

# Naive Bayes (II) and Classifier Evaluation

CS 485, Fall 2026

Applications of Natural Language Processing

Brendan O'Connor

College of Information and Computer Sciences  
University of Massachusetts Amherst



# Naive Bayes overview

- Variables

**w**: a word in the vocabulary  $V$

**x**: document text (words),  $x = (w_1, w_2, w_3, \dots)$

**y**: document's category or label

- Model Parameters

**p(y)**: prior probability of the categories

**p(w | y)**: per-category word probabilities

- Word Independence ("unigram") Assumption:

$$p(x | y) \equiv p(w_1, w_2, \dots, w_n | y) = \prod_{i=1}^n p(w_i | y)$$

- Steps to use

- 1. Training: learn  $p(y)$  and  $p(w|y)$  parameters for all classes and words, based on their counts in labeled training data
- 2. Prediction: But we want **p(y | x)**. Given learned parameters, for new doc, use Bayes Rule to predict posterior probability of class labels

# LM parameter estimation

$$\text{unsmoothed } P(w_i | y) = \frac{\text{count}(w_i, y)}{\sum_{w \in V} \text{count}(w, y)}$$

# LM parameter estimation

$$\text{unsmoothed } P(w_i | y) = \frac{\text{count}(w_i, y)}{\sum_{w \in V} \text{count}(w, y)}$$

what if we see no positive training documents containing the word “awesome”?

$$p(\text{awesome} | \text{POS}) = 0$$

# Add- $\alpha$ (pseudocount) smoothing

$$\text{unsmoothed } P(w_i | y) = \frac{\text{count}(w_i, y)}{\sum_{w \in V} \text{count}(w, y)}$$

$$\text{smoothed } P(w_i | y) = \frac{\text{count}(w_i, y) + \alpha}{\sum_{w \in V} \text{count}(w, y) + \alpha |V|}$$

what happens if we do  
add- $\alpha$  smoothing as  $\alpha$  increases?



# Classifier Evaluation

- Must assess accuracy on held-out data.
  - Train/test split
  - (Alternative: cross-validation)
- Must tune hyperparameters (e.g. pseudocount) on a "development" or "tuning" set.
  - Train/dev/test split

# Classifier Eval Metrics

- For each document: gold-standard labels, compared to predicted labels
- Accuracy: all errors count equally
- For predicting a specific class:  
are **false positives** and **false negatives** always equally important?
  - What about: asymmetric costs, rare classes...

|                             |                 | <i>gold standard labels</i>        |                       |                                               |
|-----------------------------|-----------------|------------------------------------|-----------------------|-----------------------------------------------|
|                             |                 | gold positive                      | gold negative         |                                               |
| <i>system output labels</i> | system positive | <b>true positive</b>               | <b>false positive</b> | <b>precision</b> = $\frac{tp}{tp+fp}$         |
|                             | system negative | <b>false negative</b>              | <b>true negative</b>  |                                               |
|                             |                 | <b>recall</b> = $\frac{tp}{tp+fn}$ |                       | <b>accuracy</b> = $\frac{tp+tn}{tp+fp+tn+fn}$ |

**Figure 4.4** A confusion matrix for visualizing how well a binary classification system performs against gold standard labels.

- Precision =
- Recall =
- F1 = their harmonic mean

# Multiple classes

- What about a 5-way or 100-way classification problem?
- Accuracy is an even more misleading metric
- Each category has its own Prec, Rec, and F1!

|                      |        | <i>gold labels</i> |        |      |
|----------------------|--------|--------------------|--------|------|
|                      |        | urgent             | normal | spam |
| <i>system output</i> | urgent | 8                  | 10     | 1    |
|                      | normal | 5                  | 60     | 50   |
|                      | spam   | 3                  | 30     | 200  |

- How to combine?  
Macroaveraging and Microaveraging

|                      |        | <i>gold labels</i>                            |                                                   |                                                    |                                                       |
|----------------------|--------|-----------------------------------------------|---------------------------------------------------|----------------------------------------------------|-------------------------------------------------------|
|                      |        | urgent                                        | normal                                            | spam                                               |                                                       |
| <i>system output</i> | urgent | 8                                             | 10                                                | 1                                                  | <b>precision<sub>u</sub></b> = $\frac{8}{8+10+1}$     |
|                      | normal | 5                                             | 60                                                | 50                                                 | <b>precision<sub>n</sub></b> = $\frac{60}{5+60+50}$   |
|                      | spam   | 3                                             | 30                                                | 200                                                | <b>precision<sub>s</sub></b> = $\frac{200}{3+30+200}$ |
|                      |        | <b>recall<sub>u</sub></b> = $\frac{8}{8+5+3}$ | <b>recall<sub>n</sub></b> = $\frac{60}{10+60+30}$ | <b>recall<sub>s</sub></b> = $\frac{200}{1+50+200}$ |                                                       |

- How to combine?  
Macroaveraging and Microaveraging