LING 120:
Language and Computers

Semester: FALL 2017

Instructor: Sowmya Vajjala

Iowa State University, USA

23 October 2017

Outline

- Assignment 4 is graded - questions discussion
- Assignment 5 description
- Quick recap of last week
- Evaluation of classification
- Recommended Reading before next class: 5.5.1 in the textbook.

Assignment discussion

Assignment 4 discussion - Question 1

- Different outputs corenlp.run gives: POS, NER, Dependencies, OpenIE
- POS: useful for various tasks - spell checkers, understanding the right sense of usage of a word, doing any of the other three tasks.
- NER: identifying names of persons, organizations etc. Useful to extract specific information, do question answering etc.
- Dependencies: This is again useful for question answering, generally understanding the meaning of a text, even for doing grammar correction etc.
- Open IE - question answering

Assignment 4 discussion - Question 2

- Different ways of developing a spell-checker:
  1. Dictionary + POS
  2. N-grams
  3. Relationships between words (e.g., which words appear together in context, which words do not appear together)

Assignment 4 discussion - Question 2

- Different ways of developing a spell-checker:
  1. Dictionary + POS
  2. N-grams
  3. Relationships between words (e.g., which words appear together in context, which words do not appear together)
- Spell checking for Turkish like languages (called "agglutinative")
  - generally requires you to break up those long words into constituent words, and then checking for what suffixes, or words go together in the language.
  - It is more difficult to come up with efficient spell checkers for such languages than for English.
  - May be you can write very detailed, and specific rules for Turkish; or you can use machine learning.

Assignment 5 description

- Deadline: November 4th
- 2 questions, related to text classification
  1. How will you do automatic language identification - what sort of resources do you need, what are the different steps, evaluation etc (basically, a summary of what you learnt so far in text classification, but for a different problem)
  2. How to do "opinion mining" (similar to above, but focus on designing domain specific features)
- Details on Canvas.

General Remarks

- When I ask for descriptive answers, I ask for descriptive answers.
- Sloppy writing is hard to evaluate leniently.
- Getting it wrong is okay, as long as you are able to explain your logic clearly
- ... and hopefully, you get the right answers after discussion.
- Make use of office hours if you don't know what my expectations are, or if you don't know how to write.

Recap of last week

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!!

Attendance Question from last class

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?

Attendance Question from last class

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?

- Training data: 2
- Feature extraction: 7
- Learning: 7
- Evaluation: 2

Measuring success in text classification

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!!

Questions and Terminology

- Let us say I am running a test to diagnose whether a patient has a disease or not. If disease or not is a classification problem, there are 4 possible outcomes:
  - If the test says positive, and it turns out the patient actually has the disease: TRUE POSITIVE
  - If the test says negative, and patient does not have the disease: TRUE NEGATIVE
  - If the test says positive, and the patient does not have the disease:?

Questions and Terminology

- Let us say I am running a test to diagnose whether a patient has a disease or not. If disease or not is a classification problem, there are 4 possible outcomes:
  - If the test says positive, and it turns out the patient actually has the disease: TRUE POSITIVE
  - If the test says negative, and patient does not have the disease: TRUE NEGATIVE
  - If the test says positive, and the patient does not have the disease:? FALSE POSITIVE
  - If the test says negative, and the patient has the disease:?

Questions and Terminology

- Let us say I am running a test to diagnose whether a patient has a disease or not. If disease or not is a classification problem, there are 4 possible outcomes:
  - If the test says positive, and it turns out the patient actually has the disease: TRUE POSITIVE
  - If the test says negative, and patient does not have the disease: TRUE NEGATIVE
  - If the test says positive, and the patient does not have the disease:? FALSE POSITIVE
  - If the test says negative, and the patient has the disease:? FALSE NEGATIVE
  - In this specific scenario, what is more dangerous: FALSE POSITIVE OR FALSE NEGATIVE?

Evaluating classification: Accuracy

- Prediction accuracy on test set: typically used in most machine learning evaluation for text, images, videos, all sorts of things: $\frac{TP+TN}{TP+TN+FP+FN}$
- What does this tell us?

Evaluating classification: Accuracy

- Prediction accuracy on test set: typically used in most machine learning evaluation for text, images, videos, all sorts of things: $\frac{TP+TN}{TP+TN+FP+FN}$
- What does this tell us?
- This tells us number about the overall percentage correct classifications by the classifier.

Evaluating classification: Precision and Recall

- Precision $= \frac{TP}{TP+FP}$
- In search: $(\frac{Relevant\ documents\ shown\ in\ results}{Total\ number\ of\ documents\ in\ the\ results})$
- Recall $= \frac{TP}{TP+FN}$
- In search: $(\frac{Relevant\ documents\ shown\ in\ results}{Total\ number\ of\ relevant\ documents\ in\ the\ web})$
- Recall is also referred to as "Sensitivity" in Medicine.

Evaluating classification in medical context

- Sensitivity is typically used in medicine, primarily focuses on questions such as: "does the patient actually have the disease as the test says?"
- If a high-sensitivity test predicts patient has a disease, it is very likely he really has the disease.

Evaluating classification in medical context

- Sensitivity is typically used in medicine, primarily focuses on questions such as: "does the patient actually have the disease as the test says?"
- If a high-sensitivity test predicts patient has a disease, it is very likely he really has the disease.
- Specificity is typically used to focus on the does the patient actually not have the disease?"
- if a high specificity test told you the patient does not have a disease, may be he likely does not have the disease and don't need further tests.
- FYI: Sensitivity $= \frac{TP}{TP+FN}$, Specificity$= \frac{TN}{TN+FP}$

Evaluating classification: Others

- Several other measures: `https://en.wikipedia.org/wiki/Precision_and_recall`
- What is a good evaluation measure depends on what you want out of your classifier.

Evaluating classification: Others

- Several other measures: `https://en.wikipedia.org/wiki/Precision_and_recall`
- What is a good evaluation measure depends on what you want out of your classifier.
- if you are using a spam classifier, would you want to see have more True positives (not-spam means not-spam) or more True-negatives (spam is spam) or both?

Evaluating classification: Others

- Several other measures: `https://en.wikipedia.org/wiki/Precision_and_recall`
- What is a good evaluation measure depends on what you want out of your classifier.
- if you are using a spam classifier, would you want to see have more True positives (not-spam means not-spam) or more True-negatives (spam is spam) or both?
- What is the difference between high true-positives and high true-negatives?

Evaluating classification: Confusion Matrix

- A confusion matrix is a summary of all those measures we discussed so far. You can get each of those from such a matrix.
- Let us take an example matrix for a classification problem with two categories: Correct, Wrong

actual. $\downarrow$ pred. $\rightarrow$	**Correct**	**Incorrect**
Correct	400	100
Incorrect	100	400

- You can calculate your TP, TN, FP, FN, Accuracy, Precision (whatever measure you want) !

Confusion Matrix for more than two categories

actual. ↓ pred. →	**Sports**	**Politics**	**Others**
Sports	400	50	50
Politics	75	425	0
Others	0	0	500

- How many total news items are there in this dataset?

Confusion Matrix for more than two categories

actual. ↓ pred. →	**Sports**	**Politics**	**Others**
Sports	400	50	50
Politics	75	425	0
Others	0	0	500

- How many total news items are there in this dataset?
- How many were classified correctly?

Confusion Matrix for more than two categories

actual. $\downarrow$ pred. $\rightarrow$	**Sports**	**Politics**	**Others**
Sports	400	50	50
Politics	75	425	0
Others	0	0	500

- How many total news items are there in this dataset?
- How many were classified correctly?
- How do you get True positives, True negatives etc in this case?

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!!**

Ultimate evaluation for a real-world application is customer satisfaction, increase in revenue etc. (beyond all these measures)

Attendance Exercise

Write on paper or submit on Canvas

1. Let us say you are working on classifying webpages as "appropriate" and "inappropriate" for children and you developed two classifiers.
2. Now let us say you have a test set that has 500 texts labeled "appropriate", 250 texts labeled "inappropriate".
3. Here are the confusion matrices for Classifiers A and B:

(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

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