Naïve Bayes & Maxent Models

CMSC 473/673
UMBC
September 18\textsuperscript{th}, 2017

Some slides adapted from 3SLP
Announcements: Assignment 1

Due 11:59 AM, Wednesday 9/20

< 2 days

Use submit utility with:
  class id cs473_ferraro
  assignment id a1

We must be able to run it on GL!
  Common pitfall #1: forgetting files
  Common pitfall #2: incorrect paths to files
  Common pitfall #3: 3rd party libraries
Announcements: Course Project

Official handout will be out Wednesday 9/20

Until then, focus on assignment 1

Teams of 1-3

Mixed undergrad/grad is encouraged but not required

Some novel aspect is needed

Ex 1: reimplement existing technique and apply to new domain
Ex 2: reimplement existing technique and apply to new (human) language
Ex 3: explore novel technique on existing problem
Recap from last time...
Two Different Philosophical Frameworks

\[ p(X | Y) = \frac{p(Y | X) \times p(X)}{p(Y)} \]

- **posterior probability**
- **likelihood**
- **prior probability**
- **marginal likelihood**

- Posterior Classification/Decoding
  - maximum a posteriori
- Noisy Channel Model Decoding

*there are others too (CMSC 478/678)*
Posterior Decoding:
Probabilistic Text Classification

Assigning subject categories, topics, or genres
Spam detection
Authorship identification

Age/gender identification
Language Identification
Sentiment analysis
...

\[ p(X \mid Y) = \frac{p(Y \mid X) \cdot p(X)}{p(Y)} \]

- \( p(Y \mid X) \): observation likelihood (averaged over all classes)
- \( p(X) \): class-based likelihood (language model)
- \( p(Y) \): prior probability of class

class

observed data

class
Noisy Channel Model

what I want to tell you “sports”

what you actually see “The Os lost again…”

hypothesized intent “sad stories” “sports”

rewrite according to what’s likely “sports”
Noisy Channel

Machine translation
Speech-to-text
Spelling correction
Text normalization

Part-of-speech tagging
Morphological analysis
Image captioning
...

\[ p(X \mid Y) = \frac{p(Y \mid X) \ast p(X)}{p(Y)} \]

possible (clean) output
observed (noisy) text
translation/decode model
(clean) language model
observation (noisy) likelihood
Classify or Decode with Bayes Rule

\[ \text{argmax}_X p(X \mid Y) \]
Classify or Decode with Bayes Rule

\[
\arg\max_X \frac{p(Y \mid X) \cdot p(X)}{p(Y)}
\]
Classify or Decode with Bayes Rule

\[
\arg\max_X \frac{p(Y \mid X) \ast p(X)}{p(Y)}
\]

constant with respect to \(X\)
Classify or Decode with Bayes Rule

\[ \text{argmax}_x p(Y \mid X) \ast p(X) \]
My Hobby:
Sitting down with grad students and timing how long it takes them to figure out that I'm not actually an expert in their field.

**Engineering:**
Our big problem is heat dissipation. Have you tried logarithms?

**Linguistics:**
Ah, so does this Finno-Ugric family include, say, Klingon?

**Sociology:**
Yeah, my latest work is on ranking people from best to worst.

**Literary Criticism:**
You see, the deconstruction is inextricable from not only the text, but also the self.

48 seconds 63 seconds 4 minutes

Eight papers and two books and they haven't caught on.
MY HOBBY:
SITTING DOWN WITH GRAD STUDENTS AND TIMING
HOW LONG IT TAKES THEM TO FIGURE OUT THAT
I'M NOT ACTUALLY AN EXPERT IN THEIR FIELD.

ENGINEERING:
OUR BIG PROBLEM IS HEAT DISSIPATION
HAVE YOU TRIED LOGARITHMS?

48 SECONDS

LINGUISTICS:
AH, SO DOES THIS FINNO-UGRIC FAMILY INCLUDE,
SAY, KLINGON?

63 SECONDS

SOCIOLOGY:
Yeah, my latest work is on ranking people
from best to worst.

4 MINUTES

LITERARY CRITICISM:
YOU SEE, THE DECONSTRUCTION IS INEXTRICABLE FROM NOT ONLY
THE TEXT, BUT ALSO THE SELF.

EIGHT PAPERS AND
TWO BOOKS AND THEY
HAVEN'T CAUGHT ON.
Classify or Decode with Bayes Rule

$$\arg\max_X \log p(Y \mid X) + \log p(X)$$
Classify or Decode with Bayes Rule

$$\text{argmax}_X \log p(Y \mid X) + \log p(X)$$
Classify or Decode with Bayes Rule

\[
\text{argmax}_X \log p(Y \mid X) + \log p(X)
\]

how likely is label X overall?

how well does text Y represent label X?
Classify or Decode with Bayes Rule

\[
\text{argmax}_X \log p(Y \mid X) + \log p(X)
\]

*how likely is label \(X\) overall?*

*how well does text \(Y\) represent label \(X\) ?*

For “simple” or “flat” labels:
* iterate through labels
* evaluate score for each label, keeping only the best (n best)
* return the best (or n best) label and score
Classify or **Decode** with Bayes Rule

$$\operatorname*{argmax}_X \log p(Y \mid X) + \log p(X)$$

**how well does text (complex input) Y represent text (complex output) X?**

**how likely is text (complex output) X overall?**
Classify or **Decode** with Bayes Rule

\[
\text{argmax}_x \log p(Y \mid X) + \log p(X)
\]

* iterate through labels
* evaluate score for each label, keeping only the best (n best)
* return the best (or n best) label and score

**how likely is text (complex output) X overall?**

**how well does text (complex input) Y represent text (complex output) X?**

* can be complicated
Classify or **Decode** with Bayes Rule

argmax_X \log p(Y \mid X) + \log p(X)

* iterate through each label
* evaluate score for each label, keeping only the best (n best)
* return the best (or n best) label and score

* how likely is text (complex output) X overall?

* how well does text (complex input) Y represent text (complex output) X?

* can be complicated

we’ll come back to this in October
Evaluation: the 2-by-2 contingency table

<table>
<thead>
<tr>
<th></th>
<th>Actually Correct</th>
<th>Actually Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Selected/ Guessed</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Not selected/ not guessed</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Classification Evaluation: the 2-by-2 contingency table

<table>
<thead>
<tr>
<th></th>
<th>Actually Correct</th>
<th>Actually Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Selected/ Guessed</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Not selected/ not guessed</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Classes/Choices*
Classification Evaluation: the 2-by-2 contingency table

<table>
<thead>
<tr>
<th></th>
<th>Actually Correct</th>
<th>Actually Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Selected/ Guessed</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>True Positive</strong></td>
<td>Blue (Correct (TP))</td>
<td>Blue (Guessed)</td>
</tr>
<tr>
<td><strong>Not selected/ not guessed</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Classes/Choices

- Blue: Correct
- White: Guessed
Classification Evaluation: the 2-by-2 contingency table

<table>
<thead>
<tr>
<th>Selected/ Guessed</th>
<th>Actually Correct</th>
<th>Actually Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Positive</td>
<td>Correct (TP)</td>
<td>False Positive</td>
</tr>
<tr>
<td></td>
<td>Guessed</td>
<td>(FP)</td>
</tr>
<tr>
<td>Not selected/ not guessed</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Classification Evaluation: the 2-by-2 contingency table

<table>
<thead>
<tr>
<th></th>
<th>Actually Correct</th>
<th>Actually Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Selected/ Guessed</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True Positive (TP)</td>
<td><img src="blue" alt="Correct" /> <img src="blue" alt="Guessed" /></td>
<td>False Positive (FP) <img src="white" alt="Correct" /> <img src="blue" alt="Guessed" /></td>
</tr>
<tr>
<td><strong>Not selected/ not guessed</strong></td>
<td>False Negative (FN) <img src="blue" alt="Correct" /> <img src="white" alt="Guessed" /></td>
<td></td>
</tr>
</tbody>
</table>
## Classification Evaluation: the 2-by-2 contingency table

<table>
<thead>
<tr>
<th></th>
<th>Actually Correct</th>
<th>Actually Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Selected/ Guessed</strong></td>
<td>True Positive (TP)</td>
<td>False Positive (FP)</td>
</tr>
<tr>
<td></td>
<td>Correct</td>
<td>Correct</td>
</tr>
<tr>
<td></td>
<td>Guessed</td>
<td>Guessed</td>
</tr>
<tr>
<td><strong>Not selected/ not guessed</strong></td>
<td>False Negative (FN)</td>
<td>True Negative (TN)</td>
</tr>
<tr>
<td></td>
<td>Correct</td>
<td>Correct</td>
</tr>
<tr>
<td></td>
<td>Guessed</td>
<td>Guessed</td>
</tr>
</tbody>
</table>
Classification Evaluation: Accuracy, Precision, and Recall

**Accuracy:** % of items correct

\[
\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}
\]

<table>
<thead>
<tr>
<th></th>
<th>Actually Correct</th>
<th>Actually Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Selected/Guessed</strong></td>
<td>True Positive (TP)</td>
<td>False Positive (FP)</td>
</tr>
<tr>
<td><strong>Not select/not guessed</strong></td>
<td>False Negative (FN)</td>
<td>True Negative (TN)</td>
</tr>
</tbody>
</table>
Classification Evaluation: Accuracy, Precision, and Recall

**Accuracy**: % of items correct
\[
\frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}
\]

**Precision**: % of selected items that are correct
\[
\frac{\text{TP}}{\text{TP} + \text{FP}}
\]

<table>
<thead>
<tr>
<th>Selected/Guessed</th>
<th>Actually Correct</th>
<th>Actually Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actually Correct</td>
<td>True Positive (TP)</td>
<td>False Positive (FP)</td>
</tr>
<tr>
<td>Actually Incorrect</td>
<td>False Negative (FN)</td>
<td>True Negative (TN)</td>
</tr>
<tr>
<td>Not selected/not guessed</td>
<td>False Negative (FN)</td>
<td>True Negative (TN)</td>
</tr>
</tbody>
</table>
Classification Evaluation: Accuracy, Precision, and Recall

**Accuracy**: % of items correct

\[
\frac{TP + TN}{TP + FP + FN + TN}
\]

**Precision**: % of selected items that are correct

\[
\frac{TP}{TP + FP}
\]

**Recall**: % of correct items that are selected

\[
\frac{TP}{TP + FN}
\]

<table>
<thead>
<tr>
<th>Selected/Guessed</th>
<th>Actually Correct</th>
<th>Actually Incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>True Positive (TP)</strong></td>
<td>Not select/not guessed</td>
<td>False Positive (FP)</td>
</tr>
<tr>
<td><strong>False Negative (FN)</strong></td>
<td>Selected/Guessed</td>
<td>True Negative (TN)</td>
</tr>
</tbody>
</table>
A combined measure: $F$

Weighted (harmonic) average of Precision & Recall

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}}$$
A combined measure: $F$

Weighted (harmonic) average of Precision & Recall

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}} = \frac{(1 + \beta^2) \cdot P \cdot R}{(\beta^2 \cdot P) + R}$$

(algebra (not important))
A combined measure: \( F \)

Weighted (harmonic) average of Precision & Recall

\[
F = \frac{(1 + \beta^2) \times P \times R}{(\beta^2 \times P) + R}
\]

Balanced F1 measure: \( \beta=1 \)

\[
F_1 = \frac{2 \times P \times R}{P + R}
\]
Micro- vs. Macro-Averaging

If we have more than one class, how do we combine multiple performance measures into one quantity?

**Macroaveraging**: Compute performance for each class, then average.

**Microaveraging**: Collect decisions for all classes, compute contingency table, evaluate.
### Micro- vs. Macro-Averaging: Example

#### Class 1

<table>
<thead>
<tr>
<th>Classifier: yes</th>
<th>Truth: yes</th>
<th>Truth: no</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Classifier: no</th>
<th>Truth: yes</th>
<th>Truth: no</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>970</td>
<td></td>
</tr>
</tbody>
</table>

#### Class 2

<table>
<thead>
<tr>
<th>Classifier: yes</th>
<th>Truth: yes</th>
<th>Truth: no</th>
</tr>
</thead>
<tbody>
<tr>
<td>90</td>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Classifier: no</th>
<th>Truth: yes</th>
<th>Truth: no</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>890</td>
<td></td>
</tr>
</tbody>
</table>

#### Micro Ave. Table

<table>
<thead>
<tr>
<th>Classifier: yes</th>
<th>Truth: yes</th>
<th>Truth: no</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>20</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Classifier: no</th>
<th>Truth: yes</th>
<th>Truth: no</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>1860</td>
<td></td>
</tr>
</tbody>
</table>

Macroaveraged precision: \((0.5 + 0.9)/2 = 0.7\)

Microaveraged precision: \(100/120 = .83\)

Microaveraged score is dominated by score on common classes
Language Modeling as Naïve Bayes Classifier

\[ p(X \mid Y) = \frac{p(Y \mid X) \ast p(X)}{p(Y)} \]

*posterior probability*

- class-based likelihood (language model)
- prior probability of class
- observation likelihood (averaged over all classes)

**Posterior Classification/Decoding**

*maximum a posteriori*

**Noisy Channel Model Decoding**
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!
Bag of Words Representation

\[ \gamma(\ ) = c \]

<table>
<thead>
<tr>
<th>Word</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>seen</td>
<td>2</td>
</tr>
<tr>
<td>sweet</td>
<td>1</td>
</tr>
<tr>
<td>whimsical</td>
<td>1</td>
</tr>
<tr>
<td>recommend</td>
<td>1</td>
</tr>
<tr>
<td>happy</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Language Modeling as Naïve Bayes Classifier

$$\arg\max_x p(Y \mid X) \ast p(X)$$

Start with Bayes Rule
Language Modeling as Naïve Bayes Classifier

$$\text{argmax}_x \prod_i p(Y_i|x) * p(x)$$

Adopt naïve bag of words representation $Y_i$
Language Modeling as Naïve Bayes Classifier

$$\arg\max_X \prod_i p(Y_i|X) \ast p(X)$$

Adopt naïve bag of words representation $Y_i$

Assume position doesn’t matter
Language Modeling as Naïve Bayes Classifier

$$\arg\max_X \prod_i p(Y_i | X) * p(X)$$

Adopt naïve bag of words representation $Y_i$

Assume position doesn’t matter

Assume the feature probabilities are independent given the class $X$
Multinomial Naïve Bayes: Learning

From training corpus, extract *Vocabulary*
Multinomial Naïve Bayes: Learning

From training corpus, extract *Vocabulary*

**Calculate $P(c_j)$ terms**

For each $c_j$ in $C$ do

$docs_j = \text{all docs with class } = c_j$

$$p(c_j) = \frac{|docs_j|}{\# \text{ docs}}$$
Multinomial Naïve Bayes: Learning

From training corpus, extract *Vocabulary*

**Calculate $P(c_j)$ terms**
For each $c_j$ in $C$
do $docs_j = \text{all docs with class } = c_j$

$$p(c_j) = \frac{|docs_j|}{\# \text{ docs}}$$

**Calculate $P(w_k \mid c_j)$ terms**
$Text_j = \text{single doc containing all } docs_j$
For each word $w_k$ in *Vocabulary*

$$n_k = \# \text{ of occurrences of } w_k \text{ in } Text_j$$

$$p(w_k \mid c_j) = \text{class LM}$$
Naïve Bayes and Language Modeling

Naïve Bayes classifiers can use any sort of feature

But if, as in the previous slides
  We use **only** word features
  we use **all** of the words in the text (not a subset)

Then
  Naïve Bayes has an important similarity to language modeling
Naïve Bayes as a Language Model

<table>
<thead>
<tr>
<th>Positive Model</th>
<th>Negative Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1  l</td>
<td>0.2  l</td>
</tr>
<tr>
<td>0.1  love</td>
<td>0.001  love</td>
</tr>
<tr>
<td>0.01  this</td>
<td>0.01  this</td>
</tr>
<tr>
<td>0.05  fun</td>
<td>0.005  fun</td>
</tr>
<tr>
<td>0.1  film</td>
<td>0.1  film</td>
</tr>
</tbody>
</table>
Naïve Bayes as a Language Model

Which class assigns the higher probability to $s$?

<table>
<thead>
<tr>
<th>Positive Model</th>
<th>Negative Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1 l</td>
<td>0.2 l</td>
</tr>
<tr>
<td>0.1 love</td>
<td>0.001 love</td>
</tr>
<tr>
<td>0.01 this</td>
<td>0.01 this</td>
</tr>
<tr>
<td>0.05 fun</td>
<td>0.005 fun</td>
</tr>
<tr>
<td>0.1 film</td>
<td>0.1 film</td>
</tr>
</tbody>
</table>

I love this fun film
Naïve Bayes as a Language Model

Which class assigns the higher probability to $s$?

<table>
<thead>
<tr>
<th>Positive Model</th>
<th>Negative Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1 l</td>
<td>0.2 l</td>
</tr>
<tr>
<td>0.1 love</td>
<td>0.001 love</td>
</tr>
<tr>
<td>0.01 this</td>
<td>0.01 this</td>
</tr>
<tr>
<td>0.05 fun</td>
<td>0.005 fun</td>
</tr>
<tr>
<td>0.1 film</td>
<td>0.1 film</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>l</th>
<th>love</th>
<th>this</th>
<th>fun</th>
<th>film</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.1</td>
<td>0.01</td>
<td>0.05</td>
<td>0.1</td>
</tr>
<tr>
<td>0.2</td>
<td>0.001</td>
<td>0.01</td>
<td>0.005</td>
<td>0.1</td>
</tr>
</tbody>
</table>
Naïve Bayes as a Language Model

Which class assigns the higher probability to $s$?

<table>
<thead>
<tr>
<th>Positive Model</th>
<th>Negative Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1 l</td>
<td>0.2 l</td>
</tr>
<tr>
<td>0.1 love</td>
<td>0.001 love</td>
</tr>
<tr>
<td>0.01 this</td>
<td>0.01 this</td>
</tr>
<tr>
<td>0.05 fun</td>
<td>0.005 fun</td>
</tr>
<tr>
<td>0.1 film</td>
<td>0.1 film</td>
</tr>
</tbody>
</table>

$5 \times 10^{-7} \approx P(s|\text{pos}) > P(s|\text{neg}) \approx 1 \times 10^{-9}$
Brill and Banko (2001)

With enough data, the classifier may not matter
Summary: Naïve Bayes is Not So Naïve

Very Fast, low storage requirements

Robust to Irrelevant Features

Very good in domains with many equally important features

Optimal if the independence assumptions hold

Dependable baseline for text classification (but often not the best)
But: Naïve Bayes Isn’t Without Issue

Model the posterior in one go?

Are the features really uncorrelated?

Are plain counts always appropriate?

Are there “better” ways of handling missing/noisy data? (automated, more principled)
Maximum Entropy (Log-linear) Models

\[ p(y | x) \propto \exp(\theta \cdot f(x, y)) \]

*a more general language model*
Maximum Entropy (Log-linear) Models

\[ p(x \mid y) \propto \exp(\theta \cdot f(x, y)) \]

classify in one go
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junin department, central Peruvian mountain region.
Three people have been fatally shot and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junin department, central Peruvian mountain region.
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junin department, central Peruvian mountain region.
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junin department, central Peruvian mountain region.
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junín department, central Peruvian mountain region.
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junín department, central Peruvian mountain region.
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junín, Junín department, central Peruvian mountain region.
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Júnin, Junín department, central Peruvian mountain region.

We need to score the different combinations.
Score and Combine Our Possibilities

\[
\begin{align*}
\text{score}_1 & \text{(fatally shot, ATTACK)} \\
\text{score}_2 & \text{(seriously wounded, ATTACK)} \\
\text{score}_3 & \text{(Shining Path, ATTACK)} \\
& \quad \cdots \\
\text{score}_k & \text{(department, ATTACK)} \\
& \quad \cdots
\end{align*}
\]

are all of these uncorrelated?

\[\text{COMBINE} \quad \text{posterior probability of ATTACK}\]
Score and Combine Our Possibilities

\[ \text{score}_1(\text{fatally shot, ATTACK}) \]
\[ \text{score}_2(\text{seriously wounded, ATTACK}) \]
\[ \text{score}_3(\text{Shining Path, ATTACK}) \]

\[ \text{Q: What are the score and combine functions for Naïve Bayes?} \]
Scoring Our Possibilities

Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junin department, central Peruvian mountain region.

\[
\text{score}(\text{ATTAck}) =
\]

\[
\text{score}_1(\text{fatally shot, ATTAck}) + \text{score}_2(\text{seriously wounded, ATTAck}) + \text{score}_3(\text{Shining Path, ATTAck}) + \ldots
\]
Scoring Our Possibilities

Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junin department, central Peruvian mountain region.

\[
\text{score} \left( \text{ATTACK} \right) = \text{score}_1(\text{fatally shot, ATTACK}) + \text{score}_2(\text{seriously wounded, ATTACK}) + \text{score}_3(\text{Shining Path, ATTACK}) + \ldots
\]

Learn these scores... but how?
What do we optimize?
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junin department, central Peruvian mountain region.

$\text{SNAP}(\text{score}(\text{, ATTACK}))$
What function... operates on any real number?

is never less than 0?
What function...

operates on any real number?

is never less than 0?

\[ f(x) = \exp(x) \]
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junin department, central Peruvian mountain region.
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junín department, central Peruvian mountain region.

Maxent Modeling

\[ p(\text{ATTACK}) \propto \exp(\text{score}_1(\text{fatally shot, ATTACK}) + \text{score}_2(\text{seriously wounded, ATTACK}) + \text{score}_3(\text{Shining Path, attackers})) + \text{...}) \]
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junin department, central Peruvian mountain region.

Learn the scores (but we’ll declare what combinations should be looked at)
Maxent Modeling

Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junin department, central Peruvian mountain region.

\[ p(ATTACK) \propto \exp \left( \sum \text{weight}_i \cdot \text{applies}_i \right) \]
Three people have been fatally shot, and five people, including a mayor, were seriously wounded as a result of a Shining Path attack today against a community in Junin department, central Peruvian mountain region.

Q: How do we define Z?
Normalization for Classification

\[ Z = \sum \exp\left( \text{weight}_1 \cdot \text{applies}_1(\text{fatally shot, } x) \right) + \sum \exp\left( \text{weight}_2 \cdot \text{applies}_2(\text{seriously wounded, } x) \right) + \sum \exp\left( \text{weight}_3 \cdot \text{applies}_3(\text{Shining Path, } x) \right) + \ldots \]

\[ p(x \mid y) \propto \exp(\theta \cdot f(x, y)) \]

classify doc y with label x in one go
Normalization for Language Model

\[ p(y \mid x) \propto \exp(\theta \cdot f(x, y)) \]

general class-based (X) language model of doc y
Normalization for Language Model

\[ p(y \mid x) \propto \exp(\theta \cdot f(x, y)) \]

general class-based (X) language model of doc y

Can be significantly harder in the general case
Normalization for Language Model

\[ p(y \mid x) \propto \exp(\theta \cdot f(x, y)) \]

Can be significantly harder in the general case

Simplifying assumption: maxent n-grams!