Probability, Conditional Probability & Bayes Rule
A FAST REVIEW OF DISCRETE PROBABILITY (PART 2)
Discrete random variables

- A *random variable* can take on one of a set of different values, each with an associated probability. Its value at a particular time is *subject to random variation*.
  - *Discrete* random variables take on one of a discrete (often finite) range of values
  - Domain values must be *exhaustive* and *mutually exclusive*

- For us, random variables will have a discrete, countable (usually finite) domain of *arbitrary values*.
  - Mathematical statistics usually calls these *random elements*
  - Example: Weather is a discrete random variable with domain \{sunny, rain, cloudy, snow\}.
  - *Example:* A *Boolean random variable* has the domain \{true,false\},
Probability Distribution

- *Probability distribution* gives values for all possible assignments:
  - Vector notation: Weather is one of \(<0.72, 0.1, 0.08, 0.1>\), where weather is one of \(<\text{sunny}, \text{rain}, \text{cloudy}, \text{snow}>\).
  - \(P(\text{Weather}) = <0.72,0.1,0.08,0.1>\)
  - Sums to 1 over the domain

—*Practical advice: Easy to check*
—*Practical advice: Important to check*
Factored Representations: Propositions

- **Elementary proposition** constructed by assignment of a value to a random variable:
  
  - e.g. *Weather = sunny* (abbreviated as *sunny*)
  - e.g. *Cavity = false* (abbreviated as *¬cavity*)

- **Complex proposition** formed from elementary propositions & standard logical connectives
  
  - e.g. *Weather = sunny ∨ Cavity = false*

- *We will work with event spaces over such propositions*
A word on notation

Assume *Weather* is a discrete random variable with domain \{sunny, rain, cloudy, snow\}.

- *Weather* = sunny abbreviated sunny
- \( P(\text{Weather}=\text{sunny})=0.72 \) abbreviated \( P(\text{sunny})=0.72 \)

- *Cavity* = true abbreviated cavity
- *Cavity* = false abbreviated \( \neg \text{cavity} \)

Vector notation:
- Fix order of domain elements:
  \[
  \langle \text{sunny}, \text{rain}, \text{cloudy}, \text{snow} \rangle
  \]
- Specify the probability mass function (pmf) by a vector:
  \[
  P(\text{Weather}) = \langle 0.72, 0.1, 0.08, 0.1 \rangle
  \]
Joint probability distribution

- Probability assignment to all combinations of values of random variables (i.e. all elementary events)

<table>
<thead>
<tr>
<th></th>
<th>toothache</th>
<th>(\neg) toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>cavity</td>
<td>0.04</td>
<td>0.06</td>
</tr>
<tr>
<td>(\neg) cavity</td>
<td>0.01</td>
<td>0.89</td>
</tr>
</tbody>
</table>

- The sum of the entries in this table has to be 1
- **Every question about a domain can be answered by the joint distribution**

- Probability of a proposition is the sum of the probabilities of elementary events in which it holds
  - \(P(\text{cavity}) = 0.1\) [marginal of row 1]
  - \(P(\text{toothache}) = 0.05\) [marginal of toothache column]
Conditional Probability

- \( P(\text{cavity}) = 0.1 \) and \( P(\text{cavity} \land \text{toothache}) = 0.04 \) are both prior (unconditional) probabilities.

- Once the agent has new evidence concerning a previously unknown random variable, e.g., Toothache, we can specify a posterior (conditional) probability e.g. \( P(\text{cavity} | \text{Toothache=true}) \)

\[
P(a | b) = \frac{P(a \land b)}{P(b)}
\]

[Probability of a with the Universe \( \Omega \) restricted to b]

→ The new information restricts the set of possible worlds \( \omega_i \) consistent with it, so changes the probability.

- So \( P(\text{cavity} | \text{toothache}) = 0.04/0.05 = 0.8 \)

<table>
<thead>
<tr>
<th></th>
<th>toothache</th>
<th>( \neg ) toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>cavity</td>
<td>0.04</td>
<td>0.06</td>
</tr>
<tr>
<td>( \neg ) cavity</td>
<td>0.01</td>
<td>0.89</td>
</tr>
</tbody>
</table>
Conditional Probability (continued)

- **Definition of Conditional Probability:**
  \[ P(a \mid b) = \frac{P(a \land b)}{P(b)} \]

- **Product rule gives an alternative formulation:**
  \[
  P(a \land b) = P(a \mid b) \ast P(b) \\
  = P(b \mid a) \ast P(a)
  \]

- **A general version holds for whole distributions:**
  \[ P(\text{Weather, Cavity}) = P(\text{Weather} \mid \text{Cavity}) \ast P(\text{Cavity}) \]

- **Chain rule** is derived by successive application of product rule:
  \[
  P(A,B,C,D,E) = P(A \mid B,C,D,E) \ast P(B,C,D,E) \\
  = P(A \mid B,C,D,E) \ast P(B \mid C,D,E) \ast P(C,D,E) \\
  = \ldots \\
  = P(A \mid B,C,D,E) \ast P(B \mid C,D,E) \ast P(C \mid D,E) \ast P(D \mid E) \ast P(E)
  \]
Probabilistic Inference

- **Probabilistic inference**: the computation
  - from *observed evidence*
  - of *posterior probabilities*
  - for *query propositions*.
- We use the **full joint distribution** as the “knowledge base” from which answers to questions may be derived.
- Ex: three Boolean variables *Toothache* \( (T) \), *Cavity* \( (C) \), *ShowsOnXRay* \( (X) \)

<table>
<thead>
<tr>
<th></th>
<th>( t )</th>
<th>( \neg t )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X )</td>
<td>0.108</td>
<td>0.012</td>
</tr>
<tr>
<td>( \neg X )</td>
<td>0.144</td>
<td>0.576</td>
</tr>
</tbody>
</table>

- Probabilities in joint distribution sum to 1
Probabilistic Inference II

- Probability of any proposition computed by finding atomic events where proposition is true and adding their probabilities

- \( P(\text{cavity} \lor \text{toothache}) \)
  \[= 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 \]
  \[= 0.28 \]

- \( P(\text{cavity}) \)
  \[= 0.108 + 0.012 + 0.072 + 0.008 \]
  \[= 0.2 \]

- \( P(\text{cavity}) \) is called a **marginal probability** and the process of computing this is called **marginalization**
Probabilistic Inference III

- Can also compute conditional probabilities.

- \[ P(\neg \text{cavity} \mid \text{toothache}) = \frac{P(\neg \text{cavity} \land \text{toothache})}{P(\text{toothache})} \]
  \[ = \frac{(0.016 + 0.064)}{(0.108 + 0.012 + 0.016 + 0.064)} \]
  \[ = 0.4 \]

- Denominator is viewed as a normalization constant:
  - Stays constant no matter what the value of Cavity is.
  (Book uses \( \alpha \) to denote normalization constant \( 1/P(X) \), for random variable X.)
Bayes Rule & Naïve Bayes

(some slides adapted from slides by Massimo Poesio, adapted from slides by Chris Manning)
Bayes’ Rule & Diagnosis

\[ P(a|b) = \frac{P(b|a) \cdot P(a)}{P(b)} \]

- **Useful for assessing diagnostic probability from causal probability:**

\[ P(\text{Cause}|\text{Effect}) = \frac{P(\text{Effect}|\text{Cause}) \cdot P(\text{Cause})}{P(\text{Effect})} \]
Bayes’ Rule For Diagnosis II

\[ P(Disease \mid Symptom) = \frac{P(Symptom \mid Disease) \ast P(Disease)}{P(Symptom)} \]

Imagine:

- disease = TB, symptom = coughing
- \( P(disease \mid symptom) \) is different in TB-indicated country vs. USA
- \( P(symptom \mid disease) \) should be the same
  - It is more widely useful to learn \( P(symptom \mid disease) \)

- What about \( P(symptom) \)?
  - Use \textit{conditioning} (next slide)
  - For determining, e.g., the \textit{most likely} disease given the symptom, we can just ignore \( P(symptom) \)!!! (see slide 35)
Conditioning

- **Idea**: Use *conditional probabilities* instead of joint probabilities

- \[ P(a) = P(a \land b) + P(a \land \neg b) = P(a \mid b) \cdot P(b) + P(a \mid \neg b) \cdot P(\neg b) \]

  Here:

  \[ P(\text{symptom}) = P(\text{symptom} \mid \text{disease}) \cdot P(\text{disease}) + P(\text{symptom} \mid \neg \text{disease}) \cdot P(\neg \text{disease}) \]

- More generally: \( P(Y) = \sum_z P(Y \mid z) \cdot P(z) \)

- Marginalization and conditioning are useful rules for derivations involving probability expressions.
Exponentials rear their ugly head again...

- Estimating the necessary joint probability distribution for many symptoms is infeasible
  - For $|D|$ diseases, $|S|$ symptoms where a person can have $n$ of the diseases and $m$ of the symptoms
    - $P(s|d_1, d_2, \ldots, d_n)$ requires $|S| |D|^n$ values
    - $P(s_1, s_2, \ldots, s_m)$ requires $|S|^m$ values

- These numbers get big fast
  - If $|S| = 1000$, $|D| = 100$, $n=4$, $m=7$
    - $P(s|d_1, \ldots d_n)$ requires $1000 \times 100^4 = 10^{11}$ values (-1)
    - $P(s_1 \ldots s_m)$ requires $1000^7 = 10^{21}$ values (-1)
The Solution: *Independence*

- Random variables A and B are *independent* iff
  - \( P(A \land B) = P(A) \ast P(B) \)
  - equivalently: \( P(A \mid B) = P(A) \) and \( P(B \mid A) = P(B) \)

- A and B are independent if knowing whether A occurred gives no information about B (and vice versa)

- Independence assumptions are *essential* for efficient probabilistic reasoning

\[
P(T, X, C, W) = P(T, X, C) \ast P(W)
\]

- 15 entries \((2^4-1)\) reduced to 8 \((2^3-1 + 2-1)\)
  For *n independent* biased coins, \(O(2^n)\) entries \(\rightarrow O(n)\)
Conditional Independence

- **BUT absolute** independence is rare
- Dentistry is a large field with hundreds of variables, none of which are independent. What to do?

- **A and B** are *conditionally independent given C* iff
  - \( P(A \mid B, C) = P(A \mid C) \)
  - \( P(B \mid A, C) = P(B \mid C) \)
  - \( P(A \land B \mid C) = P(A \mid C) \ast P(B \mid C) \)

- Toothache (T), Spot in Xray (X), Cavity (C)
  - None of these are independent of the other two
  - But **T and X are conditionally independent given C**
Conditional Independence II  WHY??

- If I have a cavity, the probability that the XRay shows a spot doesn’t depend on whether I have a toothache (and vice versa):
  \[ P(X|T,C) = P(X|C) \]

- From which follows:
  \[ P(T|X,C) = P(T|C) \quad \text{and} \quad P(T,X|C) = P(T|C) \ast P(X|C) \]

- By the chain rule, given conditional independence:
  \[
  P(T,X,C) = P(T|X,C) \ast P(X,C) = P(T|X,C) \ast P(X|C) \ast P(C) \\
  = P(T|C) \ast P(X|C) \ast P(C)
  \]

- \( P(\text{Toothache, Cavity, Xray}) \) has \( 2^3 - 1 = 7 \) independent entries

- Given conditional independence, chain rule yields
  \( 2 + 2 + 1 = 5 \) independent numbers
Conditional Independence III

- In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from \( \text{exponential} \) in \( n \) to \( \text{linear} \) in \( n \).

- Conditional independence is our most basic and robust form of knowledge about uncertain environments.
Another Example

- Battery is dead (B)
- Radio plays (R)
- Starter turns over (S)
- None of these propositions are independent of one another
- **BUT:** \textit{R and S are conditionally independent given B}
Naïve Bayes I

By Bayes Rule

$$P(C|T, X) = \frac{P(T, X|C)P(C)}{P(T, X)}$$

If $T$ and $X$ are \textit{conditionally independent given} $C$:

$$P(C|T, X) = \frac{P(T|C)P(X|C)P(C)}{P(T, X)}$$

This is a \textit{Naïve Bayes Model}:

\textit{All effects assumed conditionally independent given} Cause

C  
\textbf{Cause}  

T  
\textbf{Effect}_1  

X  
\textbf{Effect}_2
Bayes' Rule II

- More generally

\[ P(\text{Cause}, \text{Effect}_1, \ldots, \text{Effect}_n) = P(\text{Cause}) \prod_i P(\text{Effect}_i | \text{Cause}) \]

- Total number of parameters is \textit{linear} in \( n \)
An Early Robust Statistical NLP Application

• A Statistical Model For Etymology (Church ’85)

• Determining etymology is crucial for text-to-speech

<table>
<thead>
<tr>
<th>Italian</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>AldriGHetti</td>
<td>lauGH, siGH</td>
</tr>
<tr>
<td>IannuCCi</td>
<td>aCCept</td>
</tr>
<tr>
<td>ItaliAno</td>
<td>hAte</td>
</tr>
</tbody>
</table>
An Early Robust Statistical NLP Application

<table>
<thead>
<tr>
<th>Name</th>
<th>Accuracy</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Angeletti</td>
<td>100%</td>
<td>Italian</td>
</tr>
<tr>
<td>Iannucci</td>
<td>100%</td>
<td>Italian</td>
</tr>
<tr>
<td>Italiano</td>
<td>100%</td>
<td>Italian</td>
</tr>
<tr>
<td>Lombardino</td>
<td>58%</td>
<td>Italian</td>
</tr>
<tr>
<td>Asahara</td>
<td>100%</td>
<td>Japanese</td>
</tr>
<tr>
<td>Fujimaki</td>
<td>100%</td>
<td>Japanese</td>
</tr>
<tr>
<td>Umeda</td>
<td>96%</td>
<td>Japanese</td>
</tr>
<tr>
<td>Anagnostopoulos</td>
<td>100%</td>
<td>Greek</td>
</tr>
<tr>
<td>Demetriadiis</td>
<td>100%</td>
<td>Greek</td>
</tr>
<tr>
<td>Dukakis</td>
<td>99%</td>
<td>Russian</td>
</tr>
<tr>
<td>Annette</td>
<td>75%</td>
<td>French</td>
</tr>
<tr>
<td>Deneuve</td>
<td>54%</td>
<td>French</td>
</tr>
<tr>
<td>Baguenard</td>
<td>54%</td>
<td>Middle French</td>
</tr>
</tbody>
</table>

- A very simple statistical model (your next homework) solved the problem, despite a wild statistical assumption
Computing the Normalizing Constant $P(T,X)$

\[
P(c|T, X) + P(\neg c|T, X) = 1
\]

\[
\frac{P(T|c)P(X|c)P(c)}{P(T, X)} + \frac{P(T|\neg c)P(X|\neg c)P(\neg c)}{P(T, X)} = 1
\]

\[
P(T|c)P(X|c)P(c) + P(T|\neg c)P(X|\neg c)P(\neg c) = P(T, X)
\]
IF THERE’S TIME.....
BUILDING A SPAM FILTER USING NAÏVE BAYES
Spam or not Spam: that is the question.

From: """" <takworlId@hotmail.com>
Subject: real estate is the only way... gem oalvgkay

Anyone can buy real estate with no money down

Stop paying rent TODAY!

There is no need to spend hundreds or even thousands for similar courses

I am 22 years old and I have already purchased 6 properties using the methods outlined in this truly INCREDIBLE ebook.

Change your life NOW!

=================================================================

Click Below to order:
http://www.wholesaledaily.com/sales/nmd.htm
=================================================================
Categorization/Classification Problems

• Given:
  • A description of an instance, $x \in X$, where $X$ is the instance language or instance space.
    — *(Issue: how do we represent text documents?)*
  • A fixed set of categories:
    $C = \{c_1, c_2, \ldots, c_n\}$

• Determine:
  • The category of $x$: $c(x) \in C$, where $c(x)$ is a categorization function whose domain is $X$ and whose range is $C$.
    — *We want to automatically build categorization functions (“classifiers”).*
EXAMPLES OF TEXT CATEGORIZATION

- **Categories = SPAM?**
  - “spam” / “not spam”

- **Categories = TOPICS**
  - “finance” / “sports” / “asia”

- **Categories = OPINION**
  - “like” / “hate” / “neutral”

- **Categories = AUTHOR**
  - “Shakespeare” / “Marlowe” / “Ben Jonson”
  - The Federalist papers
A Graphical View of Text Classification

Text feature 1

Text feature 2

Graphics

Arch.

NLP

AI

Theory
Bayesian Methods for Text Classification

- Uses Bayes theorem to build a generative Naïve Bayes model that approximates how data is produced

\[ P(C \mid D) = \frac{P(D \mid C)P(C)}{P(D)} \]

Where C: Categories, D: Documents

- Uses prior probability of each category given no information about an item.

- Categorization produces a posterior probability distribution over the possible categories given a description of each document.
Maximum a posteriori (MAP) Hypothesis

- Goodbye to that nasty normalization constant!!

\[ c_{MAP} \equiv \arg\max_{c \in C} P(c \mid D) \]

\[ = \arg\max_{c \in C} \frac{P(D \mid c)P(c)}{P(D)} \]

\[ = \arg\max_{c \in C} P(D \mid c)P(c) \]

As \( P(D) \) is constant

No need to compute \( \alpha \), here \( P(D) \)!!!!
Maximum likelihood Hypothesis

If all hypotheses are a priori equally likely, we only need to consider the $P(D|c)$ term:

$$c_{ML} \equiv \arg\max_{c \in C} P(D | c)$$

Maximum Likelihood Estimate (“MLE”)
Naive Bayes Classifiers

Task: Classify a new instance \( D \) based on a tuple of attribute values \( D = \langle x_1, x_2, \ldots, x_n \rangle \) into one of the classes \( c_j \in C \)

\[
c_{MAP} = \arg\max_{c \in C} P(c \mid x_1, x_2, \ldots, x_n)
\]

\[
= \arg\max_{c \in C} \frac{P(x_1, x_2, \ldots, x_n \mid c)P(c)}{P(x_1, x_2, \ldots, x_n)}
\]

\[
= \arg\max_{c \in C} P(x_1, x_2, \ldots, x_n \mid c)P(c)
\]
Naïve Bayes Classifier: Assumption

- $P(c_j)$
  - Can be estimated from the frequency of classes in the training examples.

- $P(x_1, x_2, \ldots, x_n | c_j)$
  - Again, $O(|X| \cdot |C|)$ parameters to estimate full joint prob. distribution
  - As we saw, can only be estimated if a vast number of training examples was available.

Naïve Bayes Conditional Independence Assumption:

$$P(x_1, x_2, \ldots, x_n | c_j) = \prod_i P(x_i | c_j)$$
The Naïve Bayes Classifier

- Conditional Independence Assumption: features are independent of each other given the class:

\[ P(X_1, \ldots, X_5 \mid C) = P(X_1 \mid C) \cdot P(X_2 \mid C) \cdot \cdots \cdot P(X_5 \mid C) \]

- This model is appropriate for binary variables