Text Classification in IR

Reading:
Textbook, Chapter 13

April 20, 2010

1

Text classification: Why?

2

Example: Detecting spam

From: Alibris <books@alibris.m0.net>
Reply-to: books@alibris.m0.net
To: mm@cse.ogi.edu
Subject: Melanie, reminding you to save $10 at Alibris

HOLIDAY SPECIAL: SAVE UP TO $10 ON YOUR PURCHASES
(order now and receive by Christmas)

With the holiday season rapidly approaching, we want to remind you of our most generous sale of the year. As a valued customer, we invite you to save up to $10 off your Alibris purchases with three ways to save:

$2 off your order of $20 or more: GIFT2
$5 off your order of $50 or more: GIFT5
$10 off your order of $100 or more: GIFT10

Simply enter the coupon codes above* at checkout. But hurry, this limited time offer expires on December 16, 2003. Visit Alibris now and save!

Save money on shipping too! Now through December 9, 2003, every item listed on our site should be delivered to continental U.S. and Canadian addresses by December 24th via standard shipping (our lowest cost option) or get FREE shipping when you order $49 of In Stock books.
Don't delay, start your holiday shopping now.
http://alibris.m0.net/m/S.asp?HB10950943733X2869462X274232X

```
From: "Basil Lutz" <0eynsozueb@a-city.de>
Reply-To: "Basil Lutz" <0eynsozueb@a-city.de>
To: <mm@santafe.edu>, <bonabeau@santafe.edu>
Subject: **SPAM 10.70** This tool will make your website more productive hukm

<html>
<head>
<title>hd36 8 ekj 0O9 92O 2                          </title>
<meta http-equiv=3D"Content-Type" content=3D"text/htm; charset=3Diso-8859=
-1">
</head>

<body>
<p><font face=3D"Arial, Helvetica, sans-serif">Can your website answer que=
stions
  in real time 24 hours a day, 7 days a week? Our clients websites do and =
we're
  not talking about some stale FAQ sheet either. Add <a href=3D"http://www=
dreamscaper.co.mn@click.net-click.net.ph/click.php?id=3Ddrcomnm">live
  operator support</a> to your website today and dramatically increase you=
r revenues.</font></p>
<p> </p>
<p> </p>
<p> </p>
<p> </p>
<p> </p>
<p> </p>
<p> </p>
<p> </p>
<p> </p>
<p><a href=3D"http://www.dreamscaper.co.mn@click.net-click.net.ph/click.ph=
p?id=3Ddrcomnx">stop</a>
  sending me emails</p>
</body>
</html>
```

```
From: =?iso-8859-1?q?james=20ken?= <ja_ken2004@yahoo.fr>
Subject: URGENT ASSISTANCE
To: ja_ken2004@yahoo.fr


 FROM:JAMES KEN.

ATTN:

Dear Respectful one,


I know this proposal letter may come to you as a
surprise considering the fact that we have not had any
formal acquaintance before .but all the same I would
want you for the sake of God to give this an immediate
attention in view of the fact that the security of our
live and possession is at stake .

I am Mr JAMES KEN 28 years old from war ravaged SIERRA
LEONE but presently domiciled in Abidjan  Ivory coast
with my sister JANET who is 18 years old .My father Mr
KEN who before his untimely assassination by the
rebels was the Director of SIERRA LEONE Diamond
corporation (SLDC) .He was killed in our government
residential house along side two of my other brothers
,two house maids and one government attached security
guard fortunately for I, younger sister and mother ,we
were on a week end visit to our home town As we got
the news of the tragedy .We immediately managed to ran
into neighbouring Ivory coast for refuge .But
unfortunately .As Fate would have it ,we lost our dear
mother (may soulrest in peace) as a result of what the
Doctor called cardiac arrest .
```

As we were coming into this country ,we had some documents of a deposit of $ 11 700 000 USD (eleven million seven hundred thousand USD) made by my late father in a security and trust company .According to my father, he intended to use this fund for his international business transaction after his tenure in office but was unfortunately murdered .We had located the security company where the money is deposited with the help of an attorney and established ownership .please right now ,with the bitter experiences we had in our country and the war still going on especially in diamond area which incidentally is where we hail from .coupled with the incessant political upheavals and hostilities in this country Ivory coast ,we desire seriously to leave here and live the rest of our life into a more peaceful and politically stable country like yours Hence this proposal and request .We therefore wish you can help us in the following regards :

1)To provide us with a good bank account to transfer the money into.
2)To help us invest the money into a lucrative business .
3)To assist my sister Janet get a college admission to further her education.

Please I know that , this letter may sound strange and incredible to you but the CNN and the BBC African bulletin normally have it as their major news features .Therefore for the sake of God and humanity give an immediate positive consideration and reply to me via our e-mail address. I will willingly agree to any suitable percentage of the money you will propose as your compensation for your assistance with regards to the above .please in view of our sensitive refugee status and as we are still conscious of our father 's enemies .I would like you to give this a highly confidential approach .

Best Regards .
JAMES KEN.

Spamassassin results

```
X-Spam-Report: ---- Start SpamAssassin results
       6.70 points, 4 required;
       *  0.4 -- BODY: Offers a limited time offer
       *  0.1 -- BODY: Free Offer
       *  0.4 -- BODY: Stop with the offers, coupons, discounts etc!
       *  0.1 -- BODY: HTML font color is red
       *  0.1 -- BODY: Image tag with an ID code to identify you
       *  2.8 -- BODY: Bayesian classifier says spam probability
              is 80 to 90% [score: 0.8204]
       *  0.8 -- BODY: HTML font color is green
       *  0.3 -- BODY: FONT Size +2 and up or 3 and up
       *  0.1 -- BODY: HTML font color not within safe 6x6x6 palette
       *  0.1 -- BODY: HTML font color is blue
       *  0.3 -- BODY: Message is 70% to 80% HTML
       *  1.2 -- Date: is 6 to 12 hours after Received: date
       ---- End of SpamAssassin results
```

Spamassassin results

```
X-Spam-Report: ---- Start SpamAssassin results
       10.70 points, 4 required;
       *  0.4 -- BODY: Message is 40% to 50% HTML
       *  1.0 -- URI: URL contains username and (optional) password
       *  0.8 -- URI: Uses a username in a URL
       *  1.2 -- RBL: Received via a relay in dnsbl.njabl.org
                 [RBL check: found 78.199.241.24.dnsbl.njabl.org.,]
                 [type: 127.0.0.9]
       *  4.3 -- RBL: Received via a relay in list.dsbl.org
                 [RBL check: found 78.199.241.24.list.dsbl.org.]
       *  0.1 -- Message has X-MSMail-Priority, but no X-MimeOLE
       *  0.1 -- Message only has text/html MIME parts
       *  2.8 -- Forged mail pretending to be from MS Outlook IMO
       ---- End of SpamAssassin results
```

Spamassassin results

```
X-Spam-Report: ---- Start SpamAssassin results
       13.30 points, 4 required;
       *  0.7 -- From: ends in numbers
       *  1.4 -- Subject is indicative of a Nigerian spam
       *  1.5 -- BODY: Nigerian scam key phrase (million dollars)
       *  2.7 -- BODY: Contains urgent matter
       *  2.8 -- BODY: Bayesian classifier says spam probability is
              80 to 90% [score: 0.8584]
       *  0.7 -- Subject is all capitals
       *  0.7 -- From: contains an underline and numbers/letters
       *  2.8 -- Message body has multiple indications of Nigerian spam
       ---- End of SpamAssassin results
```

Automatic text classification: How?

- **Needed:**
  - Function mapping document into feature vector:

    $F_{feature\ extraction}(\mathbf{x})=(x_1, x_2, \ldots, x_n)$

  - Training examples:

    $S = \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^m)$

  - Learning method L:

    L: {S}→ {Classifiers}

  - Classifier C:

    C: {**x**}→ {classes}

12

- **Usual assumptions for text classification:**
  - Each document belongs to exactly one class
  - Features are *terms t* in the *vocabulary V*
  - Position of terms in a document is ignored
  - *Inductive learning hypothesis* holds

13

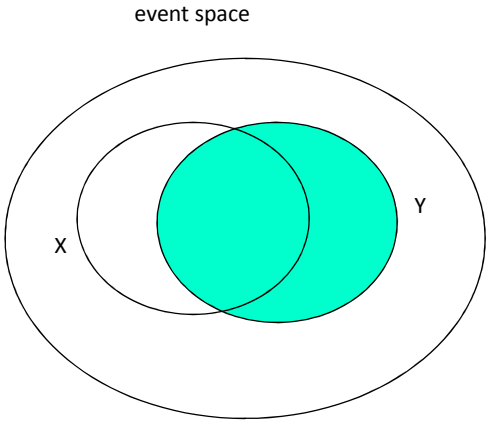
Review of Conditional Probability

- Probability of an event given the occurrence of some other event.

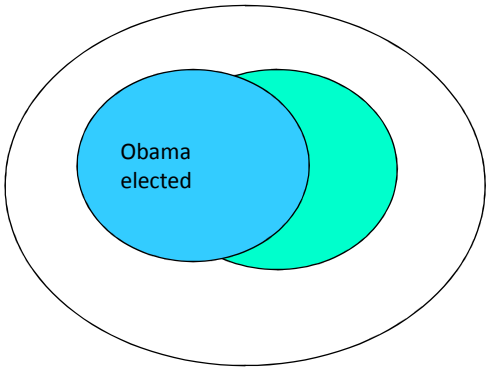
E.g.,

What is the probability that a liberal Supreme Court Justice will be appointed in the next four years, given that Barack Obama was elected President?

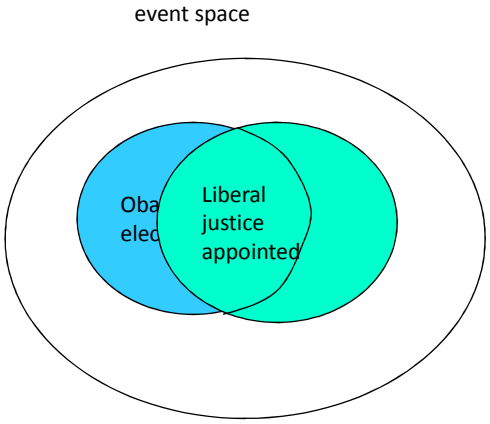
$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)}$$



$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)}$$

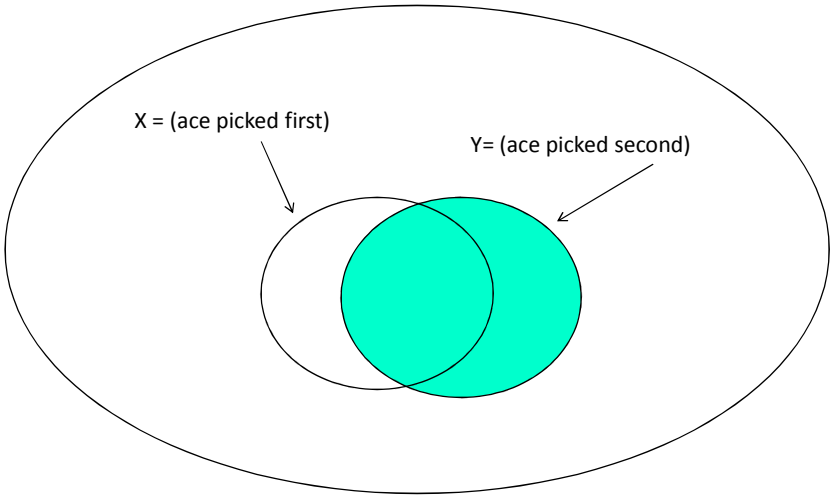


$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)} = \frac{P(X, Y)}{P(Y)}$$

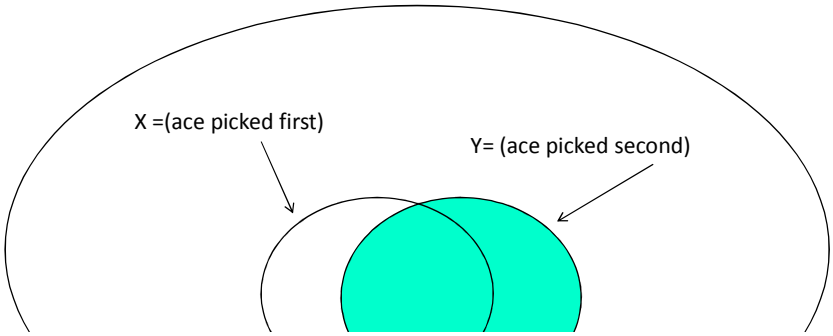


- Consider choosing a card from a well-shuffled standard deck of 52 playing cards. Given that the first card chosen is an ace, what is the probability that the second card chosen will be an ace?

event space: all possible pairs of cards



event space: all possible pairs of cards



P(X,Y) = number of pairs of aces / total number of pairs
= 4x3/52x51 = 12/2652.

P(Y) = 4 / 52
P(X | Y) = (12/2652) / (4 / 52) = 3/51.

Relationships among joint, conditional, posterior, and marginal probabilities

$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)} = \frac{P(X, Y)}{P(Y)}$$

$$P(X, Y) = P(X \mid Y)P(Y) = P(Y \mid X)P(X)$$

Bayes rule :

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}$$

The Monty Hall Problem

You are a contestant on a game show.

There are 3 doors, A, B, and C. There is a new car behind one of them and goats behind the other two.

Monty Hall, the host, asks you to pick a door, any door. You pick door A.

Monty tells you he will open a door , different from A, that has a goat behind it. He opens door B: behind it there is a goat.

Monty now gives you a choice: Stick with your original choice A or switch to C.

Should you switch?

http://math.ucsd.edu/~crypto/Monty/monty.html

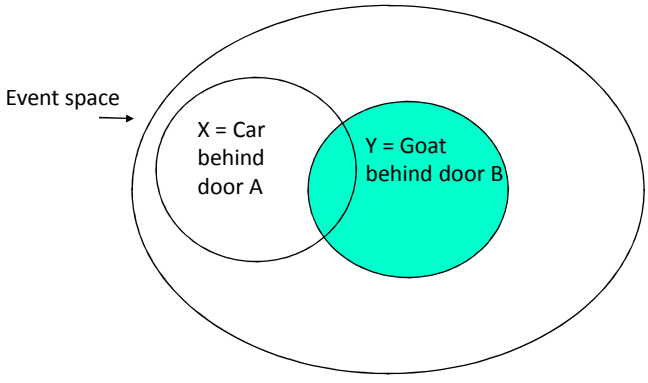
Bayesian probability formulation

Hypothesis space H:

h_1 = Car is behind door A
h_2 = Car is behind door B
h_3 = Car is behind door C

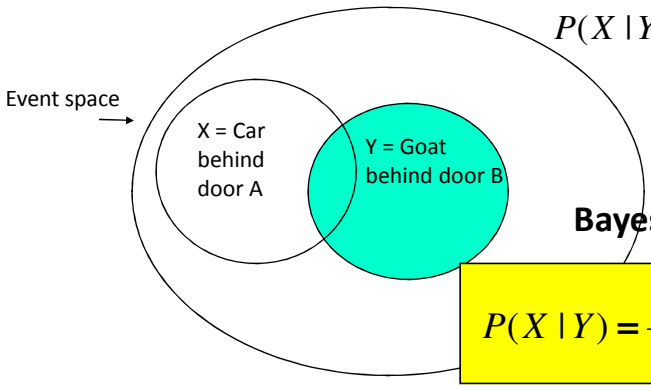
Data D = Monty opened B

What is $P(h_1 \mid D)$?
What is $P(h_2 \mid D)$?
What is $P(h_3 \mid D)$?



Event space = All possible configurations of cars and goats behind doors A, B, C

$$P(X \mid Y) = \frac{P(X \cap Y)}{P(Y)}$$



Bayes Rule:

$$P(X \mid Y) = \frac{P(Y \mid X)P(X)}{P(Y)}$$

Using Bayes' Rule to solve the Monty Hall problem

You pick door A.
Data *D* = Monty opened door B

Hypothesis space *H*:
h_1 = Car is behind door A
h_2 = Car is behind door C
h_3 = Car is behind door B

What is P(h_1 | *D*)?
What is P(h_2 | *D*)?
What is P(h_3 | *D*)?

Prior probability:
P(h_1) = 1/3 P(h_2) =1/3 P(h_3) =1/3
Likelihood:
P(D | h_1) = 1/2
P(D | h_2) = 1
P(D | h_3) = 0

P(D) = p(D|h_1)p(h_1) + p(D|h_2)p(h_2) +
p(D|h_3)p(h_3) = 1/6 + 1/3 + 0 = 1/2

By Bayes rule:
P(h1|D) = P(D|h1)p(h1) / P(D) = ½ 1/3 / ½ = 1/3

P(h2|D) = P(D|h2)p(h2) / P(D) = 1 1/3 / ½ = 2/3

So you should switch!

Independence and Conditional Independence

- Two random variables, X and Y, are independent if

$$P(X,Y) = P(X)P(Y)$$

- Two random variables, X and Y, are independent *given* Z if

$$P(X,Y \mid C) = P(X \mid C)P(Y \mid C)$$

- Examples?

Example: Spam classification

- Consider a spam-classification task, in which the documents are messages described by the following binary term attributes:

- **viagra**: true if the document contains the term "viagra"
- **million :** true if the document contains the term "million"
- **dollars:** true if the document contains the term "dollars"

28

Training data for spam classification

Document	**viagra**	**million**	**dollars**	**class**
msg 1	t	f	f	+
msg 2	f	f	f	+
msg 3	t	t	f	–
msg 4	t	f	t	+
msg 5	f	f	t	–

29

Naive Bayes Classifier

Let *classes* = {+, −}.

Assume we have a vocabulary V of terms t_i with |V| = n.

Let **x** = <a_1, a_2, ..., a_n>, where a_i=*1* if term t_i appears in document **x**, a_i= 0 otherwise.

We want to find the most probable class value, $class_{\text{MAP}}$, given the data **x**:

$$class_{MAP} = \underset{class \in \{+,-\}}{\text{argmax}}\ P(class \mid \mathbf{x})$$

$$= \underset{class \in \{+,-\}}{\text{argmax}}\ P(class \mid a_1, a_2, ..., a_n)$$

By Bayes Theorem:

$$class_{MAP} = \underset{class \in \{+,-\}}{\text{argmax}} \frac{P(a_1, a_2, ..., a_n \mid class) P(class)}{P(a_1, a_2, ..., a_n)}$$

$$= \underset{class \in \{+,-\}}{\text{argmax}} P(a_1, a_2, ..., a_n \mid class) P(class)$$

$P(class)$ can be estimated from the training data. How?

However, in general, not practical to use training data to estimate $P(a_1, a_2, ..., a_n \mid class)$. Why not?

- Naive Bayes classifier: Assume

$$P(a_1, a_2, ..., a_n \mid class) = P(a_1 \mid class) P(a_2 \mid class) \cdots P(a_n \mid class)$$

Is this a good assumption?

Given this assumption, here's how to classify an instance $\boldsymbol{x} = \langle a_1, a_2, ..., a_n \rangle$:

Naive Bayes classifier:

$$class_{MAP} = \underset{class \in \{+,-\}}{\text{argmax}} P(class) \prod_i P(a_i \mid class)$$

Estimate the values of these various probabilities over the training set.

Example: Spam classification

Document	**viagra**	**million**	**dollars**	**class**
msg 1	t	f	f	+
msg 2	f	f	f	+
msg 3	t	t	f	–
msg 4	t	f	t	+
msg 5	f	f	t	–

33

- Problem: "floating point underflow"
- Solution: Use logarithms of probabilities

$$class_{MAP} = \underset{class \in \{+,-\}}{\operatorname{argmax}} \log P(class) + \sum_i \log P(a_i \mid class)$$

34

Example: Spam classification

35

- Problem: some $P(a_i|class)$ are zero.
- Solution: "Laplace smoothing": add one to each count.
- Example

36

Bernoulli model: $P(t|c)$ = number of documents of class *c* that contain term *t* / total number of documents of class *c*

Multinomial model: $P(t|c)$ = number of "term positions" in which term *t* appears in documents of class c / total number of "term positions"

37

Homework

38