Text Classification and Naïve Bayes

Text Classification:
Practical Issues

The Real World

- Gee, I'm building a text classifier for real, now!
- What should I do?

No training data? Manually written rules

If (wheat or grain) and not (whole or bread) then
  Categorize as grain

- Need careful crafting
  - Human tuning on development data
  - Time-consuming: 2 days per class

66

Very little data?

- Use Naïve Bayes
  - Naïve Bayes is a "high-bias" algorithm (Ng and Jordan 2002 NIPS)
- Get more labeled data
  - Find clever ways to get humans to label data for you
- Try semi-supervised training methods:
  - Bootstrapping, EM over unlabeled documents, …

67

A reasonable amount of data?

- Perfect for all the clever classifiers
  - SVM
  - Regularized Logistic Regression
- You can even use user-interpretable decision trees
  - Users like to hack
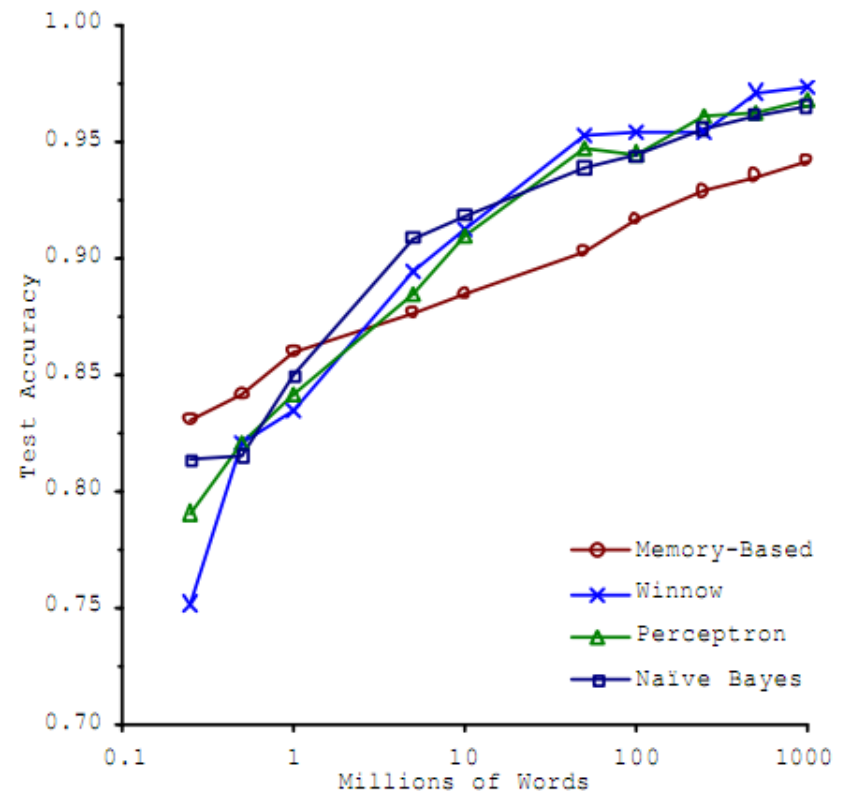 - Management likes quick fixes

A huge amount of data?

- Can achieve high accuracy!
- At a cost:
  - SVMs (train time) or kNN (test time) can be too slow
  - Regularized logistic regression can be somewhat better
- So Naïve Bayes can come back into its own again!

69

Accuracy as a function of data size

- ## With enough data
  - ### Classifier may not matter

70

Brill and Banko on spelling correction

Real-world systems generally combine:

- Automatic classification
- Manual review of uncertain/difficult/"new" cases

Underflow Prevention: log space

- Multiplying lots of probabilities can result in floating-point underflow.
- Since log(xy) = log(x) + log(y)
  - Better to sum logs of probabilities instead of multiplying probabilities.
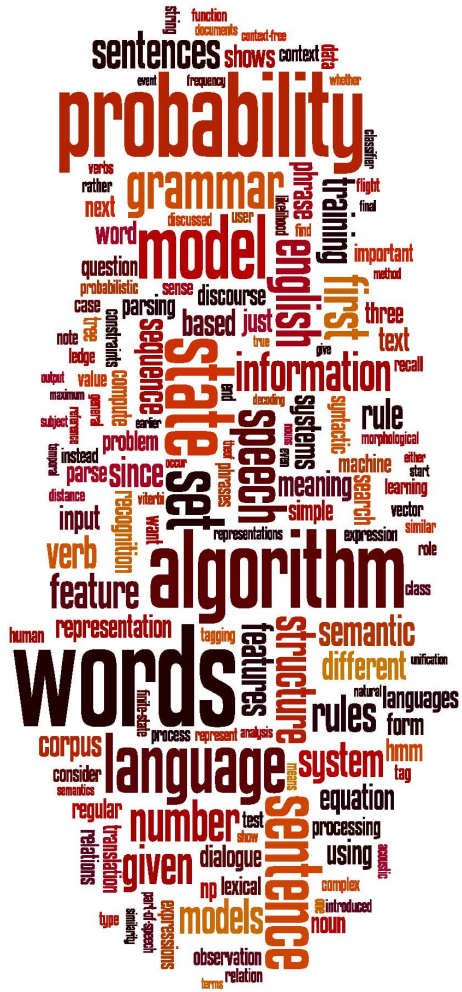- Class with highest un-normalized log probability score is still most probable.

$$c_{NB} = \operatorname*{argmax}_{c_j \in C} \log P(c_j) + \sum_{i \in positions} \log P(x_i \mid c_j)$$

- Model is now just max of sum of weights

How to tweak performance

- Domain-specific features and weights: *very* important in real performance

- Sometimes need to collapse terms:
  - Part numbers, chemical formulas, …
  - But stemming generally doesn't help

- Upweighting: Counting a word as if it occurred twice:
  - title words (Cohen & Singer 1996)
  - first sentence of each paragraph (Murata, 1999)
  - In sentences that contain title words (Ko *et al,* 2002)

73

Text Classification and Naïve Bayes

Text Classification: Practical Issues