Classical Algorithms in Text Classification

Yash Gandhi

Head of Data Science

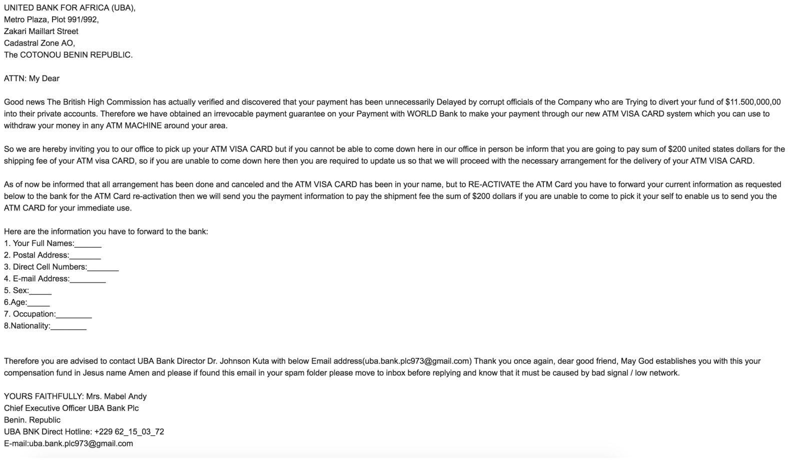
helpshift
DATA SCIENCE

Examples

Examples

- Email classification and spam filtering
- News organisation
- Fraud Detection
- Document classification
- Sentiment analysis
- Recommender Systems

Spam Email



UNITED BANK FOR AFRICA (UBA),
Metro Plaza, Plot 991/992,
Zakari Maillart Street
Cadastral Zone AO,
The COTONOU BENIN REPUBLIC.

ATTN: My Dear

Good news The British High Commission has actually verified and discovered that your payment has been unnecessarily Delayed by corrupt officials of the Company who are Trying to divert your fund of $11.500,000,00 into their private accounts. Therefore we have obtained an irrevocable payment guarantee on your Payment with WORLD Bank to make your payment through our new ATM VISA CARD system which you can use to withdraw your money in any ATM MACHINE around your area.

So we are hereby inviting you to our office to pick up your ATM VISA CARD but if you cannot be able to come down here in our office in person be inform that you are going to pay sum of $200 united states dollars for the shipping fee of your ATM visa CARD, so if you are unable to come down here then you are required to update us so that we will proceed with the necessary arrangement for the delivery of your ATM VISA CARD.

As of now be informed that all arrangement has been done and canceled and the ATM VISA CARD has been in your name, but to RE-ACTIVATE the ATM Card you have to forward your current information as requested below to the bank for the ATM Card re-activation then we will send you the payment information to pay the shipment fee the sum of $200 dollars if you are unable to come to pick it your self to enable us to send you the ATM CARD for your immediate use.

Here are the information you have to forward to the bank:
1. Your Full Names:______
2. Postal Address:_______
3. Direct Cell Numbers:_______
4. E-mail Address:________
5. Sex:_____
6.Age:_____
7. Occupation:________
8.Nationality:________

Therefore you are advised to contact UBA Bank Director Dr. Johnson Kuta with below Email address(uba.bank.plc973@gmail.com) Thank you once again, dear good friend, May God establishes you with this your compensation fund in Jesus name Amen and please if found this email in your spam folder please move to inbox before replying and know that it must be caused by bad signal / low network.

YOURS FAITHFULLY: Mrs. Mabel Andy
Chief Executive Officer UBA Bank Plc
Benin. Republic
UBA BNK Direct Hotline: +229 62_15_03_72
E-mail:uba.bank.plc973@gmail.com

Spam

Old School Approach

- Black listed domains
- Search for commonly used words in spam
- Formatting

Spam

Problems with Old School Approach

- Buying and discarding domains quickly
- Deliberate Spelling mistakes, diacritics, etc
- *What if the email was sent by a real Nigerian Prince?*
- Keep changing formatting to confuse the rules

A Plan for Spam

When I did try statistical analysis, I found immediately that it was much cleverer than I had been. It discovered, of course, that terms like "virtumundo" and "teens" were good indicators of spam. But it also discovered that "per" and "FL" and "ff0000" are good indicators of spam. In fact, "ff0000" (html for bright red) turns out to be as good an indicator of spam as any pornographic term.

-Paul Graham, `A Plan for Spam`, August 2002

BayesTheory

Theorem (Bayes' Theorem)

Bayes' theorem relates the conditional and marginal probabilities of stochastic events A and B. For events A & B with $P(A), P(B) > 0$,

$$\begin{aligned} P(B \mid A) &= \frac{P(A \mid B)P(B)}{P(A)} \\ &= \frac{P(A \mid B)P(B)}{\sum_i P(A \mid B_i)P(B_i)} \end{aligned}$$

where, the sum is over any set of mutually exclusive, exhaustive events, $\{B_i\}$, *with* $P(B_i) > 0$ *for all* i.

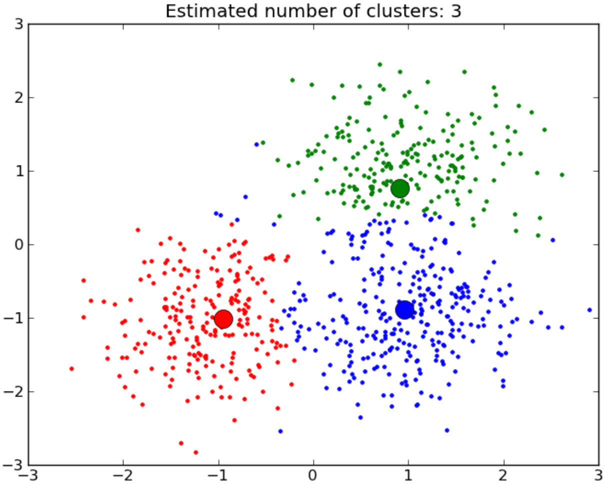
Machine Learning

Machine Learning

We use the Bayes framework for classifying Spam. But first lets talk about basics of Machine learning.
Two broad categories:

- Supervised Classification
- Unsupervised Classification

Clusters



Clusters

For any sort of learning, algorithms require historic data. For the historic data:

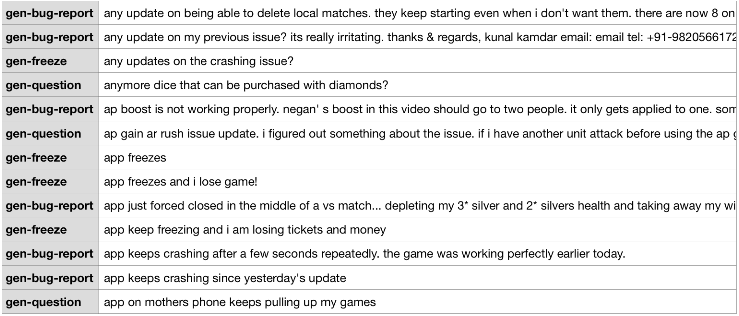
- If we know the labels for the data points apriori and based on the information we try to classify any new data point, then the process is supervised learning.
- If the labels are not known and we are only trying to find a dividing line between data points, then the process is called unsupervised learning.

Supervised Classification for texts

Important terms

- Corpus
- Label
- Training Data
- Validation Data

Corpus



gen-bug-report	any update on being able to delete local matches. they keep starting even when i don't want them. there are now 8 on
gen-bug-report	any update on my previous issue? its really irritating. thanks & regards, kunal kamdar email: email tel: +91-9820566172
gen-freeze	any updates on the crashing issue?
gen-question	anymore dice that can be purchased with diamonds?
gen-bug-report	ap boost is not working properly. negan' s boost in this video should go to two people. it only gets applied to one. som
gen-question	ap gain ar rush issue update. i figured out something about the issue. if i have another unit attack before using the ap g
gen-freeze	app freezes
gen-freeze	app freezes and i lose game!
gen-bug-report	app just forced closed in the middle of a vs match... depleting my 3* silver and 2* silvers health and taking away my wi
gen-freeze	app keep freezing and i am losing tickets and money
gen-bug-report	app keeps crashing after a few seconds repeatedly. the game was working perfectly earlier today.
gen-bug-report	app keeps crashing since yesterday's update
gen-question	app on mothers phone keeps pulling up my games

Näive Bayes Algorithm

Procedure

- Segregate Training and Testing Data set
- Training Dataset
  - Basic Preprocessing
  - Advanced Preprocessing
  - Supervised Classification Model
- Testing Dataset
  - Basic Preprocessing
  - Vocabulary Reduction
  - Classification using model

Basics of Text Analysis

Building Blocks of Text

- Documents
- Paragraphs
- Sentences
- **Words/Tokens**
- Characters

Basics of Text Analysis

Basic Preprocessing

- Remove/replace HTML tags
- Remove/replace URLS/Emails
- Remove/replace Numbers
- Remove/replace accents/diacritics/umlauts (ä,à,ö,ù)
- Tokenisation into words
- Remove Stopwords

Advanced Preprocessing for training

- Stemming
- Lemmatisation
- Spelling Correction
- Vocabulary reduction
  - Using tf-idf
  - Using Statistical significance

After this, we have now generated the vocabulary for our domain. Vocabulary contains all the essential words used in context of the domain.

Building Model

- For each data point in training data, we apply basic preprocessing and remove all the words that are out of vocabulary..
- Each data point is now a list of vocabulary words with its corresponding label. This representation is called Bag-of-words.
- Segregate data points by labels, and for each label find the distribution of words in the label.
- For each label, also store the document count

Model

	$label_1$	$label_2$	$\dots$	$label_K$
w_1	c_{11}	c_{12}	$\dots$	c_{1K}
w_2	c_{21}	c_{22}	$\dots$	c_{2K}
w_3	c_{31}	c_{32}	$\dots$	c_{3K}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
w_V	c_{V1}	c_{V2}	$\dots$	c_{VK}

	$label_1$	$label_2$	$\dots$	$label_K$
$Counts$	n_1	n_2	$\dots$	n_K

For probability of an event L_j conditioned on multiple independent events $W = \{w_1, w_2, \dots, w_K\}$,

$$\begin{aligned} P(L_j \mid w_1, w_2, \dots, w_K) &= \frac{P(w_1, w_2, \dots, w_K \mid L_j) P(L_j)}{P(w_1, w_2, \dots, w_K)} \\ &= \frac{P(w_1 \mid L_j) P(w_2 \mid L_j) \dots P(w_K \mid L_j) P(L_j)}{P(w_1) P(w_2) \dots P(w_K)} \\ &= \prod_{i=1}^{K} P(w_i \mid L_j) \frac{P(L_j)}{\prod_{i=1}^{K} P(w_i)} \end{aligned}$$

The probabilities, $P(w_1)P(w_2)\ldots P(w_K)$ are independent of L_j, hence we can simply replace the term $\frac{1}{P(w_1)P(w_2)...P(w_K)}$ by a proportionality constant, C

$$P(L_j \mid w_1, w_2, \ldots, w_K) = C \prod_{i=1}^{K} P(w_i \mid L_j)P(L_j)$$

Here we are multiplying a lot of small numbers and we may face computational errors due to underflow. Hence we take the log,

$$\log P(L_j \mid W) = \sum_{i=1}^{K} \log P(w_i \mid L_j) + \log P(L_j) + \log C$$

This Quantity is called log-likelihood. And our goal is to find the Label, L_j for which this quantity is maximum.

For prediction

- For each data point in testing data, we apply basic preprocessing and remove all the words that are out of vocabulary.
- Each data point is Bag-of-words.
- For each label, apply log likelihood formula to find out the log probability for the label.

$$\mathcal{L}_j = \sum_{i=1}^{K} \log P(w_i \mid L_j) + \log P(L_j) + \log C$$

$$\mathcal{L} = \{\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_K\}$$
$$m = max(\mathcal{L})$$

Prediction Probability

- So the Individual probabilities for labels:

$$\mathcal{P}(\mathcal{L}) = \{e^{\mathcal{L}_1}, e^{\mathcal{L}_2}, \ldots, e^{\mathcal{L}_K}\}$$

- Label corresponding to Maximum probability is the predicted label for the new data point. But there is a problem here.

Underflow!

- As these are exhaustive probabilities,

$$
\begin{aligned}
\sum_{j=1}^{K} \mathcal{P}(\mathcal{L}_K) &= 1 \\
\implies \mathcal{P}(\mathcal{L}_m) &= \frac{e^{\mathcal{L}_m}}{\sum_{j=1}^{K} e^{\mathcal{L}_j}} \\
\implies \mathcal{P}(\mathcal{L}_m) &= \frac{1}{\sum_{j=1}^{K} e^{\mathcal{L}_j - \mathcal{L}_m}}
\end{aligned}
$$

where $\mathcal{L}_m$ is the maximum value in $\mathcal{L}$