CMPT-825
Natural Language Processing

Anoop Sarkar

http://www.cs.sfu.ca/~anoop

Document Classification/Text Categorization

- There is set of classes C. Put a document into one of $|C|$ classes.
- The only information available are the words in the document.
- We will look at the naive Bayes classifier as a framework for solving this task.

Bayes Rule

- C is a random variable over classes: $c_1, \ldots, c_k, \ldots, c_{|C|}$

- Assume there are $|D|$ documents
  Each document is represented as a vector of attributes: $\mathbf{x_1}, \ldots, \mathbf{x_{|D|}}$
  X is a random variable over the vector of attributes
  $\mathbf{x} = x_1, \ldots, x_j, \ldots, x_d$

- $P(C = c_k \mid X = \mathbf{x_i}) = \frac{P(C=c_k) \times P(X=\mathbf{x_i}|C=c_k)}{P(\mathbf{x_i})}$

- $P(c_k \mid \mathbf{x_i}) = \frac{P(c_k) \times P(\mathbf{x_i}|c_k)}{P(\mathbf{x_i})}$

Naive Bayes Assumption

- $P(c_k \mid \mathbf{x_i}) = \frac{P(c_k) \times P(\mathbf{x_i} | c_k)}{P(\mathbf{x_i})}$

- $P(\mathbf{x_i} \mid c_k) = \prod_{j=1}^{d} P(x_j \mid c_k)$

- $P(c_k \mid \mathbf{x_i}) = P(c_k) \times \prod_{j=1}^{d} P(x_j \mid c_k)$

- Class priors $P(c_k)$ need to be estimated:
  Each class gets the uniform distribution

Naive Bayes Assumption

- $P(c_k \mid \mathbf{x_i}) = P(c_k) \times P(\mathbf{x_i} \mid c_k)$

- θ is the set of parameter values for this model

- A particular setting of the values of these parameters defines a probability of the data

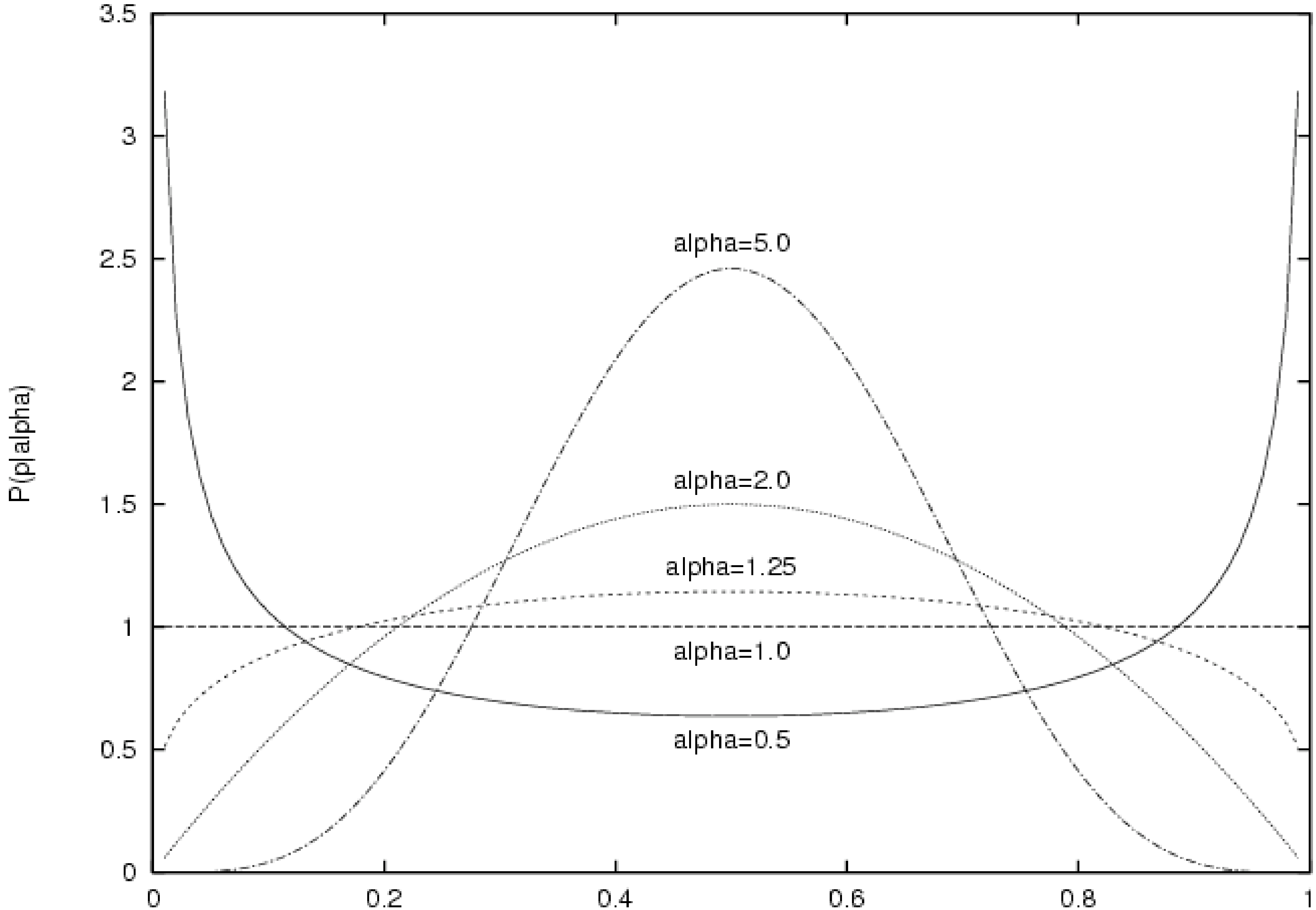
$$P(\mathbf{x_i} \mid \theta) = \sum_{k=1}^{|C|} P(c_k \mid \theta) \times \prod_{j=1}^{d} P(x_j \mid c_k; \theta)$$

Naive Bayes Parameters

- Maximum Likelihood Classifier (ML): $\widehat{\theta} = \underset{\theta}{\text{arg max}}\ P(\mathbf{x_i} \mid \theta)$

- Maximum A-Posteriori Classifier (MAP):
  $\widehat{\theta} = \underset{\theta}{\text{arg max}}\ P(\theta \mid \mathbf{x_i}) = \underset{\theta}{\text{arg max}}\ P(\mathbf{x_i} \mid \theta) \times P(\theta)$
  uses a prior over the parameter values

- Using the prior probability is a good idea.
  MAP classifiers perform better.
  Prior for multinomial distributions: **Dirichlet** prior



P(p|alpha)
alpha=5.0
alpha=2.0
alpha=1.25
alpha=1.0
alpha=0.5
3.5
3
2.5
2
1.5
1
0.5
0
0
0.2
0.4
0.6
0.8
1

Text Representation for Document Classification

- The *bag of words* approach (word order information is lost)
- Two different event models within the bag of words approach:
  - *multi-variate Bernoulli* event model
    (also called Binary Independence Model)
  - *multinomial* event model

Sample Corpus

```
But other than the fact that besuboru is played
with a ball and a bat , it ’s unrecognizable :
Fans politely return foul balls to stadium
ushers ; the strike zone expands depending
on the size of the hitter ; ties are permitted --
even welcomed -- since they honorably
sidestep the shame of defeat ; players must
abide by strict rules of conduct even in their
personal lives -- players for the Tokyo Giants ,
for example , must always wear ties when
on the road .
```

Text Representation for Document Classification

- Start with a vector with dimension equal to size of vocabulary

- *multi-variate Bernoulli* event model

  $1, 0, 1, \ldots$

- *multinomial* event model

  $0, 3, 5, \ldots$

  typical smoothing step: *Laplace prior* add one to count of each word

Naive Bayes Classifier: multi-variate Bernoulli event model

- $\underset{c_k}{\text{arg max}}\ P(c_k \mid \mathbf{x_i}) = \underset{c_k}{\text{arg max}}\ P(c_k) \times P(\mathbf{x_i} \mid c_k)$

- Let the vocabulary V be represented as a vector for each document:
  $\mathbf{x_i} = w_1, \ldots, w_t, \ldots, w_{|V|}$
  $|V|$ is the size of the vocabulary
  if $w_t \in \mathbf{x_i}$ then $B_t = 1$ else $B_t = 0$

$$P(\mathbf{x_i} \mid c_k) = \prod_{t=1}^{|V|} (B_t P(w_t \mid c_k)) + (1 - B_t)(1 - P(w_t \mid c_k))$$

Naive Bayes Classifier: multi-variate Bernoulli event model

- MAP estimate the conditional probability in NB: $\widehat{\theta}_{w_t|c_k}$

$$\widehat{\theta}_{w_t|c_k} = P(w_t \mid c_k; \theta) = \frac{1 + \sum_{i=1}^{|D|} B_t P(c_k \mid \mathbf{x_i})}{2 + \sum_{i=1}^{|D|} P(c_k \mid \mathbf{x_i})}$$

- MAP estimate for the class priors: $\widehat{\theta}_{c_k}$

$$\widehat{\theta}_{c_k} = P(c_k \mid \theta) = \frac{\sum_{i=1}^{|D|} P(c_k \mid \mathbf{x_i})}{|D|}$$

Naive Bayes Classifier: multinomial event model

- Let the vocabulary V be represented as a vector for each document:
  $\mathbf{x_i} = w_1, \ldots, w_t, \ldots, w_{|V|}$
  Let N_{ij} be the frequency of word w_j in document i
  Let L_i be the length of the document represented by $\mathbf{x_i}$

$$P(\mathbf{x_i} \mid c_k) = P(L_i) \times L_i! \times \prod_{t=1}^{|V|} \frac{P(w_t \mid c_k)^{N_{it}}}{N_{it}!}$$

- MAP estimate for conditional probability:

$$\widehat{\theta}_{w_t|c_k} = P(w_t \mid c_k; \theta) = \frac{1 + \sum_{i=1}^{|D|} N_{it} P(c_k \mid \mathbf{x_i})}{|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|D|} N_{is} P(c_j \mid \mathbf{x_i})}$$

Feature selection

- Entropy of a random variable X where $P(X = x)$ is $H(X) = -\sum_{x \in X} P(x) \times \text{log} P(x)$

- Mutual Information between two distributions is $I(X; Y) = H(X) - H(X \mid Y)$

- Feature selection using the mutual information between word (occurence) and the document class: $I(C; \mathbf{X})$

$$\begin{aligned} I(C; \mathbf{X}) &= -\sum_{c_k \in C} P(c_k) log(P(c_k)) + \\ & \sum_j \sum_{c_k \in C} P(c_k \mid w_j) \times \text{log} P(c_k \mid w_j) \end{aligned}$$

Experimental Setup

- Simple accuracy vs. recall/precision

$$\text{Recall} = \frac{\text{num of correct classes proposed}}{\text{num of classes in test data}}$$

$$\text{Precision} = \frac{\text{num of correct classes proposed}}{\text{num of classes proposed}}$$

- Multinomial event model always beats the Bernoulli event model

- Issues with document length

Software: Naive Bayes implementations

- `rainbow`: Naive Bayes classifiers for document classification based on the `bow` (bag of words) model by Andrew McCallum and collaborators.

- Datasets for document classification available from the CMU `textlearning` web page.