – replicate the performance of a human expert
- Most common approach until late 1980s
- Problems: Expensive knowledge acquisition, knowledge base maintenance, exceptions to the rules

How is this classification problem solved? (Part 2)

GAM

- Machine Learning
- Model “learns” from pre-classified examples (“Training Set”) and is then able to classify new unseen documents (“Test Set”)
- Rise in popularity (esp. in research literature!) from early 1990s
- Many types of models used
 - e.g., Neural Networks, Decision Trees, Naïve Bayes Classifier, Nearest Neighbour, Support Vector Machines, etc

- The information content of FOMC minutes, Boukus, E., Rosenberg, J., V., 2006
 - Latent Semantic Analysis on 19 years of data
 - Themes detected in FOMC minutes are correlated with current and future macroeconomic and financial market indicators
- Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web, Das, S., R., Chen, M., Y., 2006
 - 145k messages from internet chat rooms on 25 tech stocks
 - Voting scheme based on 5 individual models to create sentiment index
 - Sentiment index Granger causes stock index levels
- Machine Learning and Automated Text Classification, Sabastini, F., 2002
 - Great overview of subject
 - History of field/research, document pre-processing, models, etc

Lack of finance research in this field!

Classification

- What we want to achieve: Presented with a document we wish to best classify it
- Specifically, what is the probability of document j belonging to classification i , or

$$P(\text{Class } i \mid \text{Doc } j)$$

- Extremely difficult to estimate *directly* – far too many different combinations of Doc_j possible!
- However, there is a solution.....

Naive Bayes Classifier

GAM

Easy!

“Naive” assumption of independent terms

$$= \prod_{k=1}^K P(\text{Term}_{kj} \mid \text{Class}_i)$$

$$P(\text{Class}_i \mid \text{Doc}_j) = \frac{P(\text{Class}_i)P(\text{Doc}_j \mid \text{Class}_i)}{P(\text{Doc}_j)}$$

Too hard to estimate but it is the same for all classifications and therefore effectively a constant so we can ignore it

Case study example – UK regulatory trading statements

GAM

- Released at 7am
- Goal: Before the market opens at 8am we want a system to “read” these statements for us and predict what will happen to stock price during the day
- 3 Classifications/Categories....

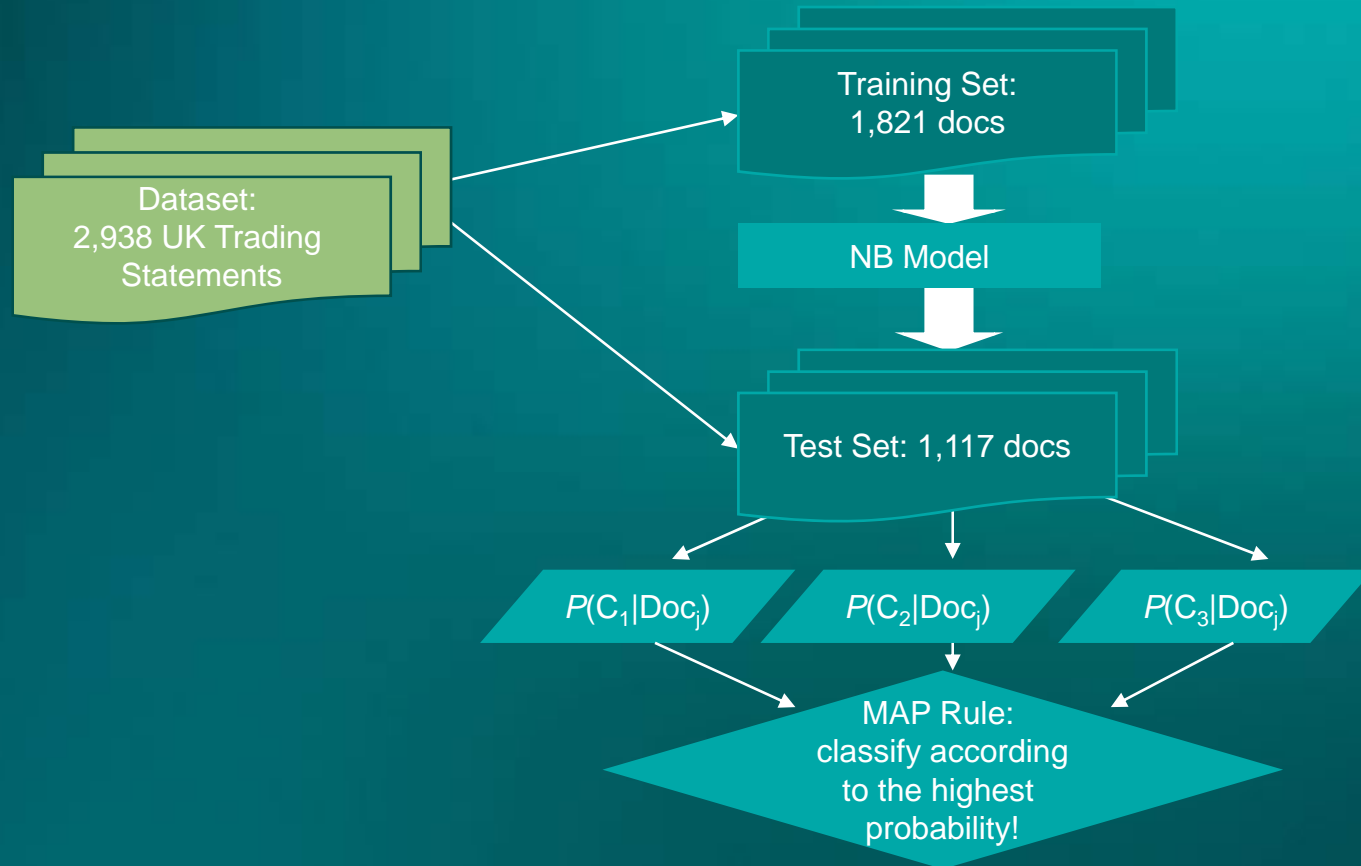
C_1 , “Outperform”: Top 1/3 of stocks ranked on market relative return on day statement is released

C_2 , “In-Line”: Middle 1/3

C_3 , “Underperform”: Bottom 1/3

(Return: Stock return minus FTSE Allshare return from close prior to statement being released to close on the day the statement is released)

Naïve Bayes – Learning and classification



Other things (1)....

- Stemming
 - All words are stemmed by the Porter (1980) algorithm
 - Reduces total number of words found in training document corpus by c.21% (from 22,927 to 18,041)

Original word	Stemmed form
“please”	“pleas”
“pleasing”	“pleas”
“pleased”	“pleas”
“finance”	“financ”
“fishing”	“fish”
“cats”	“cat”

Other things (2)....

GAM

- Term Selection
- “Document Frequency” comparison between top and bottom return quintiles....

Term	% of docs in Q1 containing term	% of docs in Q5 containing term	Difference sorted
“pleas”	71.1%	41.4%	29.8%
“strong”	52.9%	33.6%	19.4%
“ahead”	45.4%	28.8%	16.5%
.....
“impact”	12.4%	36.6%	-24.2%
“lower”	7.2%	35.9%	-28.7%
“below”	6.2%	39.3%	-33.1%

Results – Classified 1,117 docs

GAM

Classification->	Precision			Recall			C ₁ /C ₃ wrong way around
	C ₁	C ₂	C ₃	C ₁	C ₂	C ₃	
Benchmark: Random Classifier	33.3%	33.3%	33.3%	33.3%	33.3%	33.3%	22.2%
Naive Bayes Classifier	49.7%	48.6%	57.6%	69.4%	28.0%	59.1%	14.1%

Precision: Of the documents that the *classifier assigned* to C_i, what percentage were correctly classified? Measure of “exactness”. $TP / (TP + FP)$.

Recall: Of the documents that *genuinely* belong to C_i, what percentage were correctly classified? Measure of “completeness”. $TP / (TP + FN)$.

E.g., Classifier could tell us 10 docs are C₁ and could be correct in 9 cases. So Precision = 90%. This sounds good but if there actually exist 100 documents that truly belong to C₁ then the Recall is 9%, i.e., fairly poor! In this case we could say the classifier has been reasonably “exact” but not very “complete”.

Improvements: Additional Models

GAM

- Centroid
 - Takes into account frequency of term occurrence in a given document (“Term Frequency”) in addition to how many documents contain at least once occurrence (“Inverse Document Frequency”): TF-IDF
 - Average TF-IDF computed from Training Set for each classification: “Centroids”
 - Classify new documents according to nearest Centroid
- Support Vector Machines
 - TF-IDF
 - Maximises the distance (“margin”) between the sets of members and non-members of each classification in multi-dimensional vector space
 - New documents classified in accordance to which side of the “margin” they fall

Results – Classified 1,117 docs

GAM

Classification->	Precision			Recall			C ₁ /C ₃ wrong way around
	C ₁	C ₂	C ₃	C ₁	C ₂	C ₃	
Benchmark: Random Classifier	33.3%	33.3%	33.3%	33.3%	33.3%	33.3%	22.2%
Naive Bayes Classifier	49.7%	48.6%	57.6%	69.4%	28.0%	59.1%	14.1%
Centroid Classifier	51.6%	47.8%	59.7%	54.8%	44.5%	60.2%	12.3%
Support Vector Machine Classifier	51.0%	44.5%	57.0%	57.0%	35.0%	62.3%	12.4%
Classifier Committee	52.4%	47.2%	59.7%	58.9%	39.4%	62.3%	11.7%

Classifier Committee: Majority voting system using Naïve Bayes, Centroid & Support Vector Machine classifiers.

Some possible improvements...

- Use (even) more models!
 - E.g., Neural Networks, Latent Semantic Analysis, Adjective-Adverb, Decision Trees, etc
- More sophisticated models
 - Term covariance. E.g., relax the independence assumption of NB
 - Look at sentences/paragraphs instead of just words. E.g., does a sentence contain contrasting sentiment?
 - Negation. E.g., “pleased” vs. “not pleased”
 - Context sensitive. E.g., Proximity analysis

Brixton PLC

19 Aug 2008 (6.3% underperformance on the day)

GAM

“The **apocalyptic** opening lines of Bob **Dylan**’s

“All Along the Watchtower” seem to capture the beleaguered mindset of the UK commercial real estate market:

There must be some way out of here

*Said the **joker** to the **thief***

There’s too much confusion

I can’t get no relief

Businessmen they drink my wine

***Ploughmen** dig my earth*

None of them along the line

Know what any of it is worth

.... If the “**thieves**” are the funded or equity based opportunist buyers and the “**jokers**” are the owners who won’t sell, there is no “way out” of this impasse – yet.”

Conclusions

GAM

- Significant but limited success with elementary classifiers
- Machine Learning is key
- Contextual analysis is very important
- More research needed to leverage unstructured information in finance

Q & A

Appendix

GAM

Document representation

- Vector Space Model
- Documents parsed out in terms and represented as vectors (in this case binary):

$$\text{Doc}_i = \{ 1 0 0 1 0 1 1 0 0 1 \}$$

Doc number/terms	Strong	Pleased	Below	Upgrade	Ahead	...
1	1	0	0	1	1	
2	0	0	1	0	0	
3	0	0	0	0	0	
4	1	0	1	0	0	
5	0	1	0	0	1	
...						

Some existing applications/products

GAM

- Autonomy Corp Plc. Enterprise Search
- Fast Search & Transfer (Microsoft acquired 2008). Enterprise Search
- NewsScope (Reuters). “Machine readable news” with “sentiment engine” back-testable with tick data
- Monitor110 (Ceased trading 2008). Filtering of online financial information
- NewsCATS. Automated Text Categorization (ATC) prototype using a hand-made thesaurus to forecast intraday stock price trends from information contained in press releases
- Relevance (AOL acquired 2006). Live monitoring/filtering content streams
- ClearForest (Reuters acquired 2007). Bridging the gap between “unstructured text and enterprise data”

Disclaimer

GAM

Source: GAM unless otherwise stated.

The views expressed are those of the manager at the time and are subject to change. This material is intended solely for the use of the reader. It may not be reproduced or distributed to any other person. It is not an invitation to subscribe and is by way of information only. Reference to a security is not a recommendation to buy or sell that security.

This document is issued and approved by GAM London Limited (authorised and regulated by the Financial Services Authority), 12 St James's Place, London SW1A 1NX.

GAM is a member of the Julius Baer Group.