Language and Computers

Semester: FALL '17

Instructor: Sowmya Vajjala

Iowa State University, USA

20 October 2017

Outline

1. Recap of last class
2. how does learning happen?
3. reminder: Assignment 4 is due tomorrow!!

What is Text classification?

- If we have some amount of example corpus which has some kind of known categorization, text classification is the method which uses these examples to learn to categorize new texts.

What is Text classification?

- If we have some amount of example corpus which has some kind of known categorization, text classification is the method which uses these examples to learn to categorize new texts.
- Applications: spam filtering, email categorization, reachout.com forum posting severity detection, language identification, checking whether a web page is suitable for children or not etc.

Steps in Text classification?

- We need a collection of example texts with known categories (Training data)
- We need to extract "features" we want the machine to learn from these (feature extraction)
- We should take these extracted features and give them to a "learning algorithm" (training/learning phase)
- Evaluate if the "learned" classifier is doing well by "testing" it with a few more examples with known categories (test data, evaluation)
- If you are happy, start using in some real-world application!!

Recap questions

... with Random questioning

Let us take "spam classification" as our example problem.

- Where can we get the training data for this?

Recap questions

... with Random questioning

Let us take "spam classification" as our example problem.

- Where can we get the training data for this?
- What features are meaningful?

Recap questions

... with Random questioning

Let us take "spam classification" as our example problem.

- Where can we get the training data for this?
- What features are meaningful?
- How do we get these features?

Recap questions

... with Random questioning

Let us take "spam classification" as our example problem.

- Where can we get the training data for this?
- What features are meaningful?
- How do we get these features?
- What happens after feature extraction?

Recap questions

... with Random questioning

Let us take "spam classification" as our example problem.

- Where can we get the training data for this?
- What features are meaningful?
- How do we get these features?
- What happens after feature extraction?
- What according to you is a good way to evaluate the classification?

The attendance question from last class

- Problem: Understanding how to figure out which of those descriptions are by a human, or a machine, or identify unclear cases.
- Logic: Looking at the example descriptions, we should build a notion of whether those adjectives have similar meanings or opposite meanings, and use that knowledge to solve for unknown cases.
- This is an example of text classification, except that we are doing it manually, and with a small set of examples.
- Full description: `http://nacloweb.org/resources/problems/2015/N2015-G.pdf`
- Solution description: `http://nacloweb.org/resources/problems/2015/N2015-GS.pdf`

Steps in Text classification?

- **We need a collection of example texts with known categories (Training data)**
- We need to extract "features" we want the machine to learn from these (feature extraction)
- We should take these extracted features and give them to a "learning algorithm" (training/learning phase)
- Evaluate if the "learned" classifier is doing well by "testing" it with a few more examples with known categories (test data, evaluation)
- If you are happy, start using in some real-world application!!

Let us say this is my "training data"

```
ham     Is that seriously how you spell his name?
ham     I'm going to try for 2 months ha ha only joking
ham     Oops, I'll let you know when my roommate's done
ham     I see the letter B on my car
ham     Sorry, I'll call later in meeting.
spam    URGENT! You have won a 1 week FREE membership in our £100,000 Prize Jackpot!
spam    You are a winner U have been specially selected 2 receive £1000 prize or a 4* holiday (flights inc) speak to a live operator 2 claim
0871277810810
spam    Are you unique enough? Find out from 30th August.
spam    Ringtone Club: Get the UK singles chart on your mobile each week
spam    Think ur smart ? Win £200 prize this week in our weekly quiz
```

Note: In real life, training data is much larger. This is just for demonstration purposes.

Let us say this is my "training data"

```
ham     Is that seriously how you spell his name?
ham     I'm going to try for 2 months ha ha only joking
ham     Oops, I'll let you know when my roommate's done
ham     I see the letter B on my car
ham     Sorry, I'll call later in meeting.
spam    URGENT! You have won a 1 week FREE membership in our £100,000 Prize Jackpot!
spam    You are a winner U have been specially selected 2 receive £1000 prize or a 4* holiday (flights inc) speak to a live operator 2 claim
0871277810810
spam    Are you unique enough? Find out from 30th August.
spam    Ringtone Club: Get the UK singles chart on your mobile each week
spam    Think ur smart ? Win £200 prize this week in our weekly quiz
```

Note: In real life, training data is much larger. This is just for demonstration purposes.

If this is what I have, what is the next thing I should do?

Features

- We need a collection of example texts with known categories (Training data)
- **We need to extract "features" we want the machine to learn from these (feature extraction)**

What can we consider as features? (Remember: our goal is to teach the computer to identify the language of spam emails versus ham emails)

Features

- Let us start with a simple (easy) idea of features - let us take all words as features.
- Intuition: If a word occurs more frequently in Spam emails in the training data, perhaps its presence is indicative of spam in real-world emails.
- So, in my training data, are there really such words???

```
ham     Is that seriously how you spell his name?
ham     I'm going to try for 2 months ha ha only joking
ham     Oops, I'll let you know when my roommate's done
ham     I see the letter B on my car
ham     Sorry, I'll call later in meeting.
spam    URGENT! You have won a 1 week FREE membership in our £100,000 Prize Jackpot!
spam    You are a winner U have been specially selected 2 receive £1000 prize or a 4* holiday (flights inc) speak to a live operator 2 claim
0871277810810
spam    Are you unique enough? Find out from 30th August.
spam    Ringtone Club: Get the UK singles chart on your mobile each week
spam    Think ur smart ? Win £200 prize this week in our weekly quiz
```

Features

- Let us start with a simple (easy) idea of features - let us take all words as features.
- Intuition: If a word occurs more frequently in Spam emails in the training data, perhaps its presence is indicative of spam in real-world emails.
- So, in my training data, are there really such words???

```
ham     Is that seriously how you spell his name?
ham     I'm going to try for 2 months ha ha only joking
ham     Oops, I'll let you know when my roommate's done
ham     I see the letter B on my car
ham     Sorry, I'll call later in meeting.
spam    URGENT! You have won a 1 week FREE membership in our £100,000 Prize Jackpot!
spam    You are a winner U have been specially selected 2 receive £1000 prize or a 4* holiday (flights inc) speak to a live operator 2 claim
0871277810810
spam    Are you unique enough? Find out from 30th August.
spam    Ringtone Club: Get the UK singles chart on your mobile each week
spam    Think ur smart ? Win £200 prize this week in our weekly quiz
```

- How do I combine those 1000 pounds, 100K pounds, 200 pounds etc - to represent one feature?

Next step: Building up some evidence of word occurrences

How?: by writing a computer program that counts word frequencies in Spam and Ham emails of training data.

Table 5.5 Some evidence from Sandy's email collection

	Spam	*Ham*
Cash	200	3
Alice	1	50
Seth	2	34
Emily	2	25
Viagra	20	0
Credit	12	2
unicorn	0	5
Cookie	1	5
hippogriff	0	18
Pony	9	50
stallion	3	8
TOTAL	250	200

Note: This is from textbook, not for the example data I showed.

Steps in Text classification?

- We need a collection of example texts with known categories (Training data)
- We need to extract "features" we want the machine to learn from these (feature extraction)
- **We should take these extracted features and give them to a "learning algorithm" (training/learning phase)**

So, how does "learning" happen?

by observing stuff like:

- Cash appeared 203 times in total in the training data, of which 200 times was in Spam. What does this mean?

So, how does "learning" happen?

by observing stuff like:

- Cash appeared 203 times in total in the training data, of which 200 times was in Spam. What does this mean?
- Alice appeared 51 times in total, and 50 times in Ham. What is the probability that an sms is "Spam" if it has a single word "Alice"?

So, how does "learning" happen?

by observing stuff like:

- Cash appeared 203 times in total in the training data, of which 200 times was in Spam. What does this mean?
- Alice appeared 51 times in total, and 50 times in Ham. What is the probability that an sms is "Spam" if it has a single word "Alice"?
- Cookie appeared only 6 times in the entire corpus (total words: 450), and of which, it appeared 5 times. What does that mean? Is it a good feature? Does it indicate strong evidence?

So, how does "learning" happen?

by observing stuff like:

- Cash appeared 203 times in total in the training data, of which 200 times was in Spam. What does this mean?
- Alice appeared 51 times in total, and 50 times in Ham. What is the probability that an sms is "Spam" if it has a single word "Alice"?
- Cookie appeared only 6 times in the entire corpus (total words: 450), and of which, it appeared 5 times. What does that mean? Is it a good feature? Does it indicate strong evidence?
- If Cash and Alice appeared in the message (not like a bigram, but just two separate words) - what inference can we draw?

So, how does "learning" happen?

by observing stuff like:

- Cash appeared 203 times in total in the training data, of which 200 times was in Spam. What does this mean?
- Alice appeared 51 times in total, and 50 times in Ham. What is the probability that an sms is "Spam" if it has a single word "Alice"?
- Cookie appeared only 6 times in the entire corpus (total words: 450), and of which, it appeared 5 times. What does that mean? Is it a good feature? Does it indicate strong evidence?
- If Cash and Alice appeared in the message (not like a bigram, but just two separate words) - what inference can we draw?
- If credit and Alice appeared in the message, is it likely that this is spam, or not?

Once the computer calculates all these possibilities ...

Julie Jones **superb**$^{+}$ performance in the **gubernatorial debate**$^{+}$ has all but **assured**$^{+}$ her of a **major victory**$^{+}$ in the **upcoming elections**$^{+}$. **Unfortunately**$^{-}$, the evening did not go as well for her opponent John Adams, his **nervous**$^{-}$ and **uncertain**$^{-}$ performance has all but **guaranteed**$^{+}$ a **loss**$^{-}$ and put his entire **political future**$^{+}$ into question.

Figure 5.3 **Positive**$^{+}$ and **negative**$^{-}$ phrases detected by Lexalytics

How do we combine these multiple sources of information?

Different learning algorithms do this differently. I will give an overview of some of these next week.

Next week

- How do we evaluate whether a learning algorithm is working? (Monday)
- Different algorithms for learning - an introduction (Wednesday)
- Conclusion of text classification and introduction to next topic (Dialog System)
- ToDo: READ CHAPTER 5!
- ToDo: Submit Assignment 4!

Attendance Question

In the five steps in text classification I mentioned earlier today, what do you think is the most difficult step for doing spam classification? Why?