Probability: Review

• The state of the world is described using *random variables*

• Probabilities are defined over *events*
  – Sets of world states characterized by propositions about random variables
  – E.g., $D_1$, $D_2$: rolls of two dice
    • $P(D_1 > 2)$
    • $P(D_1 + D_2 = 11)$
  – $W$ is the state of the weather
    • $P(W = \text{“rainy”} \lor W = \text{“sunny”})$
Kolmogorov’s axioms of probability

- For any propositions (events) A, B:
  - $0 \leq P(A) \leq 1$
  - $P(\text{True}) = 1$ and $P(\text{False}) = 0$
  - $P(A \lor B) = P(A) + P(B) - P(A \land B)$
Joint probability distributions

- A **joint distribution** is an assignment of probabilities to every possible *atomic event*

<table>
<thead>
<tr>
<th>P(Cavity, Toothache)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cavity = false ∧ Toothache = false</td>
<td>0.8</td>
</tr>
<tr>
<td>Cavity = false ∧ Toothache = true</td>
<td>0.1</td>
</tr>
<tr>
<td>Cavity = true ∧ Toothache = false</td>
<td>0.05</td>
</tr>
<tr>
<td>Cavity = true ∧ Toothache = true</td>
<td>0.05</td>
</tr>
</tbody>
</table>
## Marginal probability distributions

<table>
<thead>
<tr>
<th>P(Cavity, Toothache)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cavity = false ∧ Toothache = false</td>
<td>0.8</td>
</tr>
<tr>
<td>Cavity = false ∧ Toothache = true</td>
<td>0.1</td>
</tr>
<tr>
<td>Cavity = true ∧ Toothache = false</td>
<td>0.05</td>
</tr>
<tr>
<td>Cavity = true ∧ Toothache = true</td>
<td>0.05</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>P(Cavity)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cavity = false</td>
<td>?</td>
</tr>
<tr>
<td>Cavity = true</td>
<td>?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>P(Toothache)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Toothache = false</td>
<td>?</td>
</tr>
<tr>
<td>Toothache = true</td>
<td>?</td>
</tr>
</tbody>
</table>
Marginal probability distributions

• Given the joint distribution $P(X, Y)$, how do we find the marginal distribution $P(X)$?

$$P(X = x) = P((X = x \land Y = y_1) \lor \ldots \lor (X = x \land Y = y_n)) = P((x, y_1) \lor \ldots \lor (x, y_n))$$

• General rule: to find $P(X = x)$, sum the probabilities of all atomic events where $X = x$. 
Conditional probability
Conditional probability

- For any two events $A$ and $B$, $P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A, B)}{P(B)}$
Conditional distributions

- A conditional distribution is a distribution over the values of one variable given fixed values of other variables.

<table>
<thead>
<tr>
<th>P(Cavity, Toothache)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Cavity = false ∧ Toothache = false</td>
<td>0.8</td>
</tr>
<tr>
<td>Cavity = false ∧ Toothache = true</td>
<td>0.1</td>
</tr>
<tr>
<td>Cavity = true ∧ Toothache = false</td>
<td>0.05</td>
</tr>
<tr>
<td>Cavity = true ∧ Toothache = true</td>
<td>0.05</td>
</tr>
</tbody>
</table>

| P(Cavity | Toothache = true) |         |
|------------|-------------------|
| Cavity = false | 0.667             |
| Cavity = true  | 0.333             |

| P(Cavity|Toothache = false) |         |
|----------------------|---------|
| Cavity = false        | 0.941   |
| Cavity = true         | 0.059   |

| P(Toothache | Cavity = true) |         |
|----------------|----------------|
| Toothache= false | 0.5           |
| Toothache = true  | 0.5           |

| P(Toothache | Cavity = false) |         |
|---------------|-----------------|
| Toothache= false | 0.889         |
| Toothache = true  | 0.111          |
Normalization trick

- To get the whole conditional distribution $P(X | y)$ at once, select all entries in the joint distribution matching $Y = y$ and renormalize them to sum to one.

<table>
<thead>
<tr>
<th>$P(Cavity, Toothache)$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$Cavity = false \land Toothache = false$</td>
<td>0.8</td>
</tr>
<tr>
<td>$Cavity = false \land Toothache = true$</td>
<td>0.1</td>
</tr>
<tr>
<td>$Cavity = true \land Toothache = false$</td>
<td>0.05</td>
</tr>
<tr>
<td>$Cavity = true \land Toothache = true$</td>
<td>0.05</td>
</tr>
</tbody>
</table>

Select

<table>
<thead>
<tr>
<th>Toothache, Cavity = false</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$Toothache = false$</td>
<td>0.8</td>
</tr>
<tr>
<td>$Toothache = true$</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Renormalize

| $P(Toothache | Cavity = false)$ |  |
|---------------------|--|
| $Toothache = false$ | 0.889 |
| $Toothache = true$  | 0.111 |
Normalization trick

• To get the whole conditional distribution $P(X | y)$ at once, select all entries in the joint distribution matching $Y = y$ and renormalize them to sum to one.

• Why does it work?

$$\frac{P(x, y)}{\sum_{x'} P(x', y)} = \frac{P(x, y)}{P(y)}$$  
by marginalization
Product rule

- Definition of conditional probability: 
  \[ P(A \mid B) = \frac{P(A, B)}{P(B)} \]

- Sometimes we have the conditional probability and want to obtain the joint:
  \[ P(A, B) = P(A \mid B)P(B) = P(B \mid A)P(A) \]
Product rule

- Definition of conditional probability:
  \[ P(A \mid B) = \frac{P(A, B)}{P(B)} \]

- Sometimes we have the conditional probability and want to obtain the joint:
  \[ P(A, B) = P(A \mid B)P(B) = P(B \mid A)P(A) \]

- The chain rule:
  \[
P(A_1, \ldots, A_n) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1, A_2) \cdots P(A_n \mid A_1, \ldots, A_{n-1})
  = \prod_{i=1}^{n} P(A_i \mid A_1, \ldots, A_{i-1})
\]
Bayes Rule

• The product rule gives us two ways to factor a joint distribution:

\[ P(A, B) = P(A \mid B)P(B) = P(B \mid A)P(A) \]

• Therefore,

\[ P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \]

• Why is this useful?
  – Can get diagnostic probability, e.g., \( P(\text{cavity} \mid \text{toothache}) \) from causal probability, e.g., \( P(\text{toothache} \mid \text{cavity}) \)

\[ P(\text{Cause} \mid \text{Evidence}) = \frac{P(\text{Evidence} \mid \text{Cause})P(\text{Cause})}{P(\text{Evidence})} \]
  – Can update our beliefs based on evidence
  – Important tool for probabilistic inference
Independence

• Two events A and B are independent if and only if $P(A, B) = P(A) P(B)$
  – In other words, $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$
  – This is an important simplifying assumption for modeling, e.g., *Toothache* and *Weather* can be assumed to be independent

• Are two *mutually exclusive* events independent?
  – No, but for mutually exclusive events we have $P(A \lor B) = P(A) + P(B)$

• **Conditional independence**: A and B are *conditionally independent* given C iff $P(A, B \mid C) = P(A \mid C) P(B \mid C)$
Conditional independence: Example

- **Toothache**: boolean variable indicating whether the patient has a toothache
- **Cavity**: boolean variable indicating whether the patient has a cavity
- **Catch**: whether the dentist's probe catches in the cavity

- If the patient has a cavity, the probability that the probe catches in it doesn't depend on whether he/she has a toothache
  \[ P(\text{Catch} \mid \text{Toothache}, \text{Cavity}) = P(\text{Catch} \mid \text{Cavity}) \]
- Therefore, Catch is conditionally independent of Toothache given Cavity
- Likewise, Toothache is conditionally independent of Catch given Cavity
  \[ P(\text{Toothache} \mid \text{Catch}, \text{Cavity}) = P(\text{Toothache} \mid \text{Cavity}) \]
- Equivalent statement:
  \[ P(\text{Toothache}, \text{Catch} \mid \text{Cavity}) = P(\text{Toothache} \mid \text{Cavity}) P(\text{Catch} \mid \text{Cavity}) \]
Conditional independence: Example

- How many numbers do we need to represent the joint probability table \( P(\text{Toothache}, \text{Cavity}, \text{Catch}) \)?
  \[ 2^3 - 1 = 7 \text{ independent entries} \]
- Write out the joint distribution using chain rule:
  \[
P(\text{Toothache}, \text{Catch}, \text{Cavity}) \\
  = P(\text{Cavity}) \ P(\text{Catch} | \text{Cavity}) \ P(\text{Toothache} | \text{Catch, Cavity}) \\
  = P(\text{Cavity}) \ P(\text{Catch} | \text{Cavity}) \ P(\text{Toothache} | \text{Cavity})
\]
- How many numbers do we need to represent these distributions?
  \[ 1 + 2 + 2 = 5 \text{ independent numbers} \]
- In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in \( n \) to linear in \( n \)
Naïve Bayes model

• Suppose we have many different types of observations (symptoms, features) that we want to use to diagnose the underlying cause

• It is usually impractical to directly estimate or store the joint distribution \( P(Cause, Effect_1, \ldots, Effect_n) \).

• To simplify things, we can assume that the different effects are conditionally independent given the underlying cause

• Then we can estimate the joint distribution as
Naïve Bayes model

• Suppose we have many different types of observations (symptoms, features) that we want to use to diagnose the underlying cause.

• It is usually impractical to directly estimate or store the joint distribution $P(Cause, Effect_1, \ldots, Effect_n)$.

• To simplify things, we can assume that the different effects are conditionally independent given the underlying cause.

• Then we can estimate the joint distribution as

$$P(Cause, Effect_1, \ldots, Effect_n) = P(Cause) \prod_i P(Effect_i \mid Cause)$$

• This is usually not accurate, but very useful in practice.
Example: Naïve Bayes Spam Filter

- **Bayesian decision theory**: to minimize the probability of error, we should classify a message as spam if $P(\text{spam} \mid \text{message}) > P(\neg \text{spam} \mid \text{message})$
  - *Maximum a posteriori (MAP)* decision

---

**Dear Sir.**

First, I must solicit your confidence in this transaction, this is by virtue of its nature as being utterly confidential and top secret. ...

---

**Ok, I know this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.**

---

**TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.**

**99 MILLION EMAIL ADDRESSES FOR ONLY $99**
Example: Naïve Bayes Spam Filter

- **Bayesian decision theory**: to minimize the probability of error, we should classify a message as spam if
  \[ P(\text{spam} \mid \text{message}) > P(\neg\text{spam} \mid \text{message}) \]
  - **Maximum a posteriori (MAP)** decision

- Apply Bayes rule to the posterior:
  \[
P(\text{spam} \mid \text{message}) = \frac{P(\text{message} \mid \text{spam})P(\text{spam})}{P(\text{message})}
  \]
  \[
P(\neg\text{spam} \mid \text{message}) = \frac{P(\text{message} \mid \neg\text{spam})P(\neg\text{spam})}{P(\text{message})}
  \]

- Notice that \( P(\text{message}) \) is just a constant normalizing factor and doesn’t affect the decision

- Therefore, to classify the message, all we need is to find
  \[ P(\text{message} \mid \text{spam})P(\text{spam}) \] and \[ P(\text{message} \mid \neg\text{spam})P(\neg\text{spam}) \]
Example: Naïve Bayes Spam Filter

• We need to find $P(\text{message} \mid \text{spam}) P(\text{spam})$ and $P(\text{message} \mid \neg\text{spam}) P(\neg\text{spam})$

• The message is a sequence of words $(w_1, \ldots, w_n)$

• *Bag of words* representation
  – The order of the words in the message is not important
  – Each word is conditionally independent of the others given message class (spam or not spam)

---

Dear Sir.

First, I must solicit your confidence in this transaction, this is by virtue of its nature as being utterly confidential and top secret. ...

---

Ok, I know this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working fine being stuck in the corner, but when I plugged it in, hit the power nothing happened.

---

TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES FOR ONLY $99
Bag of words illustration

US Presidential Speeches Tag Cloud

http://chir.ag/projects/preztags/
Bag of words illustration

US Presidential Speeches Tag Cloud
http://chir.ag/projects/preztags/
Example: Naïve Bayes Spam Filter

- We need to find $P(\text{message} \mid \text{spam}) P(\text{spam})$ and $P(\text{message} \mid \neg \text{spam}) P(\neg \text{spam})$
- The message is a sequence of words $(w_1, \ldots, w_n)$
- **Bag of words** representation
  - The order of the words in the message is not important
  - Each word is conditionally independent of the others given message class (spam or not spam)

$$P(\text{message} \mid \text{spam}) = P(w_1,\ldots,w_n \mid \text{spam}) = \prod_{i=1}^{n} P(w_i \mid \text{spam})$$

- Our filter will classify the message as spam if

$$P(\text{spam}) \prod_{i=1}^{n} P(w_i \mid \text{spam}) > P(\neg \text{spam}) \prod_{i=1}^{n} P(w_i \mid \neg \text{spam})$$
Example: Naïve Bayes Spam Filter

\[
P(\text{spam} \mid w_1, \ldots, w_n) \propto P(\text{spam}) \prod_{i=1}^{n} P(w_i \mid \text{spam})
\]

- **posterior**
- **prior**
- **likelihood**
Parameter estimation

- In order to classify a message, we need to know the prior \(P(\text{spam})\) and the likelihoods \(P(\text{word} \mid \text{spam})\) and \(P(\text{word} \mid \neg \text{spam})\)
  - These are the parameters of the probabilistic model
  - How do we obtain the values of these parameters?

<table>
<thead>
<tr>
<th>prior</th>
<th>(P(\text{word} \mid \text{spam}))</th>
<th>(P(\text{word} \mid \neg \text{spam}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>spam: 0.33</td>
<td>the: 0.0156</td>
<td>the: 0.0210</td>
</tr>
<tr>
<td>(\neg\text{spam}: 0.67)</td>
<td>to: 0.0153</td>
<td>to: 0.0133</td>
</tr>
<tr>
<td></td>
<td>and: 0.0115</td>
<td>of: 0.0119</td>
</tr>
<tr>
<td></td>
<td>of: 0.0095</td>
<td>2002: 0.0110</td>
</tr>
<tr>
<td></td>
<td>you: 0.0093</td>
<td>with: 0.0108</td>
</tr>
<tr>
<td></td>
<td>a: 0.0086</td>
<td>from: 0.0107</td>
</tr>
<tr>
<td></td>
<td>with: 0.0080</td>
<td>and: 0.0105</td>
</tr>
<tr>
<td></td>
<td>from: 0.0075</td>
<td>a: 0.0100</td>
</tr>
<tr>
<td></td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Parameter estimation

- How do we obtain the prior $P(\text{spam})$ and the likelihoods $P(\text{word} \mid \text{spam})$ and $P(\text{word} \mid \neg\text{spam})$?
  - Empirically: use training data

$$P(\text{word} \mid \text{spam}) = \frac{\text{# of word occurrences in spam messages}}{\text{total # of words in spam messages}}$$

- This is the maximum likelihood (ML) estimate, or estimate that maximizes the likelihood of the training data:

$$\prod_{d=1}^{D} \prod_{i=1}^{n_d} P(w_{d,i} \mid \text{class}_{d,i})$$

$d$: index of training document, $i$: index of a word