Scaling Words on an Ideological Space

March 7, 2013

Words as Data

- Explosion of interest in studying text data
  - The Internet search problem $\rightarrow$ multibillion dollar industry
  - Massive investment in machine learning technology to classify and predict words and documents
  - Recent collection and digitization of text
- The social and political world is filled with an immensity of text that capture meaning
  - E.g., Legislative debates, party platforms, advertisements, legal decisions, statutes, newspapers, magazines, academic journals, historical records...
- More data, bigger haystacks
  - Impossible to study all this data in human time
  - Give up nuance and subtlety, and model language instrumentally

Words as Data

- Classification, scaling, uncovering sentiment, and much more
  - Words themselves are not interesting
  - Model words to uncover latent distributions of things we care about
- Focus on two types of analysis
  - Uncovering a spatial dimension by scoring words
    - Naive Classifier (e.g., Wordscores, Bayescores)
    - Bayesian IRT (e.g., Wordfish)
  - Discovering topics within sets of documents
    - Latent Dirichlet Allocation (LDA)
- Set aside computational and sparsity issues

Uncovering ideology

- Goal is to observe whether parties offer liberal, conservative, or moderate policies to voters
    - Ideology is unobserved, but is expected to influence the patterns of words used in the platform documents
    - Specifically, liberals (L) and conservatives (C) are expected to use different subsets of words
    - And moderates (M) are expected to use a mixture of both
- Use a naive classification approach – liberal or conservative?
    - Score words based on their frequency used by liberals or conservatives in a *training set*
    - Compute probabilities used to score documents in the left out *testing set*
    - Classification probabilities interpreted as ideological scores

Word Scores

- Define $p(L|w_i)$ to be the probability that a text offers a liberal position given the word w_i. Using Bayes rule:

$$p(L|w_i) = \frac{p(w_i|L)p(L)}{p(w_i)}$$
$$= \frac{p(w_i|L)p(L)}{p(w_i|L)p(L) + p(w_i|C)p(C)}$$

- Define $W^{\{L\}}$ or $W^{\{C\}}$ to be the total number of words in a document L or C
- Define $W_i^{\{L\}}$ or $W_i^{\{C\}}$ to be the count of the ith word appearing in document L or C

Word Scores

- Assuming diffuse priors on words:

$$
\begin{aligned}
p(w_i|L) &= \frac{W_i^{\{L\}}}{W^{\{L\}}}, \qquad \text{and} \\
p(L) &= \frac{W^{\{L\}}}{W^{\{L\}} + W^{\{C\}}}
\end{aligned}
$$

  Note that $p(L)$ is a measure of the prior probability that any document is liberal, summing over all the words

- Putting this together we have

$$
p(L|w_i) = \frac{W_i^{\{L\}}}{W_i^{\{L\}} + W_i^{\{C\}}}
$$

Word Scores

- In practice, we define L documents and C documents and compute $p(L|w_i)$ and $p(C|w_i)$ for all $w_i \in \{L, C\}$
- Define S_i to be a word scoring on a scale from -1 to 1:

$$S_i = -1 \times p(L|w_i) + 1 \times p(C|w_i)$$

- Each testing document V is scored accordingly:

$$S_V = \sum_{i=1}^{N_v} \frac{W_i^{\{V\}}}{W^{\{V\}}} \times S_i$$

- Word frequencies are influential on the scoring – rare words do not contribute probabilities proportional to their informativeness

Bayes Scores Correction

- A Bayesian approach would be to consider the posterior $p(L|V)$, or the density of liberal documents L given some document V (Beauchamp 2012).

$$
\begin{aligned}
p(L|V) &= \frac{p(V|L)p(L)}{p(V)}, \qquad \text{and} \\
p(C|V) &= \frac{p(V|C)p(C)}{p(V)}
\end{aligned}
$$

- Define $p(w_i|L)$ to be the probability of encountering word w_i given we're looking at document L
    - A simplifying (and potentially strong) assumption is that words appear conditionally independent in document V
    - This gives: $p(V|L) = \prod_{i=1}^{N_V} p(w_i|L)$

Bayes Scores Correction

- Under this independence assumption

$$
\begin{aligned}
p(L|V) &= \frac{p(L)}{p(V)} \prod_{i=1}^{N_V} p(w_i|L) \\
p(C|V) &= \frac{p(C)}{p(V)} \prod_{i=1}^{N_V} p(w_i|C)
\end{aligned}
$$

- Let's take the log of the ratio of these two likelihoods:

$$
\begin{aligned}
\log \frac{p(L|V)}{p(C|V)} &= \log \frac{p(L)}{p(C)} + \sum_{i=1}^{N_V} \log \frac{p(w_i|L)}{p(w_i|C)} \\
BayesScore_V &= \sum_{i=1}^{N_V} \log \frac{p(w_i|L)}{p(w_i|C)}
\end{aligned}
$$

- Since $p(L)/p(C)$ does not depend on w_i

Shortcomings of Scoring Methods

- Useful information may be excluded
  - Word frequencies are not diffusely distributed
  - Influence of words is heterogeneous
- Lose statistical properties of uncertainty, asymptotics
- No model to fit, so model and prediction validation may be a challenge

Ideal Point Model of Words

- Wordfish model of word counts (Slapin and Proksch 2008)
  - Word counts are assumed to be Poisson distributed
- Item count test (ICT) Rasch model – cousin of IRT
  - Count the number of 'right' answers on k tests, where each word is a test
  - More correct answers suggest greater ability (α_i) on any test
  - Some tests are more discriminating (β_k) than others
  - Test-taker (γ_i) and word (δ_k) fixed effects – shifts thresholds for number of correct answers i gets on k

Ideal Point Model of Words

- Define a parameter $\lambda_{ik} = \exp\{\gamma_i + \delta_k + \beta_k \times \alpha_i\}$. Under the model

$$
\begin{aligned}
y_{ik} &\sim Poisson(\lambda_{ik}) \\
\alpha_i &\sim N(\mu_\alpha, \sigma_\alpha^2) \\
\gamma_i &\sim N(\mu_\gamma, \sigma_\gamma^2) \\
\beta_k &\sim N(\mu_\beta, \sigma_\beta^2) \\
\delta_k &\sim N(\mu_\delta, \sigma_\delta^2)
\end{aligned}
$$

- Fix all $\mu = 0$, and $\sigma = 1$

Ideal Point Model of Words

- Estimation can be done fully Bayesian or using a variant of Expectation Maximization (EM)
- For EM:
  - Stage 1: Estimate i terms, α_i and γ_i, fixing all the kth terms
  - Stage 2: Estimate k terms, β_k and δ_k, fixing all the ith terms
  - Repeat until convergence
- To incorporate prior information, define log likelihood $\log \mathcal{L}(y_{ik}|\alpha_i, \gamma_i, \beta_k, \delta_k)$. Maximize:

$$\mathcal{L}(y_{ik}|\theta) \quad - \quad \sum_{\theta} \rho_\theta \left(\frac{\mu_\theta - \theta}{\sigma_\theta} \right)^2$$

For each $\theta \in \{\alpha, \gamma, \beta, \delta\}$, and a penalty term ρ_θ

Ideal Point Model of Words

- Unlike scoring methods, the model approach can allow for additional complexity
  - Multidimensionality: $\tilde{\alpha} = \{\alpha_1, \alpha_2, ..., \alpha_g\}$, and so on
  - Complex word processes: model common v. non-common words, underlying constraint or correlation amongst words, hierarchical clustering of words in phrases, etc
- Pull out predictions about future documents or words
- Estimate measures of statistical uncertainty of our scores
- In practice, estimates are very often statistically indistinguishable
  - Favorite example here is to use OLS to score documents