Why Language is Hard: Structure and Predictions

Introduction to Data Science Algorithms
Jordan Boyd-Graber and Michael Paul
(NOT ON FINAL!)
• Not on final (but using concepts from course)
• Preview of other courses: 5622, 5832 (often cross-listed with UG)
• CLIP: focus group for second half of class
• HW1 Due today!
Most supervised algorithms are . . .

Logistic Regression
Most supervised algorithms are ...

Logistic Regression

\[ p(y | x) = \sigma(\sum_i \beta_i x_i) \]

SVM

\[ \text{sign}(\vec{w} \cdot x + b) \]

What statistical property do these (and many others) share?

\[ p(y_i, y_j | x_i, x_j) = p(y_i | x_i) p(y_j | x_j) \]

Independent!
Most supervised algorithms are . . .

**Logistic Regression**

\[
p(y \mid x) = \sigma(\sum_i \beta_i x_i)
\]

**SVM**

\[
\text{sign}(\hat{w} \cdot x + b)
\]

- What statistical property do these (and many others share)?
Most supervised algorithms are . . .

Logistic Regression

\[ p(y \mid x) = \sigma \left( \sum_i \beta_i x_i \right) \]

SVM

\[ \text{sign}(\hat{w} \cdot x + b) \]

• What statistical property do these (and many others share)?

• Hint: \[ p(y_i, y_j \mid x_i, x_j) = p(y_i \mid x_i)p(y_j \mid x_j) \]
Most supervised algorithms are ... 

<table>
<thead>
<tr>
<th>Logistic Regression</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p(y \mid x) = \sigma(\sum_i \beta_i x_i) )</td>
<td>( \text{sign}(\hat{w} \cdot x + b) )</td>
</tr>
</tbody>
</table>

- What statistical property do these (and many others share)?
- Hint: \( p(y_i, y_j \mid x_i, x_j) = p(y_i \mid x_i)p(y_j \mid x_j) \)
- Independent!
Is this how the world works?
Is this how the world works?

Also particularly relevant for 2016: correlated voting patterns
Structured Prediction

- Specifically models interactions between predictions
- Requires more robust models and algorithms
- Used in vision, language processing, finance
Structured Prediction

- Specifically models interactions between predictions
- Requires more robust models and algorithms
- Used in vision, language processing, finance
- Focus: part of speech tagging