Review: Probability

- Random variables, events
- Axioms of probability
- Atomic events
- Joint and marginal probability distributions
- Conditional probability distributions
- Product rule, chain rule
- Independence and conditional independence
Bayesian inference, Naïve Bayes model

\[ P(\text{I'M NEAR THE OCEAN} | \text{I PICKED UP A SEASHELL}) = \]
\[ \frac{P(\text{I PICKED UP A SEASHELL} | \text{I'M NEAR THE OCEAN}) P(\text{I'M NEAR THE OCEAN})}{P(\text{I PICKED UP A SEASHELL})} \]

Statistically speaking, if you pick up a seashell and don't hold it to your ear, you can probably hear the ocean.

http://xkcd.com/1236/
Bayes Rule

- The product rule gives us two ways to factor a joint probability:
  \[ P(A, B) = \]

- Therefore,
  \[ P(A | B) = \frac{P(B | A)P(A)}{P(B)} \]

- Why is this useful?
  - Can update our beliefs about A based on evidence B
    - \( P(A) \) is the prior and \( P(A|B) \) is the posterior
  - Key tool for probabilistic inference: can get diagnostic probability from causal probability
    - E.g., \( P(\text{Cavity} = \text{true} | \text{Toothache} = \text{true}) \) from \( P(\text{Toothache} = \text{true} | \text{Cavity} = \text{true}) \)
Bayes Rule example

- Marie is getting married tomorrow, at an outdoor ceremony in the desert. In recent years, it has rained only 5 days each year \((5/365 = 0.014)\). Unfortunately, the weatherman has predicted rain for tomorrow. When it actually rains, the weatherman correctly forecasts rain 90% of the time. When it doesn't rain, he incorrectly forecasts rain 10% of the time. What is the probability that it will rain on Marie's wedding?
Bayes Rule example

- Marie is getting married tomorrow, at an outdoor ceremony in the desert. In recent years, it has rained only 5 days each year (5/365 = 0.014). Unfortunately, the weatherman has predicted rain for tomorrow. When it actually rains, the weatherman correctly forecasts rain 90% of the time. When it doesn't rain, he incorrectly forecasts rain 10% of the time. What is the probability that it will rain on Marie's wedding?

\[
P(\text{rain} \mid \text{predict}) = \frac{P(\text{predict} \mid \text{rain})P(\text{rain})}{P(\text{predict})} = \frac{P(\text{predict} \mid \text{rain})P(\text{rain})}{P(\text{predict} \mid \text{rain})P(\text{rain}) + P(\text{predict} \mid \neg \text{rain})P(\neg \text{rain})}
\]
Law of total probability

\[ P(X = x) = \sum_{i=1}^{n} P(X = x, Y = y_i) \]

\[ = \sum_{i=1}^{n} P(X = x | Y = y_i) P(Y = y_i) \]
Bayes Rule example

• Marie is getting married tomorrow, at an outdoor ceremony in the desert. In recent years, it has rained only 5 days each year (5/365 = 0.014). Unfortunately, the weatherman has predicted rain for tomorrow. When it actually rains, the weatherman correctly forecasts rain 90% of the time. When it doesn't rain, he incorrectly forecasts rain 10% of the time. What is the probability that it will rain on Marie's wedding?

\[
P(\text{rain} \mid \text{predict}) = \frac{P(\text{predict} \mid \text{rain})P(\text{rain})}{P(\text{predict})}
\]

\[
= \frac{P(\text{predict} \mid \text{rain})P(\text{rain})}{P(\text{predict} \mid \text{rain})P(\text{rain}) + P(\text{predict} \mid \neg\text{rain})P(\neg\text{rain})}
\]

\[
= \frac{0.9 \times 0.014}{0.9 \times 0.014 + 0.1 \times 0.986} = \frac{0.0126}{0.0126 + 0.0986} = 0.111
\]
Bayes rule: Example

• 1% of women at age forty who participate in routine screening have breast cancer. 80% of women with breast cancer will get positive mammographies. 9.6% of women without breast cancer will also get positive mammographies. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer?

\[
P(\text{cancer} | \text{positive}) = \frac{P(\text{positive} | \text{cancer})P(\text{cancer})}{P(\text{positive})} = \frac{P(\text{positive} | \text{cancer})P(\text{cancer})}{P(\text{positive} | \text{cancer})P(\text{cancer}) + P(\text{positive} | \neg \text{cancer})P(\neg \text{Cancer})}
\]

\[
= \frac{0.8 \times 0.01}{0.8 \times 0.01 + 0.096 \times 0.99} = \frac{0.008}{0.008 + 0.095} = 0.0776
\]
DID THE SUN JUST EXPLODE?
(ITS NIGHT, SO WE`RE NOT SURE.)

This neutrino detector measures whether the sun has gone nova.

Then, it rolls two dice. If they both come up six, it lies to us. Otherwise, it tells the truth.

Let's try.
Detector! Has the sun gone nova?

Roll

Yes.

FREQUENTIST STATISTICIAN:

The probability of this result happening by chance is $\frac{1}{36} = 0.027$. Since $p < 0.05$, I conclude that the sun has exploded.

BAYESIAN STATISTICIAN:

Bet you $50 it hasn't.

\[
P(\text{nova} | \text{yes}) = \frac{P(\text{yes} | \text{nova})P(\text{nova})}{P(\text{yes})}
\]

\[
P(\neg\text{nova} | \text{yes}) = \frac{P(\text{yes} | \neg\text{nova})P(\neg\text{nova})}{P(\text{yes})}
\]

https://xkcd.com/1132/

See also: https://xkcd.com/882/
Probabilistic inference

• Suppose the agent has to make a decision about the value of an unobserved *query variable* $X$ given some observed *evidence variable(s)* $E = e$
  
  – Partially observable, stochastic, episodic environment
  
  – Examples: $X = \{\text{spam, not spam}\}$, $e = \text{email message}$
    $X = \{\text{zebra, giraffe, hippo}\}$, $e = \text{image features}$

---

**Example Email:**

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. …

**Spam Email:**

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.

---

**False Positive:**

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

99 MILLION EMAIL ADDRESSES FOR ONLY $99
Bayesian decision theory

• Let $x$ be the value predicted by the agent and $x^*$ be the true value of $X$.
• The agent has a loss function, which is 0 if $x = x^*$ and 1 otherwise.
• Expected loss for predicting $x$:

$$
\sum_{x^*} L(x, x^*)P(x^* | e)
$$

• What is the estimate of $X$ that minimizes the expected loss?
  – The one that has the greatest posterior probability $P(x|e)$
  – This is called the Maximum a Posteriori (MAP) decision
MAP decision

• Value \( x \) of \( X \) that has the highest posterior probability given the evidence \( E = e \):

\[
x^* = \arg \max_x P(X = x \mid E = e) = \frac{P(E = e \mid X = x)P(X = x)}{P(E = e)}
\]

\[
\propto \arg \max_x P(E = e \mid X = x)P(X = x)
\]

\[
P(x \mid e) \propto P(e \mid x)P(x)
\]

posterior likelihood prior

• Maximum likelihood (ML) decision:

\[
x^* = \arg \max_x P(e \mid x)
\]
Naïve Bayes model

• Suppose we have many different types of observations (symptoms, features) $E_1, \ldots, E_n$ that we want to use to obtain evidence about an underlying hypothesis $X$

• MAP decision:

$$P(X = x \mid E_1 = e_1, \ldots, E_n = e_n)$$

$$\propto P(X = x)P(E_1 = e_1, \ldots, E_n = e_n \mid X = x)$$

– If each feature $E_i$ can take on $k$ values, how many entries are in the (conditional) joint probability table $P(E_1, \ldots, E_n \mid X = x)$?
Naïve Bayes model

- Suppose we have many different types of observations (symptoms, features) $E_1, \ldots, E_n$ that we want to use to obtain evidence about an underlying hypothesis $X$

- MAP decision:

$$P(X = x \mid E_1 = e_1, \ldots, E_n = e_n) \propto P(X = x)P(E_1 = e_1, \ldots, E_n = e_n \mid X = x)$$

- We can make the simplifying assumption that the different features are conditionally independent given the hypothesis:

$$P(E_1 = e_1, \ldots, E_n = e_n \mid X = x) = \prod_{i=1}^{n} P(E_i = e_i \mid X = x)$$

  - If each feature can take on $k$ values, what is the complexity of storing the resulting distributions?
Naïve Bayes model

• Posterior:

\[ P(X = x \mid E_1 = e_1, \ldots, E_n = e_n) \]

• MAP decision:

\[ x^* = \arg \max_x P(x \mid e) \propto P(x) \prod_{i=1}^{n} P(e_i \mid x) \]

\[ \begin{array}{ccc}
\text{posterior} & \text{prior} & \text{likelihood}
\end{array} \]
Case study: Text document classification

- **MAP decision:** assign a document to the class with the highest posterior $P(\text{class} \mid \text{document})$

- **Example: spam classification**
  - Classify a message as spam if $P(\text{spam} \mid \text{message}) > P(\neg \text{spam} \mid \text{message})$

Dear Sir.

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

**X**

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

99 MILLION EMAIL ADDRESSES FOR ONLY $99

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.
Case study: Text document classification

- **MAP decision:** assign a document to the class with the highest posterior $P(\text{class} \mid \text{document})$

- We have $P(\text{class} \mid \text{document}) \propto P(\text{document} \mid \text{class})P(\text{class})$

- To enable classification, we need to be able to estimate the likelihoods $P(\text{document} \mid \text{class})$ for all classes and priors $P(\text{class})$
Naïve Bayes Representation

• Goal: estimate likelihoods $P(\text{document} \mid \text{class})$ and priors $P(\text{class})$
• Likelihood: *bag of words* representation
  – The document is a sequence of words $(w_1, \ldots, w_n)$
  – The order of the words in the document is not important
  – Each word is conditionally independent of the others given document class

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 before 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
Naïve Bayes Representation

- Goal: estimate likelihoods $P(\text{document} \mid \text{class})$ and priors $P(\text{class})$
- Likelihood: *bag of words* representation
  - The document is a sequence of words $(w_1, \ldots, w_n)$
  - The order of the words in the document is not important
  - Each word is conditionally independent of the others given document class

$$P(\text{document} \mid \text{class}) = P(w_1, \ldots, w_n \mid \text{class}) = \prod_{i=1}^{n} P(w_i \mid \text{class})$$
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/
Bag of words illustration

2007-01-23: State of the Union Address

1962-10-22: Soviet Missiles in Cuba

1941-12-08: Request for a Declaration of War

US Presidential Speeches Tag Cloud
http://chir.ag/projects/preztags/
Naïve Bayes Representation

- Goal: estimate likelihoods $P(\text{document} \mid \text{class})$ and $P(\text{class})$

- Likelihood: *bag of words* representation
  - The document is a sequence of words $(w_1, \ldots, w_n)$
  - The order of the words in the document is not important
  - Each word is conditionally independent of the others given document class

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

- Thus, the problem is reduced to estimating marginal likelihoods of individual words $P(w_i \mid \text{class})$
Parameter estimation

- Model parameters: feature likelihoods \( P(\text{word} \mid \text{class}) \) and priors \( P(\text{class}) \)
  - 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></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>( \neg \text{spam}: 0.67 )</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Parameter estimation

- Model parameters: feature likelihoods $P(\text{word} \mid \text{class})$ and priors $P(\text{class})$
  - How do we obtain the values of these parameters?
  - Need training set of labeled samples from both classes

$$P(\text{word} \mid \text{class}) = \frac{\text{# of occurrences of this word in docs from this class}}{\text{total # of words in docs from this class}}$$

- 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
Parameter estimation

- Parameter estimate:

  \[
  P(\text{word} | \text{class}) = \frac{\text{# of occurrences of this word in docs from this class}}{\text{total # of words in docs from this class}}
  \]

- Parameter smoothing: dealing with words that were never seen or seen too few times
  - Laplacian smoothing: pretend you have seen every vocabulary word one more time than you actually did

  \[
  P(\text{word} | \text{class}) = \frac{\text{# of occurrences of this word in docs from this class} + 1}{\text{total # of words in docs from this class} + V}
  \]

  \(V: \text{total number of unique words}\)
Summary: Naïve Bayes for Document Classification

- Naïve Bayes model: assign the document to the class with the highest posterior

\[ P(\text{class} \mid \text{document}) \propto P(\text{class}) \prod_{i=1}^{n} P(w_i \mid \text{class}) \]

- Model parameters:

<table>
<thead>
<tr>
<th>Prior</th>
<th>Likelihood of class 1</th>
<th>Likelihood of class K</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(\text{class}_1) )</td>
<td>( P(w_1 \mid \text{class}_1) )</td>
<td>( P(w_1 \mid \text{class}_K) )</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>( P(\text{class}_K) )</td>
<td>( P(w_n \mid \text{class}_1) )</td>
<td>( P(w_n \mid \text{class}_K) )</td>
</tr>
</tbody>
</table>
Learning and inference pipeline

**Learning**

- Training Samples
- Features
- Training
- Learned model

**Inference**

- Test Sample
- Features
- Prediction
Review: Bayesian decision making

- Suppose the agent has to make decisions about the value of an unobserved *query variable* $X$ based on the values of an observed *evidence variable* $E$.

- **Inference problem**: given some evidence $E = e$, what is $P(X | e)$?

- **Learning problem**: estimate the parameters of the probabilistic model $P(X | E)$ given a *training sample* $\{(x_1, e_1), \ldots, (x_n, e_n)\}$.