

# Using Lyrics to Classify Genres

Erin Coughlan  
CS 151 - Artificial Intelligence  
December 20, 2013

## Problem

- Music Information Retrieval
- Song recommendation systems
- Genre classification
- Given a set of song lyrics, what genre does the song belong to

## Previous Approaches

- Samples of songs - no lyrics
- Combine multiple types of data
- Various learners
  - Naïve Bayes' Text Classification
  - Support Vector Machines
  - K-Nearest Neighbors

## Previous Results

- Using only lyrics data:
  - McKay et al - 43%
  - Liang, Gu, and OConnor - 40%
- Combining types of data:
  - McKay et al - 85%

## Data

- Million Song Dataset, Last.fm Dataset, MusiXmatch Dataset

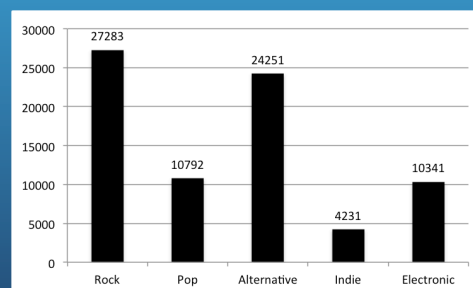
**musiXmatch**<sup>®</sup>  
not only words

**last.fm**

**MILLION  
SONG  
DATASET**

## Data

- 86,237 total songs
  - 76,898 training instances
  - 9,339 testing instances
  - (5000 validation instances)



## Naïve Bayes'

### Set of Words

$x_{1,1}=1$

$y_1=\text{spam}$

|       | a   | and | animal | are | at  | ... | zzyzx | y    |
|-------|-----|-----|--------|-----|-----|-----|-------|------|
| $x_1$ | 1   | 1   | 0      | 1   | 1   | ... | 0     | spam |
| $x_2$ | 1   | 0   | 1      | 0   | 1   | ... | 1     | no   |
| ...   | 1   | 0   | 0      | 0   | 0   | ... | 0     | spam |
| ...   | ... | ... | ...    | ... | ... | ... | ...   | ...  |
| $x_n$ | 1   | 1   | 0      | 1   | 0   | ... | ...   | no   |

$x_{n,2}=1$

$y_n=\text{no}$

Slide from Neil Rhodes, Machine Learning, HMC Fall 2013

## Multinomial Naïve Bayes'

### Bag of Words

$x_{1,1}=2$

$w_2=\text{and}$

$w_n=\text{zzyzx}$

$y_1=\text{spam}$

|       | a   | and | animal | are | at  | ... | zzyzx | y    |
|-------|-----|-----|--------|-----|-----|-----|-------|------|
| $x_1$ | 2   | 1   | 0      | 4   | 1   | ... | ...   | spam |
| $x_2$ | 5   | 0   | 1      | 0   | 2   | ... | ...   | no   |
| ...   | ... | ... | ...    | ... | ... | ... | ...   | ...  |
| $x_k$ | 10  | 8   | 0      | 1   | 0   | ... | ...   | no   |

$x_{k,2}=8$

$y_k=\text{no}$

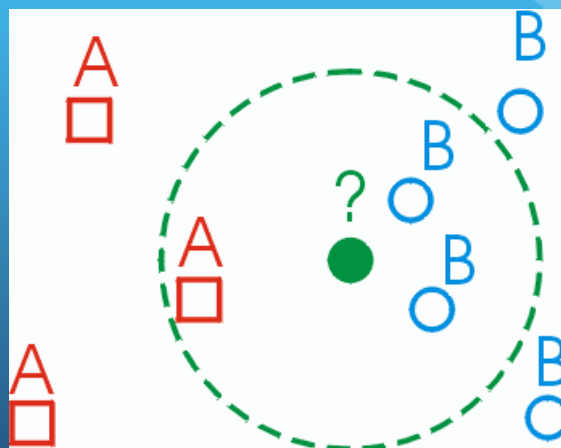
Slide from Neil Rhodes, Machine Learning, HMC Fall 2013

## Results - Naïve Bayes'

|               | Overall | Rock  | Pop   | Alternative | Indie | Electronic |
|---------------|---------|-------|-------|-------------|-------|------------|
| MultinomialNB | 31.12   | 10.62 | 26.49 | 67.70       | 13.97 | 24.89      |
| BinarizedNB   | 32.67   | 12.07 | 22.61 | 78.24       | 4.31  | 20.98      |
| BernoulliNB   | 30.60   | 9.79  | 21.81 | 66.05       | 4.61  | 35.75      |

| Prediction / true | Rock | Pop | Alternative | Indie | Electronic |
|-------------------|------|-----|-------------|-------|------------|
| Rock              | 316  | 368 | 1873        | 45    | 627        |
| Pop               | 113  | 326 | 825         | 24    | 207        |
| Alternative       | 73   | 158 | 1691        | 53    | 585        |
| Indie             | 16   | 29  | 419         | 31    | 178        |
| Electronic        | 35   | 109 | 730         | 14    | 494        |

## K-Nearest Neighbors



## Results - K-Nearest Neighbors

|     | Overall | Rock  | Pop   | Alternative | Indie | Electronic |
|-----|---------|-------|-------|-------------|-------|------------|
| KNN | 30.77   | 35.13 | 17.59 | 46.41       | 4.61  | 18.67      |

| Prediction / true | Rock | Pop | Alternative | Indie | Electronic |
|-------------------|------|-----|-------------|-------|------------|
| Rock              | 1134 | 217 | 1399        | 101   | 378        |
| Pop               | 499  | 263 | 565         | 39    | 129        |
| Alternative       | 767  | 126 | 1188        | 123   | 256        |
| Indie             | 198  | 28  | 308         | 31    | 108        |
| Electronic        | 379  | 258 | 600         | 63    | 258        |

## Summary of Results

|               | Overall | Rock  | Pop   | Alternative | Indie | Electronic |
|---------------|---------|-------|-------|-------------|-------|------------|
| MultinomialNB | 31.12   | 10.62 | 26.49 | 67.70       | 13.97 | 24.89      |
| BinarizedNB   | 32.67   | 12.07 | 22.61 | 78.24       | 4.31  | 20.98      |
| BernoulliNB   | 30.60   | 9.79  | 21.81 | 66.05       | 4.61  | 35.75      |
| KNN           | 30.77   | 35.13 | 17.59 | 46.41       | 4.61  | 18.67      |

## Conclusions

- Only achieves around 30% correctly classified
- Multiple genres in tag dataset
- Incomplete lyrics dataset
- Incorporating additional types of data
- Ensemble techniques

Questions?

## References

- Homburg, H.; Mierswa, I.; Möller, B.; Morik, K.; and Wurst, M. 2005. A benchmark dataset for audio classification and clustering. In *ISMIR*, volume 2005, 528-31.
- Li, T.; Ogihara, M.; and Li, Q. 2003. A comparative study on content-based music genre classification. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, 282-289. ACM.
- Liang, D.; Gu, H.; and O'Connor, B. 2011. Music genre classification with the million song dataset. *Machine Learning Department, CMU*.
- McKay, C.; Burgoyne, J. A.; Hockman, J.; Smith, J. B.; Vigliensoni, G.; and Fujinaga, I. 2010. Evaluating the genre classification performance of lyrical features relative to audio, symbolic and cultural features. In *ISMIR*, 213-218.