433-352 Data on the Web

Semester 2, 2007

Lecturer: Tim Baldwin

THE UNIVERSITY OF MELBOURNE

Lecture 8

Text Categorisation (2)

Bayesian Methods

- Learning and classification methods based on probability theory
- Build a **generative model** that approximates how data is produced
- Categorisation produces a posterior probability distribution over the possible categories given a description of an instance

Jackson and Moulinier (2002:pp129–134); Chakrabarti (2003:pp147–55); McCallum and Nigam (1998)

Bayes' Rule

$$P(C, X) = P(C|X)P(X) = P(X|C)P(C)$$

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)}$$

Jackson and Moulinier (2002:pp129–134); Chakrabarti (2003:pp147–55); McCallum and Nigam (1998)

Naive Bayes (NB) Classifiers

- Task: classify an instance $D = \langle x_1, x_2, ..., x_n \rangle$ according to one of the classes $c_j \in C$

$$
\begin{aligned}
c &= \text{argmax}_{c_j \in C} P(c_j | x_1, x_2, ..., x_n) \\
&= \text{argmax}_{c_j \in C} \frac{P(x_1, x_2, ..., x_n | c_j) P(c_j)}{P(x_1, x_2, ..., x_n)} \\
&= \text{argmax}_{c_j \in C} P(x_1, x_2, ..., x_n | c_j) P(c_j)
\end{aligned}
$$

Jackson and Moulinier (2002:pp129–134); Chakrabarti (2003:pp147–55); McCallum and Nigam (1998)

Simplifying Assumption

- $P(c_j)$
  - ⋆ can be estimated from the frequency of classes in the training examples **[maximum likelihood estimate]**
- $P(x_1, x_2, ..., x_n|c_j)$
  - ⋆ $O(|X|^n|C|)$ parameters (cannot be estimated in practice)
- Naive Bayes Conditional Independence Assumption:
  - ⋆ assume that the probability of observing the conjunction of attributes is equal to the product of the individual probabilities $P(x_i|c_j)$ **[hence "naive"]**

Jackson and Moulinier (2002:pp129–134); Chakrabarti (2003:pp147–55); McCallum and Nigam (1998)

The Final NB Formulation

- Applying the conditional independence assumption:

$$
\begin{aligned}
c &= \text{argmax}_{c_j \in C} P(x_1, x_2, ..., x_n | c_j) P(c_j) \\
&= \text{argmax}_{c_j \in C} P(c_j) \prod_i P(x_i | c_j)
\end{aligned}
$$

Multivariate Binomial NB

- Represent each word as a binary feature ($=$ DF model)
- Represent a document according to the word $types$ it contains
- No indication of how often a given word occurs in a given document
- “Bag of word types” document model

Jackson and Moulinier (2002:pp129–134); Chakrabarti (2003:pp147–55); McCallum and Nigam (1998)

Multivariate Binomial NB: Mechanics

$$P(D|c_i) = \prod_{j=1}^{|\mathscr{V}|} (B_j P(t_j|c_i) + (1 - B_j)(1 - P(t_j|c_i)))$$

where $B_j \in \{0, 1\}$ indicates the presence or absence of the jth term in D, $\mathscr{V}$ is the set of all terms, and

$$P(t|c_i) = \frac{1 + \sum_{k=1}^{|\mathscr{D}|} B_k P(c_i|D_k)}{2 + \sum_{k=1}^{|\mathscr{D}|} P(c_i|D_k)}$$

Jackson and Moulinier (2002:pp129–134); Chakrabarti (2003:pp147–55); McCallum and Nigam (1998)

Multivariate Binomial NB: Example

- Test document:

```
we few, we happy few, we band of brothers
```

- Test document representation:

$$\left\langle \underset{\text{aarvark}}{0}, \underset{\text{aback}}{0}, \ldots, \underset{\text{band}}{1}, \ldots \underset{\text{betwixt}}{0}, \ldots \underset{\text{brothers}}{1}, \ldots, \underset{\text{few}}{1}, \ldots, \underset{\text{happy}}{1}, \ldots \underset{\text{thee}}{0}, \ldots, \underset{\text{we}}{1}, \ldots, \underset{\text{zymogen}}{0} \right\rangle$$

- Shakespeare training document set:

  `then happy i, that love and am beloved`

  $$\langle\ 0,\ \ 0,\ \ldots,\ 0,\ \ldots\ \ 0,\ \ \ldots\ \ 0,\ \ \ldots, 0, \ldots,\ \ 1,\ \ \ldots,\ 0,\ \ldots, 0, \ldots,\ \ \ 0\ \ \rangle$$

  aardvark aback band betwixt brothers few happy thee we zymogen

  `if we shadows have offended`

  $$\langle\ 0,\ \ 0,\ \ldots,\ 0,\ \ldots\ \ 0,\ \ \ldots\ \ 0,\ \ \ldots, 0, \ldots,\ \ 0,\ \ \ldots,\ 0,\ \ldots, 1, \ldots,\ \ \ 0\ \ \rangle$$

  aarvark aback band betwixt brothers few happy thee we zymogen

- Beatles training document set:

  `we can work it out`

  $$\langle\ 0,\ \ 0,\ \ldots,\ 0,\ \ldots\ \ 0,\ \ \ldots\ \ 0,\ \ \ldots, 0, \ldots,\ \ 0,\ \ \ldots,\ 0,\ \ldots, 1, \ldots,\ \ \ 0\ \ \rangle$$

  aardvark aback band betwixt brothers few happy thee we zymogen

  `sgt pepper's lonely hearts club band`

  $$\langle\ 0,\ \ 0,\ \ldots,\ 1,\ \ldots\ \ 0,\ \ \ldots\ \ 0,\ \ \ldots, 0, \ldots,\ \ 0,\ \ \ldots,\ 0,\ \ldots, 0, \ldots,\ \ \ 0\ \ \rangle$$

  aarvark aback band betwixt brothers few happy thee we zymogen

- $P(\texttt{we}|\texttt{Shakespeare}) = \frac{1+(0\times1+1\times1+1\times0+0\times0)}{2+(1+1+0+0)} = \frac{1}{2}$
  $P(\texttt{we}|\texttt{Beatles}) = \frac{1+(0\times0+1\times0+1\times1+0\times1)}{2+(0+0+1+1)} = \frac{1}{2}$

  $P(\texttt{band}|\texttt{Shakespeare}) = \frac{1+(0\times1+0\times1+0\times0+1\times0)}{2+(1+1+0+0)} = \frac{1}{4}$
  $P(\texttt{band}|\texttt{Beatles}) = \frac{1+(0\times0+0\times0+0\times1+1\times1)}{2+(0+0+1+1)} = \frac{1}{2}$

  $P(\texttt{happy}|\texttt{Shakespeare}) = \frac{1+(1\times1+0\times1+0\times0+0\times0)}{2+(1+1+0+0)} = \frac{1}{2}$
  $P(\texttt{happy}|\texttt{Beatles}) = \frac{1+(1\times0+0\times0+0\times0+0\times0)}{2+(0+0+1+1)} = \frac{1}{4}$

- $P(D|\texttt{Shakespeare}) = ((0\times\frac{1}{4}+(1-0)\times\frac{3}{4})\times(0\times\frac{1}{4}+(1-0)\times\frac{3}{4})\times\ldots\times(1\times\frac{1}{4}+(1-1)\times\frac{3}{4})\times\ldots\times(0\times\frac{1}{4}+(1-0)\times\frac{3}{4})\times\ldots\times(1\times\frac{1}{4}+(1-1)\times\frac{3}{4})\times\ldots\times(1\times\frac{1}{4}+(1-1)\times\frac{3}{4})\times\ldots\times(1\times\frac{1}{2}+(1-1)\times\frac{1}{2})\times\ldots\times(0\times\frac{1}{4}+(1-0)\times\frac{3}{4})\times\ldots\times(1\times\frac{1}{2}+(1-1)\times\frac{1}{2})\times\ldots\times(0\times\frac{1}{4}+(1-0)\times\frac{3}{4}))$

Multinomial NB

- Represent each word as an integer
- Represent a document according to the word $tokens$ it contains
- Optionally include a term for $P(L = l_D|c_i)$ (to normalise for document length)
- "Bag of word tokens" document model
- Assumes that (a) the position of a word in the document and (b) the context of a word are irrelevant in classification

Jackson and Moulinier (2002:pp129–134); Chakrabarti (2003:pp147–55); McCallum and Nigam (1998)

Multinomial NB: Mechanics

$$P(D|c_i) = \prod_{j=1}^{|\mathscr{V}|} \frac{P(t_j|c_i)^{N_{D,t_j}}}{N_{D,t_j}!}$$

where N_{D,t_j} is the frequency of the jth term in D, $\mathscr{V}$ is the set of all terms, l_D is the length of D, and

$$P(t|c_i) = \frac{1 + \sum_{k=1}^{|\mathscr{D}|} N_{k,t} P(c_i|D_k)}{|\mathscr{V}| + \sum_{j=1}^{|\mathscr{V}|} \sum_{k=1}^{|\mathscr{D}|} N_{k,t_j} P(c_i|D_k)}$$

Jackson and Moulinier (2002:pp129–134); Chakrabarti (2003:pp147–55); McCallum and Nigam (1998)

Multinomial NB: Example

- Test document:

  `we few, we happy few, we band of brothers`

- Test document representation:

$$\langle \underset{\text{aarvark}}{0}, \underset{\text{aback}}{0}, \ldots, \underset{\text{band}}{1}, \ldots \underset{\text{betwixt}}{0}, \ldots \underset{\text{brothers}}{1}, \ldots, \underset{\text{few}}{2}, \ldots, \underset{\text{happy}}{1}, \ldots \underset{\text{thee}}{0}, \ldots, \underset{\text{we}}{3}, \ldots, \underset{\text{zymogen}}{0} \rangle$$

- Assume $|\mathscr{V}| = 100$

- Shakespeare training document set:

  `then happy i, that love and am beloved`

  $$\langle\ 0,\quad 0,\ \ldots,\ 0,\ \ldots\ 0,\ \ldots\ 0,\ \ldots,0,\ldots,\ 1,\ \ldots,\ 0,\ \ldots,0,\ldots,\ 0\ \rangle$$

  aardvark aback band betwixt brothers few happy thee we zymogen

  `if we shadows have offended`

  $$\langle\ 0,\quad 0,\ \ldots,\ 0,\ \ldots\ 0,\ \ldots\ 0,\ \ldots,0,\ldots,\ 0,\ \ldots,\ 0,\ \ldots,1,\ldots,\ 0\ \rangle$$

  aarvark aback band betwixt brothers few happy thee we zymogen

- Beatles training document set:

  `we can work it out`

  $$\langle\ 0,\quad 0,\ \ldots,\ 0,\ \ldots\ 0,\ \ldots\ 0,\ \ldots,0,\ldots,\ 0,\ \ldots,\ 0,\ \ldots,1,\ldots,\ 0\ \rangle$$

  aardvark aback band betwixt brothers few happy thee we zymogen

  `sgt pepper's lonely hearts club band`

  $$\langle\ 0,\quad 0,\ \ldots,\ 1,\ \ldots\ 0,\ \ldots\ 0,\ \ldots,0,\ldots,\ 0,\ \ldots,\ 0,\ \ldots,0,\ldots,\ 0\ \rangle$$

  aarvark aback band betwixt brothers few happy thee we zymogen

Jackson and Moulinier (2002:pp129–134); Chakrabarti (2003:pp147–55); McCallum and Nigam (1998)

- $P(\texttt{we}|\texttt{Shakespeare}) = \frac{1+(0\times1+1\times1+1\times0+0\times0)}{100+(8+5)} = \frac{2}{113}$
  $P(\texttt{we}|\texttt{Beatles}) = \frac{1+(0\times0+1\times0+1\times1+0\times1)}{100+(5+6)} = \frac{2}{111}$

  $P(\texttt{band}|\texttt{Shakespeare}) = \frac{1+(0\times1+0\times1+0\times0+1\times0)}{100+(8+5)} = \frac{1}{113}$
  $P(\texttt{band}|\texttt{Beatles}) = \frac{1+(0\times0+0\times0+0\times1+1\times1)}{100+(5+6)} = \frac{2}{111}$

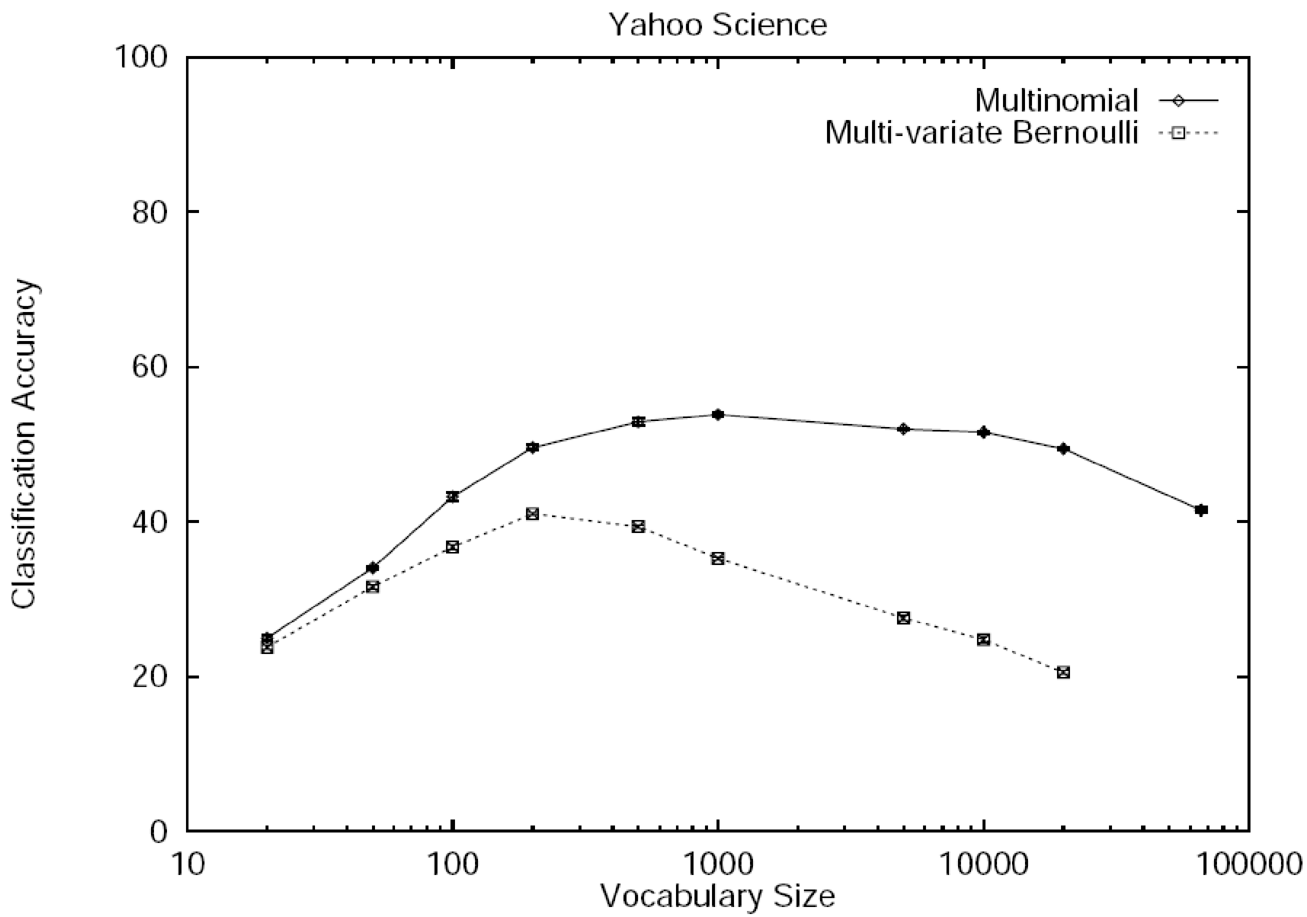
  $P(\texttt{happy}|\texttt{Shakespeare}) = \frac{1+(1\times1+0\times1+0\times0+0\times0)}{100+(8+5)} = \frac{2}{113}$
  $P(\texttt{happy}|\texttt{Beatles}) = \frac{1+(1\times0+0\times0+0\times0+0\times0)}{100+(5+6)} = \frac{1}{111}$

- $P(D|\texttt{Shakespeare}) = \frac{\frac{1}{113}^{0}}{0!} \times \frac{\frac{1}{113}^{0}}{0!} \times \ldots \times \frac{\frac{1}{113}^{1}}{1!} \times \ldots \times \frac{\frac{1}{113}^{0}}{0!} \times \ldots \times \frac{\frac{1}{113}^{1}}{1!} \times \ldots \times \frac{\frac{1}{113}^{2}}{2!} \times$
  $\ldots \times \frac{\frac{2}{113}^{1}}{1!} \times \ldots \times \frac{\frac{1}{113}^{0}}{0!} \times \ldots \times \frac{\frac{1}{113}^{3}}{3!} \times \ldots \times \frac{\frac{1}{113}^{0}}{0!} \times$

Evaluation

- In order to evaluate which of the two models performs best, what vocabulary size to use, etc, we require "held-out" test data
- Evaluation usually takes the form of simple **classification accuracy**
- Average results over multiple "splits" of training and test data for best results

Results over the Yahoo Science Dataset



McCallum and Nigam (1998)

Theoretical Properties of NB Models

- **Multiclass** classification method
- **Parametric**

  only have to store attribute–value counts/probabilities for each class, not the actual instances
- **Incremental**

  easy to add extra data to the classifier on the fly
- Simple ($\rightarrow$ fast)

Practical Properties of NB Models

- Surprising accuracy over text categorisation tasks
- Highly robust over irrelevant features
- Very good at balancing up lots of "marginally relevant" features
- Actual posterior probability estimates tend to be awry, but as a classification task, we are only interested in the relative values
- Multinomial model tends to outperform binomial
- Feature selection more important for binomial model than multinomial model – why?

Real World Practicalities

Extra Features in Text Categorisation

- There's lots more to **web** text categorisation than words:
  - ⋆ metadata
  - ⋆ domain of source page
  - ⋆ page structure
  - ⋆ link structure
  - ⋆ diachronic stability of page
  - ⋆ balance of different content types
  - ⋆ relative use of different HTML attributes
  - ⋆ well-formedness of HTML

⋮

Multi-topic Documents

- Realistically, it is possible for a document to belong to multiple categories
- Ways to model this:
  - ⋆ thresholding
  - ⋆ multiclass categories
  - ⋆ probabilistic class assignment

Summary

- How do Bayesian methods differ from NN methods?
- What are the simplifying assumptions in the NB method?
- What are the two basic variants of the Naive Bayes algorithm, and what are the strengths and weaknesses of each?

References

CHAKRABARTI, SOUMEN. 2003. *Mining the Web: Discovering Knowledge from Hypertext Data*. San Francisco, USA: Morgan Kaufmann.

JACKSON, PETER, and ISABELLE MOULINIER. 2002. *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization*. Amsterdam, Netherlands: John Benjamins.

MCCALLUM, ANDREW, and KAMAL NIGAM. 1998. A comparison of event models for Naive Bayes text classification. In *Proceedings of the AAAI-98 Workshop on Learning for Text Categorization*, Madison, USA.