ILLINOIS INSTITUTE OF TECHNOLOGY

Midterm Review

CS-585

Natural Language Processing

Derrick Higgins

Exam info

- Exam locations and times
  - Midterm: Tuesday, October 12, 6:45-8:15 PM, IIT Tower 1F6–1
  - Final exam: Tuesday, December 7, 7:30-9:30 PM
- Midterm details
  - Timed: 90 minutes
  - Multiple-choice
  - Open-book, open-notes
  - Written or printed materials OK, but no electronics
  - Bring pencils with erasers

Content overview

- Linguistics
- Math
- Information Theory
- Text Processing
- Evaluation
- Words and Word Frequency
- Machine Learning and Text Categorization

LINGUISTICS

Linguistics

- Know the different subfields of linguistics
- Know the types of NLP tasks associated with each

Phonetics

Phonology

Morphology

Syntax

Semantics

Pragmatics

Sociolinguistics

Psycholinguistics

MATH

Math

- Understanding of fundamentals: probabilities, linear algebra, information theory
- Much more about demonstrating understanding than manipulating formulas

Math

- Vector and matrix operations
  - Addition
  - Multiplication
  - Dot product
  - Transpose
- Recognize, apply and understand key functions
  - Sigmoid
  - Logit
  - Softmax
  - Norms
  - Cosine similarity

Math

- Probabilistic decompositions
  - Definition of conditional probability
  - Decomposition of joint probability given (absolute) independence of random variables
  - Decomposition of joint probability given conditional independence of random variables
  - Bayes' rule
- Probabilistic reasoning

INFORMATION THEORY

Information Theory

- Information content for an observation from a known distribution
- Entropy of a known distribution
- Bits vs. nats
- Entropy vs. perplexity

TEXT PROCESSING

Text processing

- Be able to recognize the function of all of the Unix text processing tools discussed in class
- Understand how STDIN and STDOUT work, and how programs are composed with the pipe (|) operator to perform complex operations

cat
head
tail
cut
paste
tr
sed
grep
sort
uniq
wc

EVALUATION

- Understand how specific measures are used to assess the performance of a model, or the agreement between annotators
- Understand the concept of generalization to unseen data, and how to adjust model parameters to improve generalization

- Accuracy
- True/False Positives
- True/False Negatives
- Precision
- Recall
- F-measure (F1)
- Kappa

WORDS & WORD FREQUENCY

Words and word frequencies

- Be familiar with different lexical representation types
  - Words
  - Stems
  - Lemmas
  - Word pieces/BPE
  - Ngrams
  - Multi-word expressions
- Understand methods for representing words as vectors
  - One-hot encoding
  - Hashing trick
  - Latent semantic analysis
  - word2vec

Words and word frequencies

- Understand the properties of a word frequency distribution derived from a set of texts, and how frequencies/counts differ for words in different parts of the distribution
  - Zipf's law
  - Vocabulary size : corpus size
- Be familiar with different vectorization methods for creating document vectors out of word vectors; their advantages and disadvantages
  - Binary vectorizer
  - Count vectorizer
  - Tf*idf vectorizer

TEXT CATEGORIZATION AND MACHINE LEARNING

Text categorization and machine learning

- Be familiar with the loss functions used for optimizing different types of NLP models
  - Binary/categorical cross-entropy
  - Squared error
- Be familiar with general-purpose optimization algorithms such as gradient descent, and practical challenges in using them
  - Stochastic / batch / minibatch gradient descent
  - Momentum
  - Adagrad / adadelta

Text categorization and machine learning

- Understand the bag-of-words representation for documents, why it is chosen for some tasks, and for what sorts of tasks it is useful
- Understand the concept of model regularization, and specific types of regularization for different model types
- Have a detailed understanding of the naïve Bayes classifier, and the probabilistic assumptions behind it
- Understand and be able to identify key machine learning concepts
  - Representation learning
  - Supervised vs. unsupervised learning
  - Generative vs. discriminative models
  - Neural network building blocks

NLP tasks

- Understand the different possible technical approaches to tasks we have discussed, and the resources required for them
  - Word sense disambiguation
  - Sentiment analysis
  - Latent semantic analysis
  - word2vec