NLP Notes Info 159/259

Language

- Language is inherently hard to deal with
  - Speech Acts
  - Conversational Implicature
  - Shared Knowledge
  - Variation/Indexicality
- NLP involved representing language for some end
  - Dialogue
  - Translation
  - Speech Recognition
  - Text Analysis
- Certain Views of NLP
  - Information Theoretic View
    - Involves encoding and then subsequently decoding language
  - Rational Speech View
    - Communication involved recursive reasoning
    - How can X choose words that make sense to Y
  - Pragmatic View
    - Meaning is co-constructed by the people involved and the context of the speech
  - Whorfian View
    - Structure of language influences thought

Text Classification

- Sentiment Analysis
  - Is the entire text positive or negative?
  - Difficult as it is the measure of a speaker's private state, which is unobservable
- Representation for SA
  - Only positive/negative words
  - Only words in isolation
  - Conjunctions of words
  - Higher-order linguistic structure (syntax)
- Independence
  - $P(A,B) = P(A) * P(B)$
- Data Likelihood
  - Gives a strategy for choosing the single best parameter
- Conditional Probability
  - $P(X=x \mid Y=y)$
  - In this scenario you can use Bayes Rule

- Bayes Rule
  - $\boldsymbol{P(X=x \mid Y=y)} = (P(X=x) P(Y=y \mid X=x)) / (P(X=x) P(Y=y \mid X=x))$
  - Training such a classifier means working with a summation of the denominator

Text Classification Methods

- Naïve Bayes
  - There is flexibility about the probability distributions used in NB
  - Includes Multinomial, Bernoulli, Normal, Poisson
- Multinomial Naïve Bayes
  - Discrete Distribution for modeling count data (words counts, proportion out of a whole)
- Bernoulli Naïve Bayes
  - Binary Event (True or False)
  - One parameter: P (Probability of an event occurring)
  - In terms of speech analysis, can put breaks within the writing to point out that the statement is negative in whole, even if a singular part is not
- Generative vs. Discriminative Models
  - Generative models specify a joint distribution over the labels and the data, can use this to generate new data
    - Care about P(Y|X) but need to model more
  - Discriminative models specify the conditional distribution of the label y given the data x, focused on how to discriminate between the classes
    - Only care about the probability of P(Y|X)
- Logistic Regression
  - This is a method to engage in a discriminative model
  - Does not assume that features are independent like Naïve Bayes does
- Features
  - Can encode your own understanding of domain understanding of the problem
- Conditional Likelihood
  - For the training data, want probability of true label y for each data point x to be high
  - Lets us pick parameters B that maximize the probability that the training data is true
- Feature Selection
  - We can threshold features by minimum count, but this would throw away information
  - We can take a probabilistic approach and encode a prior belief that all B should be 0, unless there is evidence otherwise
  - As a ***practical*** part of this, when calculating thorugh the gradient, only loop through nonzero valued features

Neural Networks

- Discrete, high-dimensional representation of inputs
    - Non-linear interactions of inputs
    - Multiple layers to capture hierarchical structure
- Design of Neural Networks
    - Tremendous flexibility on design choices
    - Articulate model structure and use the chain rule to derive parameter updates
- Regularization
    - Increasing the number of parameters
        - Increases the possibility for overfitting to training data
- Convolutional Networks
    - Same operation is applied to different regions of the input

Hypothesis Testing (Similar to Data8, less specific)

- Measures confidence in what we can say about a null from a sample
- Decide at which point is a sample statistic unusual enough to reject the null hypothesis
- Decide on the level of significance of p
    - Testing evaluates whether the sample statistic falls in the region that rejects the null in the rejection region defined by p
- Many such tests rely on parametric assumptions
    - Permutation test and the bootstrap do not depend on these assumptions

Language Modelling

- Provide a way to quantify the likelihood of a sequence
    - Can use multi-class logistic regression for language modelling by treating the vocabulary as the output space