Spring 2009 **CSC 466:** Knowledge Discovery from Data Alexander Dekhtyar

# Knowledge Discovery in Data: Naïve Bayes

## Overview

**Naïve Bayes** methodology refers to a *probabilistic approach* to information discovery in data. The idea behind the **Naïve Bayes** technique is applicable to the following problems studied in this course:

1. Classification/Supervised Learning.
2. Information Retrieval.
3. Collaborative Filtering/Recommender System.

## Naïve Bayesian Classification

Recall the nature of the *classification problem*:

**Classification Problem.** Given a (training) dataset $D = (A_1, \ldots, A_n, C)$, construct a **classification/prediction function** that correctly predicts the class label $C$ for every record in $D$.

One way of predicting the *class label* $c(d)$ for a record $d \in D$ is to **estimate the probabilities** $Pr(c(d) = c_1), Pr(c(d) = c_2), \ldots, Pr(c(d) = c_k)$, and **pick as the prediction the class with the highest probability**.

**Note:** In general, we *do not need to know* the exact probabilities, it **suffices to know their order (ranks)**.

**Naïve Bayes** is a method of estimating the probabilities (ranks) of a record belonging to each class.

**Classifiers and Observations.** Imagine that a classifier you are building is a black box with an input, which can observe a record from a dataset $D$, and an output, on which *you can observe* the class label $c \in dom(C)$[1]

**Notation.** **Naïve Bayes** method encodes $Pr(c(d) = c_i)$ as the conditional probability of observing $c_i$ in the output when $d$ is observed in the input:

$$Pr(d = c_i) = Pr(c_i|d).$$

$d = (a_1, \ldots, a_n)$. Therefore,

$$Pr(c_i|d) = Pr(c_i|A_1 = a_1, A_2 = a_2, \ldots, A_n = a_n).$$

By definition of **conditional probability**:

$$Pr(X|Y) = \frac{Pr(X \wedge Y)}{Pr(Y)}.$$

**Bayes Theorem** for conditional probabilities:

$$Pr(X|Y) \cdot Pr(Y) = Pr(Y|X) \cdot Pr(X).$$

When $Pr(Y) \neq 0$ and $Pr(X) \neq 0$, Bayes Theorem can be rewritten as:

$$Pr(X|Y) = \frac{Pr(Y|X) \cdot Pr(X)}{Pr(Y)}.$$

**Naïve Bayes Step 1: The Bayes Part (Applying Bayes Theorem):** According to Bayes Theorem, the conditional probability $Pr(c_i|d)$ of observing an object of category $c_i$ given values in vector $d$ can be represented as:

$$Pr(c_i|d) = \frac{Pr(d|c_i) \cdot Pr(c_i)}{Pr(d)},$$

or

$$Pr(c_i|d) = Pr(c_i|A_1 = a_1, \ldots A_n = a_n) = \frac{Pr(A_1 = a_1, \ldots, A_n = a_n|c_i) \cdot Pr(c_i)}{Pr(A_1 = a_1, \ldots A_n = a_n)}.$$

Here,

- $P(d|c_i) = Pr(A_1 = a_1, \ldots, A_n = a_n|c_i)$ is the probability of observing vector $d$ given that the classifier recognized the vector as belonging to class $c_i$. (i.e., the probability of observing $d$ among all vectors of class $c_i$).
- $P(c_i)$ is the probability of observing a vector that belongs to class $c_i$ (textsfprior probability of class $c_i$).

[1] See the Classification/Supervised Learning lecture notes for the notation.

- $P(d)$ is the probability of observing vector $d$ in the input. (prior probability of vector $d$).

**What's the point?** We have just reduced the problem of estimating $Pr(c_i|d)$ to the problem of estimating $P(d|c_i)$ (as well as $Pr(c_i)$ and $Pr(d)$).

**Naïve Bayes Step 2: Simplification.** We need to estimate three probabilities: $Pr(d|c_i), Pr(c_i)$ and $Pr(d)$. We observe the following:

1. **Estimating $Pr(c_i)$.** The probability of observing a vector from class $c_i$ can be estimated as the fraction of the training set $D$ that belongs to this class.

   In other words, let $D_i = \{d \in D | c(d) = c_i\}$. Then,

$$Pr(c_i) = \frac{|D_i|}{|D|}.$$

2. **Dealing with $Pr(d)$.** Vector $d$, when it occurs in the training set can be associated with one of $k$ classes. This means that we can represent $Pr(d)$ as the sum of conditional probabilities of observing $d$ given that a specific class $c_1, \ldots, c_k$ has been observed:

$$Pr(d) = Pr(d|c_1) \cdot Pr(c_1) + \ldots Pr(d|c_k) \cdot Pr(c_k) = \sum_{i=1}^{k} Pr(d|c_i) \cdot Pr(c_i).$$

   **We observe that $Pr(d)$ remains constant among $Pr(c_1|d)$, ..., $Pr(c_k|d)$:**

$$Pr(c_1|d) = \frac{Pr(d|c_1) \cdot Pr(c_1)}{Pr(d)}; Pr(c_2|d) = \frac{Pr(d|c_2) \cdot Pr(c_2)}{Pr(d)}; \ldots; Pr(c_k|d) = \frac{Pr(d|c_k) \cdot Pr(c_k)}{Pr(d)}.$$

   Therefore, ***if we are interested solely in* ranking *the probability estimates*, we can ignore $Pr(d)$ and concentrate on estimating**

$$Pr(d|c_i) \cdot Pr(c_i).$$

**Naïve Bayes: Step 3: the Naïve Part (estimating conditional probability):** We now need to estimate

$$Pr(d|c_i) = Pr(A_1 = a_1, \ldots, A_n = a_n | c_i).$$

The **naïve** part of the method comes from the **conditional independence assumption.**

**Conditional Independence Assumption.** Two random variables $X$ and $Y$ are said to be **conditionally independent** iff for all $x \in dom(X), y \in dom(Y)$,

$$Pr(X = x|Y = y) = Pr(X = x),$$

i.e., if the probability of observing any value of $X$ is not affected by having had observed a specific value of $Y$.

The corrolary to the conditional independence property is that the joint probability of two random variables is equal to the product of their marginal probabilities:

$$Pr(X = x \wedge Y = y) = Pr(X = x) \cdot Pr(Y = y).$$

**Estimating $Pr(d|c_i)$ using conditional independence assumption.** We apply **conditional independence assumption** to

$$Pr(d|c_i) = Pr(A_1 = a_1, \ldots, A_n = a_n|c_i).$$

That is, we assume that each attribute variable $A_i$ is conditionally independent of $A_1, \ldots, A_{i-1}, A_{i+1}, \ldots, A_n$ given $C$, the class variable:

$$Pr(A_i = a_i|A_1 = a_1, \ldots A_{i-1} = a_{i-1}, A_{i+1} = a_{i+1}, \ldots, A_n, C = c_j) = Pr(A_i = a_i|C = c_j).$$

Therefore:

$$Pr(A_1 = a_1, \ldots, A_n = a_n|c_i) = Pr(A_1 = a_1|c_i) \cdot \ldots \cdot Pr(A_n = a_n|c_i) = \prod_{j=1}^{n} Pr(A_j = a_j|c_i).$$

$Pr(A_j = a_j|c_i)$ is the probability of observing a record with $A_j = a_j$ in class $c_i$.

**What's the point?** We have now reduced the problem of estimating the conditional probability $Pr(c_i|d)$ to the problem of estimating the family of probabilities $Pr(A_j = a_j|c_i)$.

**Naı̈̈ve Bayes Step 4: Estimating Probabilities.** We estimate $Pr(A_j = a_j|c_i)$ as follows.

Let $D_i = \{d \in D|c(d) = c_i\}$ be the set of all records of class $c_i$. Let $D_{ij} = \{d \in D_i|d.A_j = a_j\}$ be the set of all records of class $c_i$ with $a_j$ as the value of the attribute $A_j$. Then, our estimate for $Pr(A_j = a_j|c_i)$ is

$$Pr(A_j = a_j|c_i) = \frac{|D_{ij}|}{|D_i|},$$

i.e., the percentage of records in class $c_i$ that have $a_j$ as the value of $A_j$.

**Naïve Bayes Step 5: Predict the class** . Compute estimates

$$Pr(d|c_1) \cdot Pr(c_1), \ldots, Pr(d|c_k) \cdot Pr(c_k),$$

using Steps 1–4 from above.

**Predict:**

$$c(d) = arg \max_{i=1,\ldots,k} (Pr(d|c_i) \cdot Pr(c_i)).$$

### Combining it all together

**Naïve Bayes for classification:**

**1. Use Bayes Theorem.**

$$Pr(c_i|d) = Pr(c_i|A_1 = a_1, \ldots A_n = a_n) = \frac{Pr(A_1 = a_1, \ldots, A_n = a_n|c_i) \cdot Pr(c_i)}{Pr(A_1 = a_1, \ldots A_n = a_n)}.$$

**2. Simplify.**

$$Pr(c_i|A_1 = a_1, \ldots A_n = a_n) \sim Pr(A_1 = a_1, \ldots, A_n = a_n|c_i) \cdot Pr(c_i).$$

**3. Apply Independence Assumption.**

$$Pr(c_i|A_1 = a_1, \ldots A_n = a_n) \sim Pr(A_1 = a_1|c_i) \cdot \ldots \cdot Pr(A_n = a_n|c_i) \cdot Pr(c_i) = Pr(c_i) \cdot \prod_{j=1}^{n} Pr(A_j = a_j|c_i).$$

**4. Estimate Probabilities.**

$$Pr(c_i) = \frac{|D_i|}{|D|}.$$

$$Pr(A_j = a_j|c_i) = \frac{|D_{ij}|}{|D_i|}.$$

$$Pr(c_i|A_1 = a_1, \ldots A_n = a_n) \sim \frac{|D_i|}{|D|} \cdot \prod_{j=1}^{n} \frac{|D_{ij}|}{|D_i|} = \frac{|D_{i1}| \cdot \ldots \cdot |D_{in}|}{|D| \cdot |D_i|^{n-1}}.$$

**5. Predict.**

$$c(d) = arg \max_{i=1,\ldots,k} (Pr(d|c_i) \cdot Pr(c_i)) = arg \max_{i=1,\ldots,k} \frac{|D_{i1}| \cdot \ldots \cdot |D_{in}|}{|D| \cdot |D_i|^{n-1}}.$$

## Naïve Bayes for Information Retrieval

The **Naiïve Bayes** model for Information Retrieval is commonly referred to as **Probabilistic IR**, **Binary Independence Retrieval** or Statistical Language Model.

**IR Problem.** Recall that the main question IR studies is formulated as follows:

> Given a document collection $D$ and a query $q$ find all documents in $D$ that are **relevant** to $q$.

**IR and Classification.** The IR problem can be viewed as a classification problem of the following form:

> Given information about a user query $q$ and some document $d \in D$, classify $d$ as either **relevant to** $q$ or **not relevant to** $q$.

**Note:** In actuality, this is a **partially supervised learning** problem: we **know** the classes but we **do not have** (at the outset) a training set.

**IR and probabilities.** We can transform the problem of classifying a document $d$ as **relevant** or **not relevant** into a problem of estimating the probabilities

$$Pr(R|d,q) \text{ and } Pr(N|d,q) \qquad (Pr(N|d,q) = 1 - Pr(R|d,q))\,.$$

Here,

$Pr(R|d,q)$ is the probability of classifying document $d$ as **relevant** for query $q$.
$Pr(N|d,q)$ is the probability of classifying document $d$ as **not relevant** for query $q$.

**Step 0: Choose IR model.** In lieu of actual document $d$ and query $q$, we use their **binary vector representations**: $d = (w_1, \ldots, w_N)$, $q = (q_1, \ldots, q_N)$, where $d_i = 1$ ($q_i = 1$) if term $t_i$ is in $d$ ($q$) and is 0 otherwise.

**Step 1: Apply the Bayes Theorem.**

$$Pr(R|d,q) = \frac{Pr(d|R,q) \cdot Pr(R|q)}{Pr(d|q)}.$$

$$Pr(N|d,q) = \frac{Pr(d|N,q) \cdot Pr(N|q)}{Pr(d|q)}.$$

Here,

- $Pr(d|R,q)$ is the probability of a relevant to $q$ document being $d$ (the probability of observing $d$ given that a relevant document was returned in response to query $q$);
- $Pr(R|q)$ is the probability of retrieving a relevant document given query $q$;
- $Pr(d|q)$ is the probability of retrieving $d$ in response to $q$;
- $Pr(d|N,q)$ is the probability of a not relevant to $q$ document being $d$ (the probability of observing $d$ given that a non-relevant document was returned, and in response to query $q$);
- $Pr(N|q)$ is the probability of retrieving a not relevant document given query $q$.

Note, that we assume $P(d|q) \neq 0$.

**Step 2. Switch to Odds Ratio.** **Naïve Bayes** method for classification switches from estimating original conditional probability to estimating just the numerator of the fraction.

In the case of IR, we only have two classes, so we can switch to estimating the **odds ratio** of $d$ being relevant vs. not relevant given query $q$ to achieve the same basic effect of not needing to estimate $Pr(d|q)$:

$$O(R|d,q) = \frac{Pr(R|d,q)}{Pr(N|d,q)} = \frac{\frac{Pr(d|R,q) \cdot Pr(R|q)}{Pr(d|q)}}{\frac{Pr(d|N,q) \cdot Pr(N|q)}{Pr(d|q)}} = \frac{Pr(d|R,q) \cdot Pr(R|q)}{Pr(d|N,q) \cdot Pr(N|q)}.$$

**Step 3. Simplify the Odds Ratio.** Note, that

$$\frac{Pr(R|q)}{Pr(N|q)},$$

the odds ratio of retrieving a relevant vs. not relevant document given query $q$ **does not depend on document** $d$, and therefore is a **constant** for each query $q$.

Thus, we reduce our task to estimating the ratio

$$\frac{Pr(d|R,q)}{Pr(d|N,q)} \sim O(R|d,q).$$

**Step 4. The Naïve Assumption.** We assume **conditional independence** of terms in $d$ given $q$. This allows us to use the following substitutions:

$$Pr(d|R,q) = Pr(d[1] = w_1, \ldots, d[N] = w_N|R,q) = \prod_{i=1}^{N} Pr(d[i] = w_i|R,q).$$

$$Pr(d|N,q) = Pr(d[1] = w_1, \ldots, d[N] = w_N|N,q) = \prod_{i=1}^{N} Pr(d[i] = w_i|N,q).$$

Therefore, we obtain,

$$\frac{Pr(d|R,q)}{Pr(d|N,q)} = \prod_{i=1}^{N} \frac{Pr(d[i] = w_i|R,q)}{Pr(d[i] = w_i|N,q)},$$

or

$$O(R|d,q) = O(R|q) \cdot \prod_{i=1}^{N} \frac{Pr(d[i] = w_i|R,q)}{Pr(d[i] = w_i|N,q)}.$$

**Step 5. Separate probabilities by term occurrence/absense.** We note the $w_i$ in the equation above can take only two values: $1$ (term $t_1$ is in the document) and $0$ (term $t_i$ is not in the document). We can, then rewrite the last formula as follows:

$$O(R|d,q) = O(R|q) \cdot \prod_{i:w_i=1} \frac{Pr(d[i] = 1|R,q)}{Pr(d[i] = 1|N,q)} \cdot \prod_{i:w_i=0}^{N} \frac{Pr(d[i] = 0|R,q)}{Pr(d[i] = 0|N,q)}.$$

**Step 6. Perform Information Retrieval (term matching).** Consider the following notation:

Denote as $p_i$ the probability, $Pr(d[i] = 1|R,q)$, of a document, relevant to $q$ containing term $t_i$.

Denote as $u_i$ the probability, $Pr(d[i] = 1|R,q)$, of a document, not relevant to $q$ containing term $t_i$.

Assuming this notation, we can construct the following probability matrix:

| Document: | Relevant to $q$ (R) | Not relevant to $q$ (N) |
|---|---|---|
| Term present: $w_i = 1$ | $p_i$ | $u_i$ |
| Term absent: $w_i = 0$ | $1 - p_i$ | $1 - u_i$ |

We can rewrite our estimate of the odds ratio in the new terms as follows:

$$O(R|d,q) = O(R|q) \cdot \prod_{i:w_i=1} \frac{p_i}{u_i} \cdot \prod_{i:w_i=0} \frac{1-p_i}{1-u_i}.$$

(this, just changes the notation, nothing else.)

**Now,** make the following ***simplifying assumption***:

> If term $t_i$ is **not present** in query $q$, i.e., if $q_i = 0$, then assume
>
> $$p_i = u_i.$$
>
> (Terms not in query have equal chance of appearing in relevant and non relevant documents).

Using this assumption, we can filter out some of the $\frac{p_i}{u_i}$ and $\frac{1-p_i}{1-u_i}$ terms from the odds ratio formula:

$$O(R|d,q) = O(R|q) \cdot \prod_{i:q_i=w_i=1} \frac{p_i}{u_i} \cdot \prod_{i:q_i=1,w_i=0} \frac{1-p_i}{1-u_i}.$$

We can rewrite this formula as:

$$O(R|d,q) = O(R|q) \cdot \prod_{i:q_i=w_i=1} \frac{p_i(1-u_i)}{u_i(1-p_i)} \cdot \prod_{i:q_i=1} \frac{1-p_i}{1-u_i}.$$

Note, that in such form, the expression

$$\prod_{i:q_i=1} \frac{1-p_i}{1-u_i}$$

is a constant given a query (it does not depend on the document). So,

$$O(R|d,q) = K_q \cdot \prod_{i:q_i=w_i=1} \frac{p_i(1-u_i)}{u_i(1-p_i)},$$

where $K_q = O(R|q) \cdot \prod_{i:q_i=1} \frac{1-p_i}{1-u_i}$ is constant for $q$.

**Step 7. Switch to Retrieval Status Value.** Our goal is to estimate

$$\prod_{i:q_i=w_i=1} \frac{p_i(1-u_i)}{u_i(1-p_i)}.$$

We can, instead, estimate the **retrieval status value** of document $d$ w.r.t. query $q$, denoted as $RSV_d$ and defined as follows:

$$RSV_d = \log\left(\prod_{i:q_i=w_i=1} \frac{p_i(1-u_i)}{u_i(1-p_i)}\right) = \sum_{i:q_i=w_i=1} \log \frac{p_i(1-u_i)}{u_i(1-p_i)}.$$

**Note:** By switching to logarithms, **we preserve monotonicity** of our estimates ($O(R|d,q) > O(R|d',q)$ **iff** $RSV_d > RSV_{d'}$). However, we replace the **product** with the **sum**.

We define $c_i$ to be the **log odds ratios for the term** $t_i$ **in the query** $q$:

$$c_i = \log \frac{p_i(1-u_i)}{u_i(1-p_i)} = \log \frac{p_i}{1-p_i} + \log 1 - u_i u_i.$$

**Similarity computation.** We compute the similarity between a document and a query as its Retrieval Status Value:

$$sim(d,q) = RSV_d = \sum_{i:q_i=w_i=1} c_i.$$

**Step 8. Parameter Estimates** We need a way of estimating $c_i$s or their $p_i$ and $u_i$ components.

**Theory.** Suppose we are able to observe the set $D_q$ of all document **relevant** to query $q$. Let $|D_q| = S$. Additionally, let $D_{qi} = \{d \in D_q | d[i] = 1\}$ and let $|D_{qi}| = s$ (i.e., $s$ out of $S$ relevant documents contain term $t_i$).

Then, we can estimate $p_i$ and $u_i$ as follows:

$$p_i = \frac{s}{S-s}$$

$$u_i = \frac{df_i - s}{|D| - df_i - s + S}$$

$$c_i = \log \frac{s \cdot (|D| - df_i - s + S)}{(df_i - s) \cdot (S - s)}.$$

**Smoothing.** We can *smoothe* the $c_i$ estimate to avoid zeroes:

$$c_i = \log \frac{(s+0.5) \cdot (|D| - df_i - s + S + 0.5)}{(df_i - s + 0.5) \cdot (S - s + 0.5)}.$$

**Practice.** In practice, the **answer set** (i.e., the list of relevant documents) is rarely available (especially, since queries are dynamic).

The estimates are **kludged** in the following manner:

$$\log \frac{1-u_i}{u_i} = \log \frac{|D| - df_i}{df_i} \approx \log |D| df_i = idf_i.$$

This assumes that the number of relevant documents is **much smaller** than the total size of the collection (and thus, $|D| - df_i$ is almost $|D|$.)

**Estimating $p_i$.** Estimating $p_i$s is **harder**. In practive the estimates are **kludged** in the following manner:

- Set $p_i = c$ for some value $c \in (0, 1)$. Typical estimate of this sort is $p_i = 0.5$.
- Tie $p_i$ to $df_i$:

$$p_i = \frac{1}{3} + \frac{2}{3} \cdot \frac{df_i}{N}.$$

## References