LING 520: Computational Analysis of English

Semester: FALL '16

Instructor: Sowmya Vajjala

Iowa State University, USA

11 October 2016

Class Outline

- POS tagging - review questions
- Text Classification: What does that mean?
- What are some applications of text classification?
- What does learning mean for a machine?
- How do you quantify what a machine learnt?
- Practice exercise

Questions about POS Tagging

- Did you have a look at various taggers in NLTK? Did you see if they assign different tags for the same sentence?

Questions about POS Tagging

- Did you have a look at various taggers in NLTK? Did you see if they assign different tags for the same sentence?
- Did you finish doing last question of Assignment 3? What are your general observations?

Questions about POS Tagging

- Did you have a look at various taggers in NLTK? Did you see if they assign different tags for the same sentence?
- Did you finish doing last question of Assignment 3? What are your general observations?
- Did anyone try to use Regular Expressions for POS tagging?

Questions about POS Tagging

- Did you have a look at various taggers in NLTK? Did you see if they assign different tags for the same sentence?
- Did you finish doing last question of Assignment 3? What are your general observations?
- Did anyone try to use Regular Expressions for POS tagging?
- Did anyone find a tagger that can be used in Python code and is not a part of NLTK?

Questions about POS Tagging

- Did you have a look at various taggers in NLTK? Did you see if they assign different tags for the same sentence?
- Did you finish doing last question of Assignment 3? What are your general observations?
- Did anyone try to use Regular Expressions for POS tagging?
- Did anyone find a tagger that can be used in Python code and is not a part of NLTK?
- Did anyone manage to get access and successfully run Biber tagger for tagging new sentences?

Pen and paper exercise from thursday

Exercises 5.2–5.4 in J&M

- Using the tags from PTB and CLAWS7 tagsets, tag the following sentences manually, ignoring the punctuation. Compare your tags with your neighbors to discuss the agreement.
- Sentences: (you can also compare yourself to Stanford POS tagger, for example)
  1. It is a nice night.
  2. This crap game is over a garage in Fifty-second Street...
  3. ...Nobody ever takes the newspapers she sells...
  4. He is a tall, skinny guy with a long, sad, mean-looking kisser, and a mournful voice.
  5. ... I am sitting in Mindy's restaurant putting on the gefillte fish, which is a dish I am very fond of ...
  6. When a guy and a doll get to talking pecks back and forth at each other, why there you are indeed.

What is text classification?

What is text classification

- "Classification" in general refers to grouping entities into pre-defined set of classes, based on some properties.

What is text classification

- "Classification" in general refers to grouping entities into pre-defined set of classes, based on some properties.
- Text classification is one of the methods of processing textual data, where the purpose is to categorize the text into one of the pre-defined set of categories, based on the language used.

What is text classification

- "Classification" in general refers to grouping entities into pre-defined set of classes, based on some properties.
- Text classification is one of the methods of processing textual data, where the purpose is to categorize the text into one of the pre-defined set of categories, based on the language used.
- Let us say I have four categories of textual data: book reviews, movie reviews, electronics reviews and other reviews on amazon.com. The process of taking a review, and assigning it to one of these four categories - is text classification.
- Note: "Text" can be documents, sentences or even words.

Where is text classification useful?

- detecting whether the new email you got is spam or not spam. (spam classification)
- automatically detecting whether a movie review is positive or negative (opinion mining, sentiment analysis etc.)
- identifying if a news article is about "sports" or "politics" or "cinema" or "science"
- identifying whether a given word in the sentence refers to a person name or not.
- identifying if a group of words form a multi-word expression or not.

Text Classification - Applications

- Think for 5 minutes, and try to list with 5 applications of text classification, atleast 3 of which should be relevant for CALL.

Text Classification - Applications

- Think for 5 minutes, and try to list with 5 applications of text classification, atleast 3 of which should be relevant for CALL.
- Some examples: classifying learner errors into different types (spelling, non-spelling, for example); classifying the learners into proficiency levels;
- General applications: sentiment analysis of product reviews on amazon, grouping search results into categories, recommending news articles related to what you are reading etc.

What is difficult about text classification?

- Let us take this problem of classifying English learners as: beginner, intermediate and advanced.
- Let us say we even have 1000 example texts for each category, classified by expert ESL teachers.
- Now, how do we go about developing a classifier for doing this automatically?

What is difficult about text classification?

- Let us take this problem of classifying English learners as: beginner, intermediate and advanced.
- Let us say we even have 1000 example texts for each category, classified by expert ESL teachers.
- Now, how do we go about developing a classifier for doing this automatically?
- Should we look at words? develop language models? error models? somehow capture syntax?

What is difficult about text classification?

- Let us take this problem of classifying English learners as: beginner, intermediate and advanced.
- Let us say we even have 1000 example texts for each category, classified by expert ESL teachers.
- Now, how do we go about developing a classifier for doing this automatically?
- Should we look at words? develop language models? error models? somehow capture syntax?
- Should we combine everything? How? How do we even work with 3000 documents and come up with patterns??

What does it mean to "learn" to classify?

What is Machine learning?

- If you show a computing machine several examples of something, it can "learn" the patterns in these examples and try to identify identical occurrences for new data.

What is Machine learning?

- If you show a computing machine several examples of something, it can "learn" the patterns in these examples and try to identify identical occurrences for new data.
- One example: If I show the machine several articles about Donald Trump, and several articles about Hillary, the machine should be able to learn some patterns seen in both these categories. Patterns can be use of certain words and phrases, use of statistics, syntactic structures in speeches etc.

What is Machine learning?

- If you show a computing machine several examples of something, it can "learn" the patterns in these examples and try to identify identical occurrences for new data.
- One example: If I show the machine several articles about Donald Trump, and several articles about Hillary, the machine should be able to learn some patterns seen in both these categories. Patterns can be use of certain words and phrases, use of statistics, syntactic structures in speeches etc.
- Basic setup for machine learning: we have access to some set of examples, called "training set". Our goal is to make the machine learn what we want it to learn from those examples.

What is Machine learning?

- If you show a computing machine several examples of something, it can "learn" the patterns in these examples and try to identify identical occurrences for new data.
- One example: If I show the machine several articles about Donald Trump, and several articles about Hillary, the machine should be able to learn some patterns seen in both these categories. Patterns can be use of certain words and phrases, use of statistics, syntactic structures in speeches etc.
- Basic setup for machine learning: we have access to some set of examples, called "training set". Our goal is to make the machine learn what we want it to learn from those examples.
- As an approximation of what it learnt, we test how it does on a "test set". If we are satisfied, we start using this learnt model in real life.

Types of Machine learning?

Broadly, there are two types of machine learning:

- Supervised learning: when we know our categories
- Unsupervised learning: when we want to find out hidden/unknown groupings.

Note: This is oversimplification. If you really want to know more, enroll in a machine learning course. Coursera has a great introductory course by Andrew Ng (great does not mean easy).

Types of Machine learning?

Can you think of one "supervised" learning and one "unsupervised" learning scenario for corpus data?

Types of Machine learning?

Can you think of one "supervised" learning and one "unsupervised" learning scenario for corpus data? Supervised learning: one example is classifying all news articles into either "sports" or "non-sports"
Unsupervised learning: one example is identifying what are the dominant topics discussed on Twitter in the past 10 days.

How does "learning" happen?

Two aspects:

- Designing features for the machine to learn
- Developing or using an existing learning algorithm that can learn a classification function based on the values of all these features.
- An example "function" is learning weights for individual variables in linear regression.

Feature Design

- What in our opinion can be useful properties to check patterns for a given classification problem?
- Let us take the problem of classifying an email into spam or non-spam. What can be useful properties to design such a system?

Feature Design

- What in our opinion can be useful properties to check patterns for a given classification problem?
- Let us take the problem of classifying an email into spam or non-spam. What can be useful properties to design such a system?
- One more: let us say we want to classify English writing into "beginner", "intermediate" and "advanced". What can be the possible things to look at?

Feature Design

- What in our opinion can be useful properties to check patterns for a given classification problem?
- Let us take the problem of classifying an email into spam or non-spam. What can be useful properties to design such a system?
- One more: let us say we want to classify English writing into "beginner", "intermediate" and "advanced". What can be the possible things to look at?
- All these properties that can be relevant to perform machine based classification are called "features".
- The process automating feature extraction from your data (text or any other form) is called feature engineering.

Feature Design continued

There are two ways of doing feature engineering.

- 1. Kitchen sink strategy: In Spam classification example, consider all words or bi/tri grams as features, and leave it to the learning algorithm to choose what works.
- Advantage: Easy to do feature engineering, because we do not have to worry about what among those features is relevant.
- 2. Hand-crafted: Choosing specific features such as: "Use of all caps", "use of words from list X" for the same problem.
- Advantage: It is easy to understand which features are useful for the classifier and which are not. Disadvantage: Coming up with such specific features can be time consuming.

Learning Algorithm

- Goal of a learning algorithm is to take a feature representation of the training data (texts) and come up with a function that can assign weights to individual features, and use this function to predict the category for any new text it sees.
- Let us say I have 3 features: num. Nouns, num. Verbs, num. Adjectives. I have two categories: A and B. I have 1000 example texts (500 labeled A, 500 labeled B).
- A learning algorithm can learn something like this:
  1. Prediction = 0.3*numNN - 0.9*numVB + 1.1*numADJ
  2. If Prediction <=1, category is A. else, category is B.

Note: This is just one example function I created from air. There are 100s of learning algorithms, and machine learning researchers come up with new ways to learn everyday.

Measuring "Learning" success - Evaluating text classification

Measuring Success in Learning

Multiple ways. Depends on the nature of your dataset, and your application.

- Prediction accuracy on test set: typically used in most ML evaluation for text, images, videos, all sorts of things
- Revenue increase - in e-commerce applications
- False positive rate (Type 1 Error), False negatives (Type 2 error) - typically in medical applications
- Precision (TP/(TP+FP)), Recall (TP/(TP+FN)), F-score (2PR/(P+R)) - typically in information retrieval, text classification

How does this process work in real life

1. You start with designing some features, depending on your understanding of the data.
2. Develop a classification model using these features. Evaluate how it is doing on a held-out test set or using cross validation (what?)

How does this process work in real life

1. You start with designing some features, depending on your understanding of the data.
2. Develop a classification model using these features. Evaluate how it is doing on a held-out test set or using cross validation (what?)
3. Study the performance, decide whether to improve the learning algorithm or the feature representation. Decide on specific improvements.
4. Keep repeating these 3 steps until you are happy with what you have.

A small exercise

1. Here is a scenario: You are assigned the task of designing a system that recommends age-appropriate news items to children below 10 years of age. Is this a classification task?

A small exercise

1. Here is a scenario: You are assigned the task of designing a system that recommends age-appropriate news items to children below 10 years of age. Is this a classification task?
2. If we have to design a system that should classify an item on google news website into "appropriate" or "inappropriate", what do you need first?

A small exercise

1. Here is a scenario: You are assigned the task of designing a system that recommends age-appropriate news items to children below 10 years of age. Is this a classification task?
2. If we have to design a system that should classify an item on google news website into "appropriate" or "inappropriate", what do you need first?
3. Assuming someone already sat and labeled 1000 news stories as appropriate, and 500 items as inappropriate, what do you need next?

A small exercise

1. Here is a scenario: You are assigned the task of designing a system that recommends age-appropriate news items to children below 10 years of age. Is this a classification task?
2. If we have to design a system that should classify an item on google news website into "appropriate" or "inappropriate", what do you need first?
3. Assuming someone already sat and labeled 1000 news stories as appropriate, and 500 items as inappropriate, what do you need next?
4. What "features" will you look for in these documents to model "appropriateness" and "inappropriateness"? Discuss in groups of 3. Spend a few minutes on this.

A small exercise-2

1. Let us say you use some of these features. You take two off-the-shelf learning algorithms (let us say methodA, methodB) for text classification and develop classifiers. Now let us say you have a test set that has 500 texts labeled "appropriate", 250 texts labeled "inappropriate".

2.

(a) pred. →	**App.**	**Inapp.**
App.	490	10
Inapp.	200	50

(b) pred. →	**App.**	**Inapp.**
App.	400	100
Inapp.	50	200

Table: Confusion matrices for two scenarios

3. What is the classification accuracy for A and B respectively?

A small exercise-2

1. Let us say you use some of these features. You take two off-the-shelf learning algorithms (let us say methodA, methodB) for text classification and develop classifiers. Now let us say you have a test set that has 500 texts labeled "appropriate", 250 texts labeled "inappropriate".

2.

(a) pred. →	**App.**	**Inapp.**
App.	490	10
Inapp.	200	50

(b) pred. →	**App.**	**Inapp.**
App.	400	100
Inapp.	50	200

Table: Confusion matrices for two scenarios

3. What is the classification accuracy for A and B respectively?
4. According to you, which one is doing better? A or B? Why?

Next Class

- Two classification algorithms: Naive Bayes classifier, k-Nearest neighbours
- Practice exercises with NLTK