Introduction to Machine Learning

Probabilistic Modeling

Sources of uncertainty

- Incompleteness
- Incorrectness

Probability

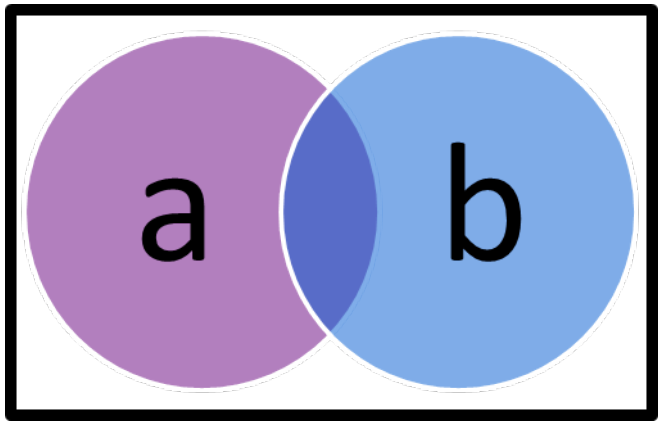
$$\text{Probability(event)} = P(\text{event}) = \frac{\#\text{instances of the event}}{\text{total \#instances}}$$

Sources of probabilities

- Frequency
- Consider the probability that the sun will still exist tomorrow.

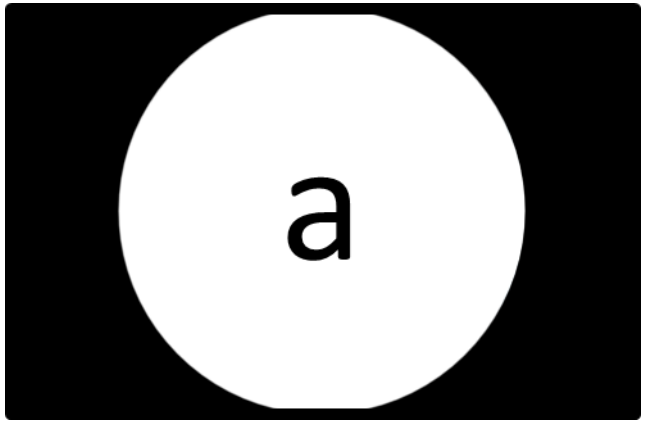
Axioms of probability

- $0 \leq P(\text{Event}) \leq 1$
- Disjunction, P(a or b) = P(a) + P(b) – P(a and b)



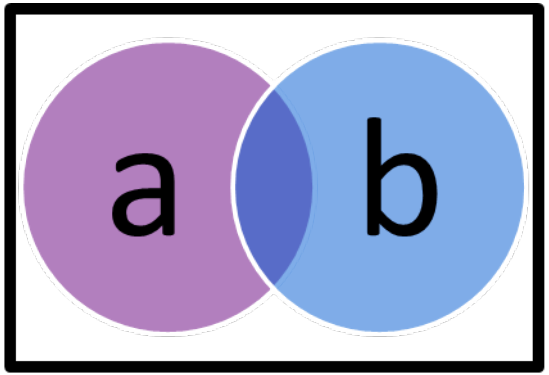
Negation

- P(not a) = 1 – P(a)



Conditional probability and conjunction

- P(a|b) = P(a and b) / P(b)



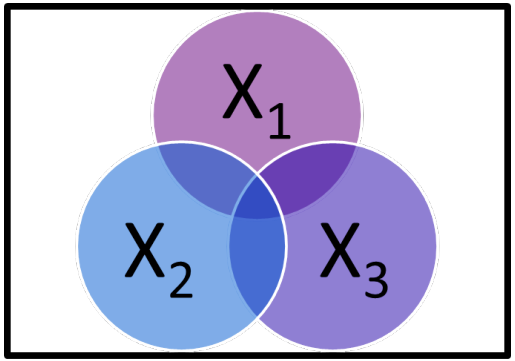
Conditional probability and conjunction

- $P(a \text{ and } b) = P(a) \times P(b|a)$
- $P(a \text{ and } b) = P(b) \times P(a|b)$

- If a and b are independent events
  - $P(a \text{ and } b) = P(a) \times P(b)$

More than 2 variables

- Chain rule



Bayes' rule

- Given a hypothesis (H) and evidence (E), what is P(H|E)?

Naïve Bayes Classifier (NBC)

- Predict class label with highest probability

Data for spam filtering

- date
- time
- recipient path
- IP number
- sender
- encoding
- many more features

Delivered-To: alex.smola@gmail.com
Received: by 10.216.47.73 with SMTP id s51cs361171web;
Tue, 3 Jan 2012 14:17:53 -0800 (PST)
Received: by 10.213.17.145 with SMTP id s17mr2519891eba.147.1325629071725;
Tue, 03 Jan 2012 14:17:51 -0800 (PST)
Return-Path: <alex+caf_=alex.smola=gmail.com@smola.org>
Received: from mail-ey0-f175.google.com (mail-ey0-f175.google.com [209.85.215.175])
by mx.google.com with ESMTPS id n4si29264232eef.57.2012.01.03.14.17.51
(version=TLSv1/SSLv3 cipher=OTHER);
Tue, 03 Jan 2012 14:17:51 -0800 (PST)
Received-SPF: neutral (google.com: 209.85.215.175 is neither permitted nor denied by best guess record for domain of alex+caf_=alex.smola=gmail.com@smola.org) client-ip=209.85.215.175;
Authentication-Results: mx.google.com; spf=neutral (google.com: 209.85.215.175 is neither permitted nor denied by best guess record for domain of alex+caf_=alex.smola=gmail.com@smola.org)
smtp.mail=alex+caf_=alex.smola=gmail.com@smola.org; dkim=pass (test mode) header.i=@googlemail.com
Received: by eaal1 with SMTP id l1so15092746eaa.6
for <alex.smola@gmail.com>; Tue, 03 Jan 2012 14:17:51 -0800 (PST)
Received: by 10.205.135.18 with SMTP id ie18mr5325064bkc.72.1325629071362;
Tue, 03 Jan 2012 14:17:51 -0800 (PST)
X-Forwarded-To: alex.smola@gmail.com
X-Forwarded-For: alex@smola.org alex.smola@gmail.com
Delivered-To: alex@smola.org
Received: by 10.204.65.198 with SMTP id k6cs206093bki;
Tue, 3 Jan 2012 14:17:50 -0800 (PST)
Received: by 10.52.88.179 with SMTP id bh19mr10729402vdb.38.1325629068795;
Tue, 03 Jan 2012 14:17:48 -0800 (PST)
Return-Path: <althoff.tim@googlemail.com>
Received: from mail-vx0-f179.google.com (mail-vx0-f179.google.com [209.85.220.179])
by mx.google.com with ESMTPS id dt4si11767074vdb.93.2012.01.03.14.17.48
(version=TLSv1/SSLv3 cipher=OTHER);
Tue, 03 Jan 2012 14:17:48 -0800 (PST)
Received-SPF: pass (google.com: domain of althoff.tim@googlemail.com designates 209.85.220.179 as permitted sender) client-ip=209.85.220.179;
Received: by vcbf13 with SMTP id f13so11295098vcb.10
for <alex@smola.org>; Tue, 03 Jan 2012 14:17:48 -0800 (PST)
DKIM-Signature: v=1; a=rsa-sha256; c=relaxed/relaxed;
d=googlemail.com; s=gamma;
h=mime-version:sender:date:x-google-sender-auth:message-id:subject
:from:to:content-type;
bh=WCbdZ5sXac25dpH02XcRyDOdts993hKwsAVXpGrFh0w=;
b=WK2B2+ExWnf/gvTkw6uUvKuP4XeoKnlJq3USYTm0RARK8dSFjyOQsIHeAP9Yssxp6O
7ngGoTzYqd+ZsyJfvQcLAWp1PCJhG8AMcnqWkx0NMeoFvIp2HQooZwxSOCx5ZRgY+7qX
ulbbdna4IUDXj6UFe16SpLDCkptd8OZ3gr7+o=
MIME-Version: 1.0
Received: by 10.220.108.81 with SMTP id e17mr24104004vcp.67.1325629067787;
Tue, 03 Jan 2012 14:17:47 -0800 (PST)
Sender: althoff.tim@googlemail.com
Received: by 10.220.17.129 with HTTP; Tue, 3 Jan 2012 14:17:47 -0800 (PST)
Date: Tue, 3 Jan 2012 14:17:47 -0800
X-Google-Sender-Auth: 6bwi6D17HjZIkxOEol38NZzyeHs
Message-ID: <CAFJJHDGPBW+SdZg0MdAABiAKydDk9tpeMoDijYGjoGO-WC7osg@mail.gmail.com>
Subject: CS 281B. Advanced Topics in Learning and Decision Making
From: Tim Althoff <althoff@eecs.berkeley.edu>
To: alex@smola.org
Content-Type: multipart/alternative; boundary=f46d043c7af4b07e8d04b5a7113a

Naïve Bayes Classifier (NBC)

- Why is the chain rule a bad idea here?

- Chain rule, need to estimate $k2^D-1$, $O(2^D)$, parameters

Naïve Bayes assumption

- The features are independent given the class label

- Can we use this to simplify the Bayes classifier?

- Naïve Bayes assumption, need to estimate k2d, O(kd), parameters

Naïve Bayes derivation

- Bayes rule, $P(b \mid a) = \frac{P(a \mid b)P(b)}{P(a)}$
- Apply to classification, $P(y_C|x_1,..,x_D) = \frac{P(x_1,..,xD|y_c)P(yC)}{P(x_1,..,xD)}$
- Output label with greatest probability, $argmax_{yc \in Y} \frac{P(x_1,..,xD|y_c)P(yC)}{P(x_1,..,xD)}$
- Remove denominator (why?), $argmax_{yc \in Y} P(x_1,..,xD|y_c)P(yC)$
- Apply naïve Bayes assumption, $argmax_{y_c \in Y} P(y_c) \prod_{i=1...D} P(x_i|y_c)$

Using NBC

- Training
  - For each target value (class value) y_c estimate y_c and $P(x|y_c)$
- For each attribute value x_i of each attribute x estimate $P(x_i|y_c)$
- Classify new instance

$$y_{NBC} = argmax_{y_c \in Y} P(y_c) \prod_{i=1 \ldots D} P(x_i|y_c)$$

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

NBC subtleties

- Conditional independence is often violated
- Some attribute values may not appear
- Products can get very small
- Continuous data

Gaussian naïve Bayes

- Assume continuous-valued feature x_i follows Gaussian distribution
- Compute mean and variance of x_i for each class
- Compute probability attribute has value v given mean and variance
- $P(x_i = v|y_C) = \frac{1}{\sqrt{2\pi\sigma_C^2}} \mathrm{e}^{\frac{(v-\mu_C)^2}{2\sigma_C^2}}$

Smoothing

- Example
  - Outlook=Sunny, Temperature=Cool, Humidity=High, Wind=Strong
- NBC(Example) = ?
- Laplace smoothing

Let's try this out

Now let's use it for text mining

comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

Extracting features from t

It was the best of times
It was the worst of times
It was the age of wisdom
It was the age of foolishness

Stop words

- a
- an
- the
- of
- on
- with
- from
- at
- to

- Different for each language

Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

Term frequency-inverse document frequency

- How important is a word to a particular document?
- Term frequency
- Inverse document frequency
- TF-IDF score

Let's try this out