Naive Bayes Classifiers and Document Classification

Brandon Malone

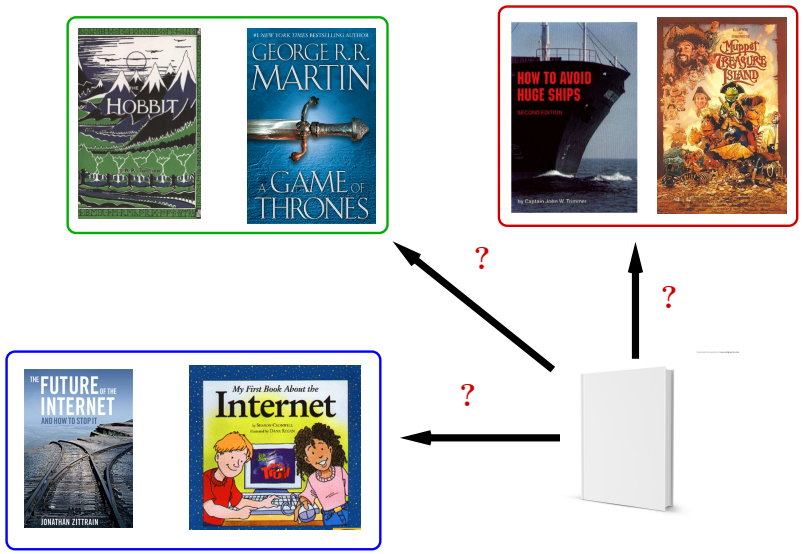
Much of this material is adapted from notes by Hiroshi Shimodaira

Many of the images were taken from the Internet

January 24, 2014

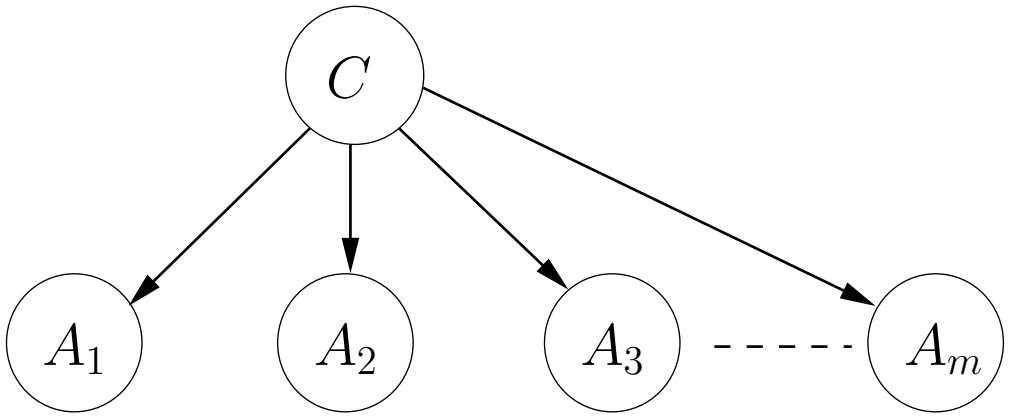
Document Classification

Suppose we have a large number of books. Some are about fantasy, some are about technology, and some are about the high seas.



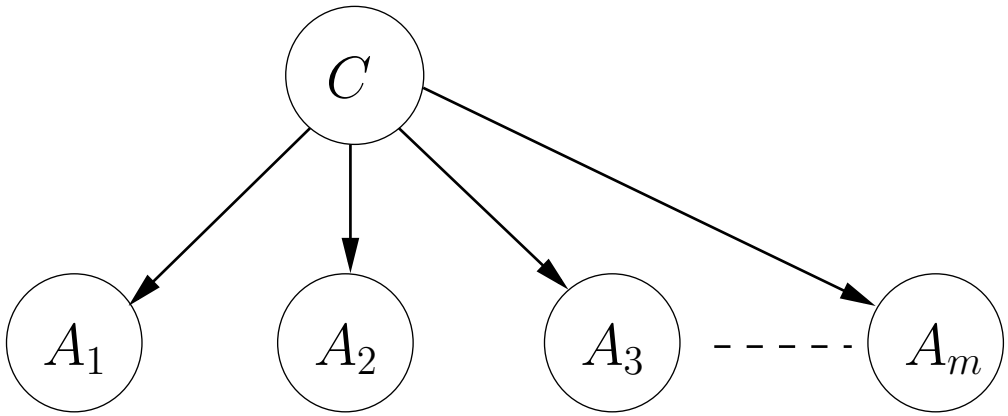
We are given a new book. How can we (automatically) tell which topic the book belongs to?

The Naive Bayes Classifier



What are the conditional independencies asserted by this structure?

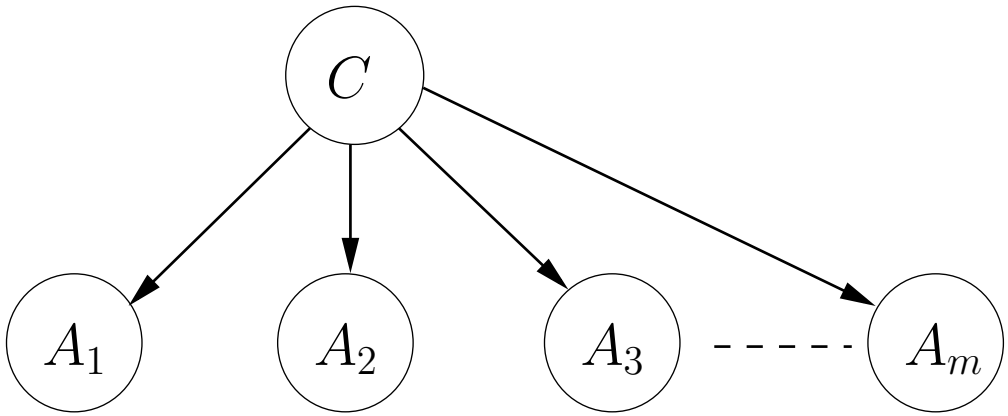
The Naive Bayes Classifier



What are the conditional independencies asserted by this structure?

All of the **attributes** (A_is, sometimes called "features") are independent, given the **class**.

The Naive Bayes Classifier

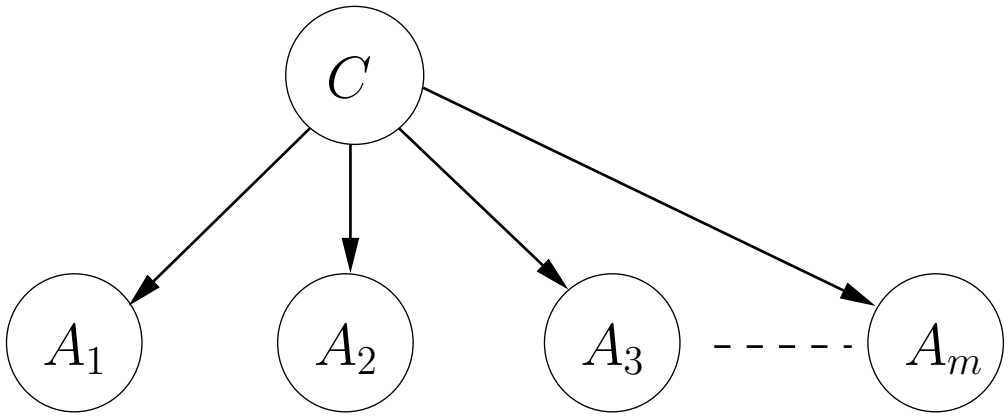


What are the conditional independencies asserted by this structure?

All of the **attributes** (A_is, sometimes called "features") are independent, given the **class**.

If all variables are binary, how many parameters do we need?

The Naive Bayes Classifier



What are the conditional independencies asserted by this structure?

All of the **attributes** (A_is, sometimes called "features") are independent, given the **class**.

If all variables are binary, how many parameters do we need?

1 for the class, plus 2 for each attribute.

Counting distinct permutations

How many distinct sequences can we make?

Counting distinct permutations

How many distinct sequences can we make?

There are 16 letters, so there are $16! \approx 2 \times 10^{13}$ permutations.

Counting distinct permutations

How many distinct sequences can we make?

There are 16 letters, so there are $16! \approx 2 \times 10^{13}$ permutations.

We can choose the "I"s 4! different ways but have the same permutation.

Counting distinct permutations

How many distinct sequences can we make?

There are 16 letters, so there are $16! \approx 2 \times 10^{13}$ permutations.

We can choose the "I"s 4! different ways but have the same permutation.

$$\frac{n!}{n_1!n_2!\ldots n_d!} = \frac{n!}{n_M!n_I!n_S!n_P!n_T!n_A!n_E!} = \frac{16!}{1!4!5!2!2!1!1!} \approx 1.8 \times 10^9$$

Creating a distribution for the items

Suppose we now attach probabilities to each of the d items.

$$\sum_{t=1}^{d} p_t = 1 \qquad p_t > 0, \text{ for all } t$$

We can view creating our sequence as a series of independent draws from this distribution.

If order important, then the probability of our example is

$$p_M \times p_I \times p_S \times p_S \cdots \times p_E = p_M^{n_M} \times p_I^{n_I} \times p_S^{n_S} \ldots p_E^{n_E} = \prod_{t=1}^{d} p_t^{n_d}$$

What if order is not important?

Creating a distribution for the items

What if order is not important?

Say $\mathbf{n} = (n_1, \ldots n_d)$ gives the number of each item we drew. Then

$$P(\mathbf{n}) = P(\text{drawing } \mathbf{n} \text{ one way}) \times \text{number of ways to draw } \mathbf{n}$$

Creating a distribution for the items

What if order is not important?

Say $\mathbf{n} = (n_1, \ldots n_d)$ gives the number of each item we drew. Then

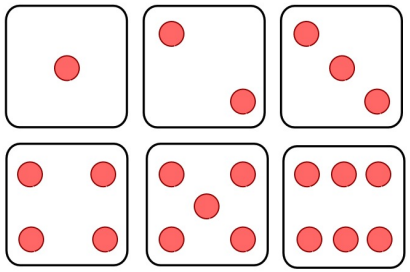
$$P(\mathbf{n}) = P(\text{drawing } \mathbf{n} \text{ one way}) \times \text{number of ways to draw } \mathbf{n}$$

$$P(\mathbf{n}) = \prod_{t=1}^{d} p_t^{n_d} \times \frac{n!}{n_1! n_2! \ldots n_d!}$$

This is called the **multinomial distribution**.

Estimating the probabilities from data

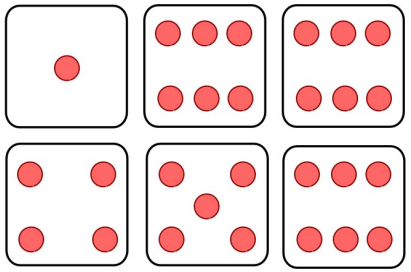
Suppose we roll a die 6 times, and we get...



What probabilities might we attach to each number?

Estimating the probabilities from data

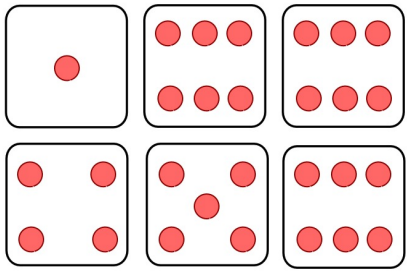
Suppose we roll a die 6 times, and we get...



What probabilities might we attach to each number?

Estimating the probabilities from data

Suppose we roll a die 6 times, and we get...



What probabilities might we attach to each number?

$$p_t = \frac{n_t}{\sum_{u=1}^{d} n_u}$$

These are called the **maximum likelihood parameters**.

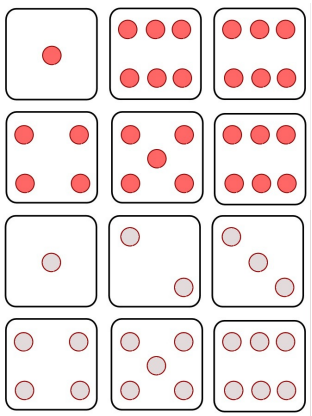
The zero probability problem

Suppose we use the maximum likelihood parameters. What is the probability of rolling a 3?

The zero probability problem

Suppose we use the maximum likelihood parameters. What is the probability of rolling a 3?

A simple correction is to add a "pseudocount" to each item.



$$p_t = \frac{n_t + 1}{d + \sum_{u=1}^{d} n_u}$$

This is sometimes called "smoothing," and we will return to this problem.

Documents as bags of words

We can view documents as a **bag of words**, in which we discard the order among words and simply count occurrences.

So a document D^i is $\mathbf{n}_i = (n_{i,1}, \ldots n_{i,d})$, where $n_{i,t}$ gives the count of word t in D^i.

Our **vocabulary** consists of d words.

A generative model for a document

Suppose we want to create a document (bag of words) of K words.

Further, suppose we are given the distribution for the vocabulary (p_t for each word).

A generative model for a document

Suppose we want to create a document (bag of words) of K words.

Further, suppose we are given the distribution for the vocabulary (p_t for each word).

A simple technique is to draw from the distribution K times.

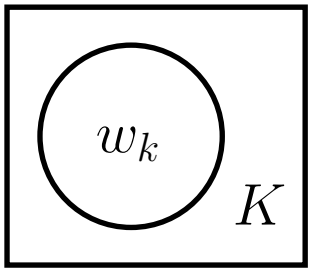


Figure: A simple **generative model** using **plate notation**

Documents about a topic

Suppose I want to write a book about fantasy.

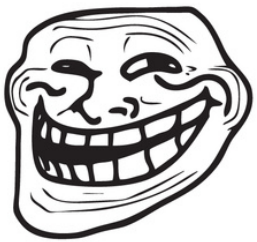
Am I likely to use the same words as if I were writing a book about technology?

Documents about a topic

Suppose I want to write a book about fantasy.

Am I likely to use the same words as if I were writing a book about technology?

Maybe...



Documents about a topic

Suppose I want to write a book about fantasy.

Am I likely to use the same words as if I were writing a book about technology?

... but probably not.

So the probability distribution of my words *depends* upon the topic of my book.

A generative model for documents about a topic

Suppose we want to create a document of K words *about fantasy*.

Further, suppose we are given the distribution for the vocabulary *given that the topic is fantasy* ($P(w_t|C = \text{fantasy})$ for each word).

A generative model for documents about a topic

Suppose we want to create a document of K words *about fantasy*.

Further, suppose we are given the distribution for the vocabulary *given that the topic is fantasy* ($P(w_t|C = \text{fantasy})$ for each word).

A simple technique is to draw from the distribution K times.

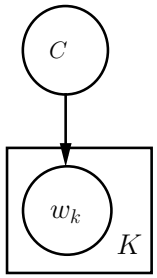


Figure: A conditional generative model using plate notation

We assume the word probabilities are independent given the topic!

This is called the **(multinomial) naive Bayes classifier**.

A generative model for documents about a topic

Suppose we want to create a document of K words *about fantasy*.

Further, suppose we are given the distribution for the vocabulary *given that the topic is fantasy* ($P(w_t|C = \text{fantasy})$ for each word).

A simple technique is to draw from the distribution K times.

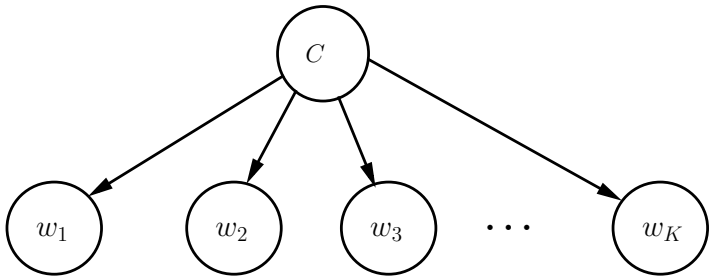


Figure: A conditional generative model as an explicit graphical model

We assume the word probabilities are independent given the topic!

This is called the **(multinomial) naive Bayes classifier**.

Reasoning forward about documents

Suppose we are given a naive Bayes classifier ($Pr(w_t|C)$ for all words and topics and $Pr(C)$ for all topics).

Further, suppose we are given a document $D_i = \mathbf{n}_i$ and are told that it is about fantasy.

What is the **likelihood** of this document, $Pr(\mathbf{n}_i|C = \text{fantasy})$?

Reasoning forward about documents

Suppose we are given a naive Bayes classifier ($Pr(w_t|C)$ for all words and topics and $Pr(C)$ for all topics).

Further, suppose we are given a document $D_i = \mathbf{n}_i$ and are told that it is about fantasy.

What is the **likelihood** of this document, $Pr(\mathbf{n}_i|C = \text{fantasy})$?

$P(\mathbf{n}_i|C = \text{fantasy}) = P(\text{drawing } \mathbf{n}_i \text{ one way}|C = \text{fantasy}) \times \text{number of ways to draw } \mathbf{n}_i$

Reasoning forward about documents

Suppose we are given a naive Bayes classifier ($Pr(w_t|C)$ for all words and topics and $Pr(C)$ for all topics).

Further, suppose we are given a document $D_i = \mathbf{n}_i$ and are told that it is about fantasy.

What is the **likelihood** of this document, $Pr(\mathbf{n}_i|C = \text{fantasy})$?

$$P(\mathbf{n}_i|C = \text{fantasy}) = P(\text{drawing } \mathbf{n}_i \text{ one way}|C = \text{fantasy}) \times \text{number of ways to draw } \mathbf{n}_i$$

$$= \prod_{t=1}^{d} Pr(w_t|C = \text{fantasy})^{n_d} \times \frac{n!}{n_1!n_2!\ldots n_d!}$$

Reasoning backward about documents

Suppose we are given a naive Bayes classifier ($Pr(w_t|C)$ for all words and topics and $Pr(C)$ for all topics).

Further, suppose we are given a document $D_i = \mathbf{n}_i$.

What is the **posterior probability** that this document is about fantasy, $Pr(C = \text{fantasy}|\mathbf{n}_i)$?

Reasoning backward about documents

Suppose we are given a naive Bayes classifier ($Pr(w_t|C)$ for all words and topics and $Pr(C)$ for all topics).

Further, suppose we are given a document $D_i = \mathbf{n}_i$.

What is the **posterior probability** that this document is about fantasy, $Pr(C = \text{fantasy}|\mathbf{n}_i)$?

$$Pr(C = \text{fantasy}|\mathbf{n}_i) = \frac{Pr(\mathbf{n}_i|C = \text{fantasy}) \times Pr(C = \text{fantasy})}{Pr(\mathbf{n}_i)}$$

Reasoning backward about documents

Suppose we are given a naive Bayes classifier ($Pr(w_t|C)$ for all words and topics and $Pr(C)$ for all topics).

Further, suppose we are given a document $D_i = \mathbf{n}_i$.

What is the **posterior probability** that this document is about fantasy, $Pr(C = \text{fantasy}|\mathbf{n}_i)$?

$$Pr(C = \text{fantasy}|\mathbf{n}_i) = \frac{Pr(\mathbf{n}_i|C = \text{fantasy}) \times Pr(C = \text{fantasy})}{Pr(\mathbf{n}_i)}$$

$$= \frac{\prod_{t=1}^{d} Pr(w_t|C = \text{fantasy})^{n_d} \times \frac{n!}{n_1!n_2!\ldots n_d!} \times Pr(C = \text{fantasy})}{Pr(\mathbf{n}_i)}$$

Reasoning backward about documents

Suppose we are given a naive Bayes classifier ($Pr(w_t|C)$ for all words and topics and $Pr(C)$ for all topics).

Further, suppose we are given a document $D_i = \mathbf{n}_i$.

What is the **posterior probability** that this document is about fantasy, $Pr(C = \text{fantasy}|\mathbf{n}_i)$?

$$Pr(C = \text{fantasy}|\mathbf{n}_i) = \frac{Pr(\mathbf{n}_i|C = \text{fantasy}) \times Pr(C = \text{fantasy})}{Pr(\mathbf{n}_i)}$$

$$= \frac{\prod_{t=1}^{d} Pr(w_t|C = \text{fantasy})^{n_d} \times \frac{n!}{n_1!n_2!\ldots n_d!} \times Pr(C = \text{fantasy})}{Pr(\mathbf{n}_i)}$$

Do we need to know the exact probability for classification?

Reasoning backward about documents

Suppose we are given a naive Bayes classifier ($Pr(w_t|C)$ for all words and topics and $Pr(C)$ for all topics).

Further, suppose we are given a document $D_i = \mathbf{n}_i$.

What is the **posterior probability** that this document is about fantasy, $Pr(C = \text{fantasy}|\mathbf{n}_i)$?

$$Pr(C = \text{fantasy}|\mathbf{n}_i) = \frac{Pr(\mathbf{n}_i|C = \text{fantasy}) \times Pr(C = \text{fantasy})}{Pr(\mathbf{n}_i)}$$

$$= \frac{\prod_{t=1}^{d} Pr(w_t|C = \text{fantasy})^{n_d} \times \frac{n!}{n_1!n_2!\ldots n_d!} \times Pr(C = \text{fantasy})}{Pr(\mathbf{n}_i)}$$

Do we need to know the exact probability for classification?

$$Pr(C = \text{fantasy}|\mathbf{n}_i) \propto \prod_{t=1}^{d} Pr(w_t|C = \text{fantasy})^{n_d} \times \frac{n!}{n_1!n_2!\ldots n_d!} \times Pr(C = \text{fantasy})$$

The topic of $\mathbf{n}_i$ is $\arg\max_k Pr(C = k|\mathbf{n}_i)$.

Learning naive Bayes classifiers from data

Suppose we are given N documents and their topics. How can we learn a naive Bayes classifier from this?

- $Pr(C = k)$. The (smoothed) proportion of documents which belong to topic k

$$Pr(C = k) = \frac{N_k + 1}{N + T}$$

- $Pr(w_t|C = k)$. The (smoothed) proportion of the times w_t appears in a document from topic k.

$$Pr(w_t|C = k) = \frac{1 + \text{number of times } w_t \text{ appears in a document from topic } k}{d + \text{number of words in all documents from topic } k}$$

$$Pr(w_t|C = k) = \frac{1 + \sum_{i \text{ such that } z_i = k} \mathbf{n}_{i,t}}{d + \sum_{s=1}^{d} \sum_{i \text{ such that } z_i = k}^{N} \mathbf{n}_{i,t}}$$

- N_k. The number of documents from topic k
- T. The number of topics
- z_i. An indicator which gives the topic k of $\mathbf{n}_i$
- $n_{i,t}$. The number of times word w_t appears in document $\mathbf{n}_i$

Class work

Convert the documents in the **corpus** on the handout into their bag of words representation.

Construct the naive Bayes classifier for the corpus.

Calculate the likelihood, or **conditional distributions**, for each document in the corpus ($Pr(\mathbf{n}_i|C = z_i)$).

Calculate the posterior probability, or **classification distribution**, for the unlabeled documents ($Pr(C = k|\mathbf{n}_i)$).

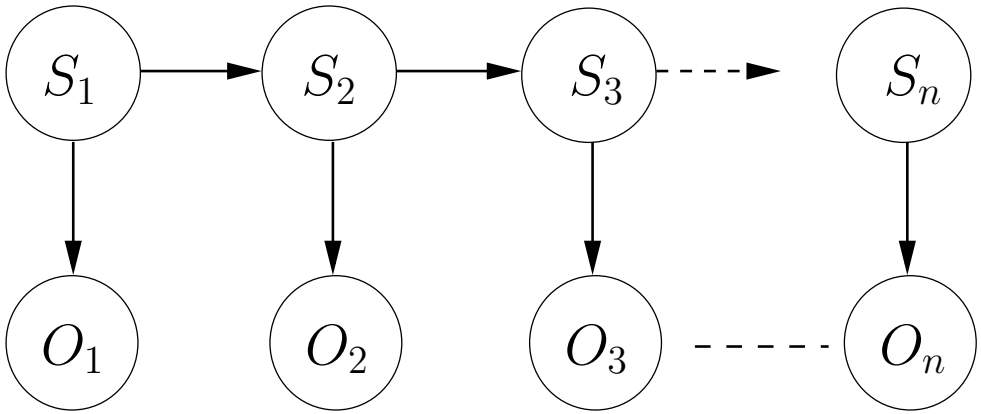
Recap

During this section, we discussed

- The multinomial distribution
- Estimating the (smoothed) parameters for a multinomial distribution from data
- The multinomial bag of words representation of text documents
- Independence assumptions in a naive Bayes classifier (NBC)
- Calculating likelihood using an NBC
- Calculating posterior probability using an NBC
- Learning an NBC from data

Next in probabilistic models

- Markov models for modeling time series and sequences
- Hidden Markov models for gene prediction



- Forward-backward algorithm for finding the most likely instantiation of a set of hidden variables