40.016: The Analytics Edge

Week 10 Lecture 2

TEXT ANALYTICS (PART 2)

Term 5, 2022

Outline

Outline

Text Analytics

- Process of (automatically) deriving high-quality information from text – translating large volumes of unstructured text into quantitative data to uncover insights, trends, and patterns.
- Common tasks are
  (1) text categorization and summarization (Wednesday: Enron corpus and Homework: emails)
  (2) sentiment analysis (Monday: tweets)
- Text analytics build on several steps, e.g.,
  - processing the text
  - finding patterns
  - learning a classification model

Outline

The Enron Corporation

- Enron Corporation was an American energy and services company established in 1985 and based in Houston
- It was originally involved in transmitting and distributing electricity and natural gas throughout the US.
- Before its bankruptcy in 2001, Enron employed nearly 20,000 staff, with claimed revenues of about $100 billion (in 2000)
- By the end of 2001, it was revealed that the reported financial conditions were based on systematic accounting fraud (the Enron scandal)

The California energy crisis

- The California energy crisis (2000-2001) was a situation in which California had a shortage of electric supply caused by market manipulations and illegal shutdowns of pipelines.
- The state suffered from multiple blackouts, while Enron gamed the market, which was deregulated.
- The Federal Energy Regulatory Commission (FERC) investigated Enron involvement in the crisis and this ultimately led to a $1.52 billion settlement.

The California energy crisis (cont'd)

- As part of the investigations, FERC released the emails from some of the executives at Enron.
- One approach is to search for keywords like "electricity bid", "energy schedule" in emails, and then carefully review which ones are responsive.
- Predictive coding using text analytics as an alternate and newer way to do this.
- The data come from a 2010 Text Retrieval Conference (attorneys labeled emails responsive to energy schedule or bids).

The Enron Corpus Dataset

- Publicly-available dataset of the email messages sent or received by about 150 Enron senior managers
- “Rare” kind of dataset (This corpus is one of the few publicly-available mass collection of real emails available for study, as such collections are often bond by privacy laws)

The Enron Corpus

Our goal: identify which emails are responsive to content related to energy bids. (0: not responsive; 1: responsive)

- In energy market, electricity suppliers offer to sell electricity for a given bid price.
- Terms like “electricity bid” or “energy schedule” may have enough predictive power to learn an accurate classifier.

To this purpose, we will use **predictive coding**.

Predictive coding

- Predictive coding allows software to take information entered by people and generalize it to a larger group of documents.
- Traditionally, one searches a set of documents to identify documents relevant to a case.
- Rather, predictive coding considers the context.
- Instead of having legal experts going through all documents, one uses a training set, which has been categorized by legal experts.
- Then, the use of text analytics helps predict the categorization of future documents.

Summary

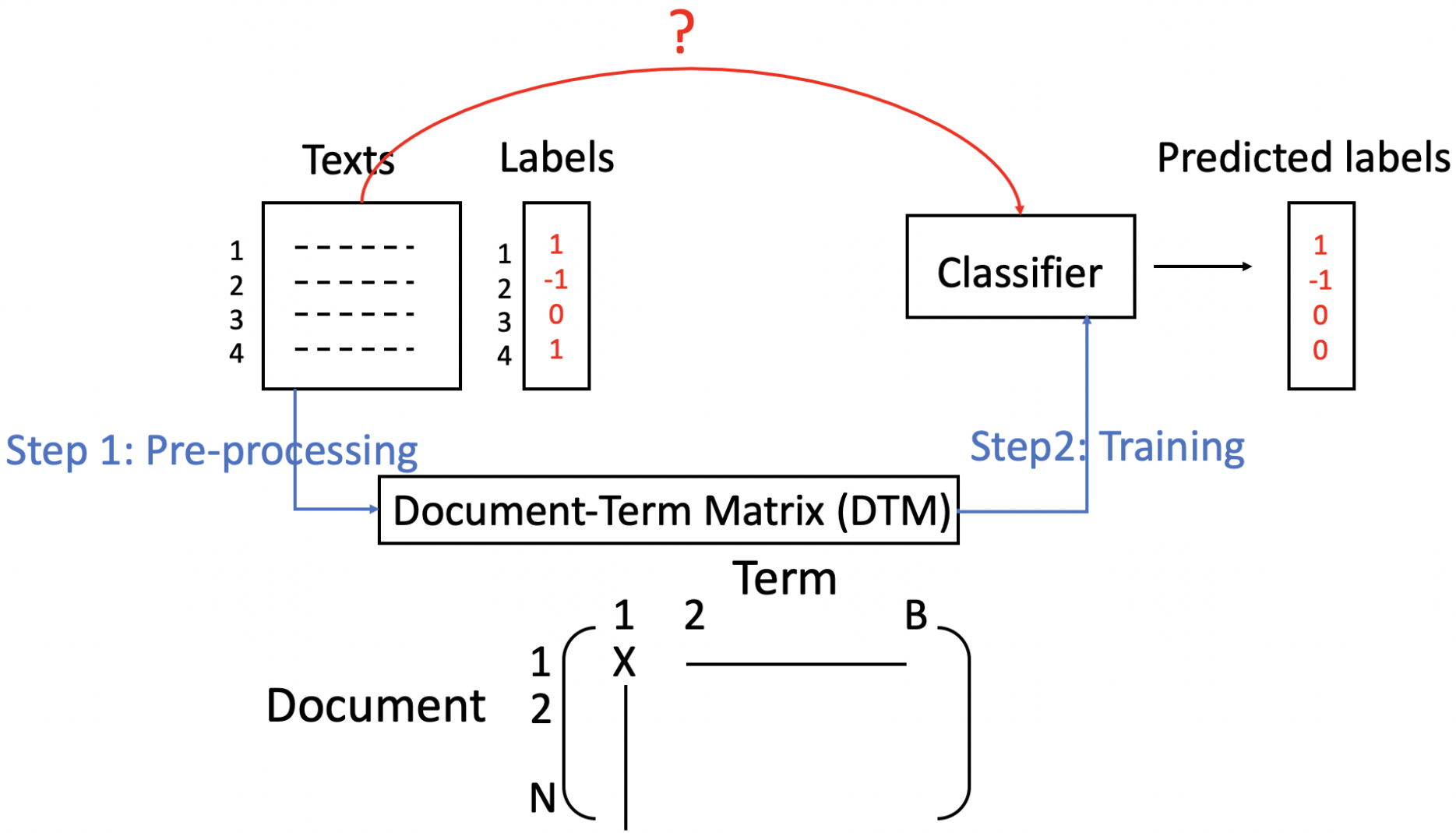
- **Data:** large unstructured datasets containing emails and their corresponding classification (responsive or non-responsive).
- **Model:** A classification model (e.g., logistic regression, CART, Random Forest, Naive Bayes classifier) that predicts the class of an email based on its text.
- **Value and Decision:** the model replaces the expensive option of manually searching all emails.

Outline

Modelling workflow

- Pre-processing
  - Convert text to lower case
  - Remove stopwords
  - Remove punctuation
  - (Remove numbers)
  - Stemming
  - Create DTM
  - Removing sparse terms
- Preparing the DTM for model learning
- Train and test a classifier

Modelling workflow (cont'd)



Outline

Recall Bayes Theorem

- $P(A)$ is the probability of event A
- $P(B)$ is the probability of event B
- $P(A \mid B)$ is the probability of observing event A if B is true
- $P(B \mid A)$ is the probability of observing event B if A is true.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Example 1

Chris Wiggins, an associate professor of applied mathematics at Columbia University, posed the following question:

"A patient goes to see a doctor. The doctor performs a test with 99% reliability – that is, 99% of people who are sick test positive and 99% of the healthy people test negative. The doctor knows that only 1% of the people in the country are sick.

Now the question is: if the patient tests positive, what is the chance that the patient is sick?"

`https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/bs704_probability6.html`

Example 1

Chris Wiggins, an associate professor of applied mathematics at Columbia University, posed the following question:

"A patient goes to see a doctor. The doctor performs a test with 99% reliability – that is, 99% of people who are sick test positive and 99% of the healthy people test negative. The doctor knows that only 1% of the people in the country are sick.

Now the question is: if the patient tests positive, what is the chance that the patient is sick?"

The intuitive answer is 99%, but the correct answer is 50%...."

`https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/bs704_probability6.html`

Example 1 (cont'd)

Take the example of 10,000 people:

	Sick (A)	Healthy	
Test positive (B)	99	99	198
Test negative	1	9801	9802
	100	9900	10000

- $P(A) = 0.01$ is the probability of sickness.
- We know the reliability of the test is 99%, i.e., $P(B \mid A) = 99/100 = 0.99$ and $P(\text{Test negative} \mid \text{Healthy}) = 0.99$.
- What is $P(A \mid B)$?
- From the table above, we can see that given a positive test, the probability of sickness is $99/198 = 50\%$.

Example 1 (cont'd)

- The solution to this question can easily be calculated using Bayes theorem.
- Bayes stated that the probability one tests positive AND are sick is the product of the likelihood that one tests positive GIVEN that he is sick and the "prior" probability that he is sick.

$$P(A \cap B) = P(B \mid A)P(A) \Big(= P(A \mid B)P(B) \Big)$$

$$\Rightarrow \; P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} = \frac{0.99 \times 0.01}{0.0198} = 50\%.$$

- Bayes theorem allows one to compute a conditional probability based on the available information.

Example 2

Suppose a patient exhibits symptoms that make her physician concerned that she may have a particular disease. The disease is relatively rare in this population, with a prevalence of 0.2%. The physician recommends a screening test that costs $250 and requires a blood sample.

Before agreeing to the screening test, the patient wants to know what will be learned from the test, specifically she wants to know the probability of disease, given a positive test result.

The physician reports that the screening test is widely used and has a reported sensitivity of 85%. In addition, the test comes back positive 8% of the time and negative 92% of the time.

`https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_probability/bs704_probability6.html`

Example 2 (cont'd)

	Diseased (A)	Not Diseased	
Test positive (B)	17	783	800
Test negative	3	9,197	9,200
	20	9,980	10,000

- $P(A) = 0.002$
- $P(B \mid A) = 0.85$, i.e., the probability of screening positive, given the presence of disease is 85% (the sensitivity of the test)
- $P(B) = 0.08$, i.e., the probability of screening positive overall is 8%.

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} = \frac{0.85 \times 0.002}{0.08} = 2.125\%$$

Example 2 (cont'd)

- If the patient undergoes the test and it comes back positive, there is a 2.125% chance that she has the disease.
- Also, note, however, that without the test, there is a 0.2% chance that she has the disease.
- In view of this, do you think the patient should have the screening test?

Recall chain rule of conditional probability

- **Two events.** The chain rule for two random events A and B says

$$P(A \cap B) = P(B \mid A) \cdot P(A)$$

- **More than two events.** For more than two events $A_1, A_2, \cdots, A_n$, the chain rule extends to the formula

$$\begin{aligned} P(A_n \cap A_{n-1} \cap \cdots \cap A_1) &= P(A_n \mid A_{n-1} \cap \cdots \cap A_1) \cdot P(A_{n-1} \cap \cdots \cap A_1) \\ &= P(A_n \mid A_{n-1} \cap \cdots \cap A_1) \cdot P(A_{n-1} \mid A_{n-2} \cap \cdots \cap A_1) \cdot P(A_{n-2} \cap \cdots \cap A_1) \\ &= \cdots = \prod_{k=1}^{n} P(A_k \mid \bigcap_{j=1}^{k-1} A_j) \end{aligned}$$

Recall chain rule of conditional probability (cont'd)

With four events ($n = 4$), the chain rule is

$$\begin{aligned} P(A_4 \cap A_3 \cap A_2 \cap A_1) &= P(A_4 \mid A_3 \cap A_2 \cap A_1) P(A_3 \cap A_2 \cap A_1) \\ &= P(A_4 \mid A_3 \cap A_2 \cap A_1) P(A_3 \mid A_2 \cap A_1) P(A_2 \cap A_1) \\ &= P(A_4 \mid A_3 \cap A_2 \cap A_1) P(A_3 \mid A_2 \cap A_1) P(A_2 \mid A_1) P(A_1) \end{aligned}$$

Naive Bayes classifier

- The Naive Bayes classifier is a simple classification rule based on Bayes' rule, which can be used with the bag of words model.
- Given a document to be classified, represented by a vector (a row of DTM), a Naive Bayes classifier assigns to this document probabilities
  - $P(\text{positive} \mid \text{a row of DTM})$
  - $P(\text{negative} \mid \text{a row of DTM})$
  - $P(\text{neutral} \mid \text{a row of DTM})$

Naive Bayes classifier (cont'd)

Given a problem instance to be classified, represented by a vector $\{x_1, ..., x_p\}$ of predictors, a Naive Bayes classifier assigns to this instance probabilities $P(C_k \mid x_1, \cdots, x_p)$ for each of the K possible classes or outcomes C_k.

Using **Bayes' rule**, lets rewrite this as:

$$P(C_k \mid x_1, \cdots, x_p) = \frac{P(C_k)P(x_1, \cdots, x_p \mid C_k)}{P(x_1, \cdots, x_p)}$$

- $P(x_1, \cdots, x_p)$: constant Z, because it does not depend on C_k and the values of $x_1, \cdots, x_p$ are given
- $P(C_k \mid x_1, \cdots, x_p)$: Posterior probability
- $P(C_k)$: Priori probability

Naive Bayes classifier (cont'd)

Let's focus on the numerator $P(C_k)P(x_1, \cdots, x_p \mid C_k)$, which is equal to the joint probability $P(C_k, x_1, \cdots, x_p) = P(x_1, \cdots, x_p, C_k)$. Using the chain rule, we can write:

$$\begin{aligned} P(x_1, \cdots, x_p, C_k) &= P(x_1 \mid x_2, \cdots, x_p, C_k) \times P(x_2, \cdots, x_p, C_k) \\ &= P(x_1 \mid x_2, \cdots, x_p, C_k) \times P(x_2 \mid x_3 \cdots, x_p, C_k) \\ &\times \cdots \times P(x_{p-1} \mid x_p, C_k) \times P(x_p \mid C_k) \times P(C_k) \end{aligned}$$

Naive Bayes classifier (cont'd)

Now, we make a naive hypothesis of conditional independence of the features, that is, x_i is conditionally independent of every other feature x_j (with $i \neq j$). With this hypothesis in place, we can write:

$$
\begin{aligned}
P(x_1, \cdots, x_p, C_k) &= P(x_1 \mid x_2, \cdots, x_p, C_k) \times P(x_2 \mid x_3 \cdots, x_p, C_k) \\
&\times \cdots \times P(x_{p-1} \mid x_p, C_k) \times P(x_p \mid C_k) \times P(C_k) \\
&= P(x_1 \mid \cancel{x_2, \cdots, x_p}, C_k) \times P(x_2 \mid \cancel{x_3 \cdots, x_p}, C_k) \\
&\times \cdots \times P(x_{p-1} \mid \cancel{x_p}, C_k) \times P(x_p \mid C_k) \times P(C_k) \\
&= P(C_k) \cdot \prod_{i=1}^{p} P(x_i \mid C_k)
\end{aligned}
$$

Naive Bayes classifier (cont'd)

Recalling that $P(x_1, \cdots, x_p)$ is a constant Z, we can finally write:

$$P(C_k \mid x_1, \cdots, x_p) = \frac{1}{Z} \underbrace{P(C_k) \prod_{i=1}^{p} P(x_i \mid C_k)}_{P(x_1, \cdots, x_p, C_k)}$$

The prediction rule is to assign a class k such that:

$$\arg\max_{k=1,\cdots,K} P(C_k) \prod_{i=1}^{p} P(x_i \mid C_k),$$

which is estimated from the training set and then applied to the test set.

Naive Bayes classifier (cont'd)

To estimate the $P(x_i \mid C_k)$, one can, for example, use a Gaussian distribution. Denoting with μ_{ki} and σ^2_{ki} the mean and variance of x_i in class k (i.e., x_{ki}), we get:

$$P(x_i \mid C_k) = \frac{1}{\sqrt{2\pi\sigma^2_{ki}}} e^{-\frac{(x_{ki}-\mu_{ki})^2}{2\sigma^2_{ki}}}.$$

x_1	$C_k = \{0,1\}$
2	0
4	0
3	0
5	1
9	1

$$P(x_1 | C_k) = ?$$

$$\mu = \frac{2+4+3}{3} = 3$$

$$\sigma^2 = (2-3)^2 + (4-3)^2 = 2$$

$$\mu = \frac{5+9}{2} = 7$$

$$\sigma^2 = (5-7)^2 + (9-7)^2 = 8$$

Example

Training set

Gender	Height (Foot)	Weight (lb)	Shoe size (Inch)
Male	6	180	12
Male	5.92	190	11
Male	5.58	170	12
Male	5.92	165	10
Female	5	100	6
Female	5.5	150	8
Female	5.42	130	7
Female	5.75	150	9

Testing set

Gender	Height (Foot)	Weight (lb)	Shoe size (Inch)
Unknown	6	130	8

Example (cont'd)

$$\text{Posterior(Male)} = \frac{P(\text{M})P(\text{Height} \mid \text{M})P(\text{Weight} \mid \text{M})P(\text{ShoeSize} \mid \text{M})}{Z}$$

$$\text{Posterior(Female)} = \frac{P(\text{F})P(\text{Height} \mid \text{F})P(\text{Weight} \mid \text{F})P(\text{ShoeSize} \mid \text{F})}{Z}$$

- $P(\text{M}) = \frac{4}{8} = 0.5 \quad P(\text{F}) = \frac{4}{8} = 0.5$
- Assume Height, Weight, Shoe size follow Gaussian distributions.

Gender	Height (Foot)		Weight (lb)		Shoe size (Inch)	
	Mean	Variance	Mean	Variance	Mean	Variance
Male	5.855	0.035	176.25	123	11.25	0.92
Female	5.4175	0.097	132.5	558	7.5	1.67

$$P(\text{Height} \mid \text{M}) = \frac{1}{\sqrt{2\pi \times 0.035}} e^{-\frac{(6-5.855)^2}{2\times 0.035}} \ldots$$

- For the testing data, Posterior(Male) $<$ Posterior(Female) $\rightarrow$ Female!

Back to R!

In R, Naive Bayes Classifier can be carried out with the `e1071` package.

`model <- naiveBayes(formula,data)`

`model$apriori` To see $P(C_1), \cdots, P(C_K)$

`model$tables$`x_i To see mean and standard deviation of $P(x_i \mid C_k)$

`predict <- predict(model,newdata=test,type="class")` Prediction

Advantages and disadvantages

Advantages

- Works with binary and multi-class problems
- Computationally efficient

Disadvantages

- Assumption of conditional independence of the predictors, which may (or may not) lead to poor performance
- Assumption of Gaussian distribution for $p(x_i \mid C_k)$, $i = 1, \cdots, p$

References

- Teaching notes.