Probability and Bayes Theorem

Relevant Readings: Sections 6.1, 6.2, 6.9 in Mitchell

CS495 - Machine Learning, Fall 2009

Final project

- Start dreaming up possible applications of concept learning to create agents (like the checkers example) for your final project

Probability

- If A is an event, we will denote the probability that A occurs as $\Pr(A)$

Probability

- If A is an event, we will denote the probability that A occurs as $\Pr(A)$
  - Note: Mitchell uses the notation $P(A)$

Probability

- If A is an event, we will denote the probability that A occurs as $\Pr(A)$
  - Note: Mitchell uses the notation $P(A)$
  - All probabilities are in the range $[0, 1]$

Probability

- If A is an event, we will denote the probability that A occurs as $\Pr(A)$
  - Note: Mitchell uses the notation $P(A)$
  - All probabilities are in the range $[0, 1]$
  - If an event is impossible, it has probability 0

Probability

- If A is an event, we will denote the probability that A occurs as $\Pr(A)$
  - Note: Mitchell uses the notation $P(A)$
  - All probabilities are in the range $[0, 1]$
  - If an event is impossible, it has probability 0
  - If an event is certain, it has probability 1

Probability

- If A is an event, we will denote the probability that A occurs as $\Pr(A)$
  - Note: Mitchell uses the notation $P(A)$
  - All probabilities are in the range $[0, 1]$
  - If an event is impossible, it has probability 0
  - If an event is certain, it has probability 1
  - Examples

Probability

- If A is an event, we will denote the probability that A occurs as $\Pr(A)$
  - Note: Mitchell uses the notation $P(A)$
  - All probabilities are in the range $[0, 1]$
  - If an event is impossible, it has probability 0
  - If an event is certain, it has probability 1
  - Examples
- Conditional probability

Probability

- If A is an event, we will denote the probability that A occurs as $\Pr(A)$
  - Note: Mitchell uses the notation $P(A)$
  - All probabilities are in the range $[0, 1]$
  - If an event is impossible, it has probability 0
  - If an event is certain, it has probability 1
  - Examples
- Conditional probability
  - $\Pr(A \mid B)$ is read: "probability of A, given B"

Probability

- If A is an event, we will denote the probability that A occurs as $\Pr(A)$
  - Note: Mitchell uses the notation $P(A)$
  - All probabilities are in the range $[0, 1]$
  - If an event is impossible, it has probability 0
  - If an event is certain, it has probability 1
  - Examples
- Conditional probability
  - $\Pr(A \mid B)$ is read: "probability of A, given B"
  - If A and B are events, then the conditional probability $\Pr(A \mid B)$ is defined to be $\Pr(A \mid B) = \frac{\Pr(A \wedge B)}{\Pr(B)}$

Probability

- If A is an event, we will denote the probability that A occurs as $\Pr(A)$
  - Note: Mitchell uses the notation $P(A)$
  - All probabilities are in the range $[0, 1]$
  - If an event is impossible, it has probability 0
  - If an event is certain, it has probability 1
  - Examples
- Conditional probability
  - $\Pr(A \mid B)$ is read: "probability of A, given B"
  - If A and B are events, then the conditional probability $\Pr(A \mid B)$ is defined to be $\Pr(A \mid B) = \frac{\Pr(A \wedge B)}{\Pr(B)}$
  - Examples

Probability

- If A is an event, we will denote the probability that A occurs as $\Pr(A)$
  - Note: Mitchell uses the notation $P(A)$
  - All probabilities are in the range $[0, 1]$
  - If an event is impossible, it has probability 0
  - If an event is certain, it has probability 1
  - Examples
- Conditional probability
  - $\Pr(A \mid B)$ is read: "probability of A, given B"
  - If A and B are events, then the conditional probability $\Pr(A \mid B)$ is defined to be $\Pr(A \mid B) = \frac{\Pr(A \wedge B)}{\Pr(B)}$
  - Examples

Testing for a rare illness:

- Suppose we have tested somone for the illness

Testing for a rare illness:

- Suppose we have tested somone for the illness
  - Let I be the event that the person has illness

Testing for a rare illness:

- Suppose we have tested somone for the illness
  - Let I be the event that the person has illness
  - Let $\oplus$ be the event that the test came out positive

Testing for a rare illness:

- Suppose we have tested somone for the illness
  - Let I be the event that the person has illness
  - Let $\oplus$ be the event that the test came out positive
  - Let $\ominus$ be the event that the test came out negative

Testing for a rare illness:

- Suppose we have tested somone for the illness
  - Let I be the event that the person has illness
  - Let $\oplus$ be the event that the test came out positive
  - Let $\ominus$ be the event that the test came out negative
- We know our test is very accurate:

Testing for a rare illness:

- Suppose we have tested somone for the illness
  - Let I be the event that the person has illness
  - Let $\oplus$ be the event that the test came out positive
  - Let $\ominus$ be the event that the test came out negative
- We know our test is very accurate:
  - $\Pr(\oplus \mid I) = .98$

Testing for a rare illness:

- Suppose we have tested somone for the illness
  - Let I be the event that the person has illness
  - Let $\oplus$ be the event that the test came out positive
  - Let $\ominus$ be the event that the test came out negative
- We know our test is very accurate:
  - $\Pr(\oplus \mid I) = .98$
  - $\Pr(\ominus \mid I) = .02$

Testing for a rare illness:

- Suppose we have tested somone for the illness
  - Let I be the event that the person has illness
  - Let $\oplus$ be the event that the test came out positive
  - Let $\ominus$ be the event that the test came out negative
- We know our test is very accurate:
  - $\Pr(\oplus \mid I) = .98$
  - $\Pr(\ominus \mid I) = .02$
  - $\Pr(\oplus \mid \neg I) = .03$

Testing for a rare illness:

- Suppose we have tested somone for the illness
  - Let I be the event that the person has illness
  - Let $\oplus$ be the event that the test came out positive
  - Let $\ominus$ be the event that the test came out negative
- We know our test is very accurate:
  - $\Pr(\oplus \mid I) = .98$
  - $\Pr(\ominus \mid I) = .02$
  - $\Pr(\oplus \mid \neg I) = .03$
  - $\Pr(\ominus \mid \neg I) = .97$

Testing for a rare illness:

- Suppose we have tested somone for the illness
  - Let I be the event that the person has illness
  - Let $\oplus$ be the event that the test came out positive
  - Let $\ominus$ be the event that the test came out negative
- We know our test is very accurate:
  - $\Pr(\oplus \mid I) = .98$
  - $\Pr(\ominus \mid I) = .02$
  - $\Pr(\oplus \mid \neg I) = .03$
  - $\Pr(\ominus \mid \neg I) = .97$
- Imagine now that $\oplus$ happens; the test came out positive

Testing for a rare illness:

- Suppose we have tested somone for the illness
  - Let I be the event that the person has illness
  - Let $\oplus$ be the event that the test came out positive
  - Let $\ominus$ be the event that the test came out negative
- We know our test is very accurate:
  - $\Pr(\oplus \mid I) = .98$
  - $\Pr(\ominus \mid I) = .02$
  - $\Pr(\oplus \mid \neg I) = .03$
  - $\Pr(\ominus \mid \neg I) = .97$
- Imagine now that $\oplus$ happens; the test came out positive
  - Which is a more likely hypothesis: that the person has the illness or not?

Testing for a rare illness:

- Suppose we have tested somone for the illness
  - Let I be the event that the person has illness
  - Let $\oplus$ be the event that the test came out positive
  - Let $\ominus$ be the event that the test came out negative
- We know our test is very accurate:
  - $\Pr(\oplus \mid I) = .98$
  - $\Pr(\ominus \mid I) = .02$
  - $\Pr(\oplus \mid \neg I) = .03$
  - $\Pr(\ominus \mid \neg I) = .97$
- Imagine now that $\oplus$ happens; the test came out positive
  - Which is a more likely hypothesis: that the person has the illness or not?
  - What if I told you that $\Pr(I) = .008$ and $\Pr(\neg I) = .992$ (i.e. the illness is very rare)? Does that change your mind?

Testing for a rare illness:

- Suppose we have tested somone for the illness
  - Let I be the event that the person has illness
  - Let $\oplus$ be the event that the test came out positive
  - Let $\ominus$ be the event that the test came out negative
- We know our test is very accurate:
  - $\Pr(\oplus \mid I) = .98$
  - $\Pr(\ominus \mid I) = .02$
  - $\Pr(\oplus \mid \neg I) = .03$
  - $\Pr(\ominus \mid \neg I) = .97$
- Imagine now that $\oplus$ happens; the test came out positive
  - Which is a more likely hypothesis: that the person has the illness or not?
  - What if I told you that $\Pr(I) = .008$ and $\Pr(\neg I) = .992$ (i.e. the illness is very rare)? Does that change your mind?

Bayes Theorem

- Take any events A and B

Bayes Theorem

- Take any events A and B
- We know by conditional probability (product rule) that $\Pr(A \mid B)\Pr(B) = \Pr(A \wedge B) = \Pr(B \mid A)\Pr(A)$

Bayes Theorem

- Take any events A and B
- We know by conditional probability (product rule) that $\Pr(A \mid B) \Pr(B) = \Pr(A \wedge B) = \Pr(B \mid A) \Pr(A)$
- Divide both sides by $\Pr(B)$ (assume it's nonzero):

Bayes Theorem

- Take any events A and B
- We know by conditional probability (product rule) that $\Pr(A \mid B)\Pr(B) = \Pr(A \wedge B) = \Pr(B \mid A)\Pr(A)$
- Divide both sides by $\Pr(B)$ (assume it's nonzero):
  - $\Pr(A \mid B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$

Bayes Theorem

- Take any events A and B
- We know by conditional probability (product rule) that $\Pr(A \mid B)\Pr(B) = \Pr(A \wedge B) = \Pr(B \mid A)\Pr(A)$
- Divide both sides by $\Pr(B)$ (assume it's nonzero):
  - $\Pr(A \mid B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$
  - This equation is known as Bayes Theorem

Bayes Theorem

- Take any events A and B
- We know by conditional probability (product rule) that $\Pr(A \mid B)\Pr(B) = \Pr(A \wedge B) = \Pr(B \mid A)\Pr(A)$
- Divide both sides by $\Pr(B)$ (assume it's nonzero):
  - $\Pr(A \mid B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$
  - This equation is known as Bayes Theorem
- Apply Bayes Theorem to the illness question:

Bayes Theorem

- Take any events A and B
- We know by conditional probability (product rule) that $\Pr(A \mid B)\Pr(B) = \Pr(A \wedge B) = \Pr(B \mid A)\Pr(A)$
- Divide both sides by $\Pr(B)$ (assume it's nonzero):
  - $\Pr(A \mid B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$
  - This equation is known as Bayes Theorem
- Apply Bayes Theorem to the illness question:
  - $A = I$

Bayes Theorem

- Take any events A and B
- We know by conditional probability (product rule) that $\Pr(A \mid B)\Pr(B) = \Pr(A \wedge B) = \Pr(B \mid A)\Pr(A)$
- Divide both sides by $\Pr(B)$ (assume it's nonzero):
  - $\Pr(A \mid B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$
  - This equation is known as Bayes Theorem
- Apply Bayes Theorem to the illness question:
  - $A = I$
  - $B = \oplus$

Bayes Theorem

- Take any events A and B
- We know by conditional probability (product rule) that $\Pr(A \mid B)\Pr(B) = \Pr(A \wedge B) = \Pr(B \mid A)\Pr(A)$
- Divide both sides by $\Pr(B)$ (assume it's nonzero):
  - $\Pr(A \mid B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$
  - This equation is known as Bayes Theorem
- Apply Bayes Theorem to the illness question:
  - $A = I$
  - $B = \oplus$
  - $\Pr(I \mid \oplus) = \frac{\Pr(\oplus|I)\Pr(I)}{\Pr(\oplus)} = \frac{(.98)(.008)}{\Pr(\oplus)}$

Bayes Theorem

- Take any events A and B
- We know by conditional probability (product rule) that $\Pr(A \mid B)\Pr(B) = \Pr(A \wedge B) = \Pr(B \mid A)\Pr(A)$
- Divide both sides by $\Pr(B)$ (assume it's nonzero):
  - $\Pr(A \mid B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$
  - This equation is known as Bayes Theorem
- Apply Bayes Theorem to the illness question:
  - $A = I$
  - $B = \oplus$
  - $\Pr(I \mid \oplus) = \frac{\Pr(\oplus|I)\Pr(I)}{\Pr(\oplus)} = \frac{(.98)(.008)}{\Pr(\oplus)}$
  - $\Pr(\neg I \mid \oplus) = \frac{\Pr(\oplus|\neg I)\Pr(\neg I)}{\Pr(\oplus)} = \frac{(.03)(.992)}{\Pr(\oplus)}$

Bayes Theorem

- Take any events A and B
- We know by conditional probability (product rule) that $\Pr(A \mid B)\Pr(B) = \Pr(A \wedge B) = \Pr(B \mid A)\Pr(A)$
- Divide both sides by $\Pr(B)$ (assume it's nonzero):
  - $\Pr(A \mid B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}$
  - This equation is known as Bayes Theorem
- Apply Bayes Theorem to the illness question:
  - $A = I$
  - $B = \oplus$
  - $\Pr(I \mid \oplus) = \frac{\Pr(\oplus|I)\Pr(I)}{\Pr(\oplus)} = \frac{(.98)(.008)}{\Pr(\oplus)}$
  - $\Pr(\neg I \mid \oplus) = \frac{\Pr(\oplus|\neg I)\Pr(\neg I)}{\Pr(\oplus)} = \frac{(.03)(.992)}{\Pr(\oplus)}$
  - So the more likely hypothesis is that the person *doesn't* have the illness!

Bayes Theorem

- Take any events A and B
- We know by conditional probability (product rule) that $\Pr(A \mid B) \Pr(B) = \Pr(A \wedge B) = \Pr(B \mid A) \Pr(A)$
- Divide both sides by $\Pr(B)$ (assume it's nonzero):
  - $\Pr(A \mid B) = \frac{\Pr(B|A) \Pr(A)}{\Pr(B)}$
  - This equation is known as Bayes Theorem
- Apply Bayes Theorem to the illness question:
  - $A = I$
  - $B = \oplus$
  - $\Pr(I \mid \oplus) = \frac{\Pr(\oplus|I) \Pr(I)}{\Pr(\oplus)} = \frac{(.98)(.008)}{\Pr(\oplus)}$
  - $\Pr(\neg I \mid \oplus) = \frac{\Pr(\oplus|\neg I) \Pr(\neg I)}{\Pr(\oplus)} = \frac{(.03)(.992)}{\Pr(\oplus)}$
  - So the more likely hypothesis is that the person *doesn't* have the illness!

Naive Bayes

- Hey wait, maybe we can apply Bayes Theorem concept to machine learning!

Naive Bayes

- Hey wait, maybe we can apply Bayes Theorem concept to machine learning!
- One way to doing this is the called Naive Bayesian Classifier

Naive Bayes

- Hey wait, maybe we can apply Bayes Theorem concept to machine learning!
- One way to doing this is the called Naive Bayesian Classifier
  - It's simple

Naive Bayes

- Hey wait, maybe we can apply Bayes Theorem concept to machine learning!
- One way to doing this is the called Naive Bayesian Classifier
  - It's simple
  - It's computationally efficient

Naive Bayes

- Hey wait, maybe we can apply Bayes Theorem concept to machine learning!
- One way to doing this is the called Naive Bayesian Classifier
  - It's simple
  - It's computationally efficient
  - It can be remarkably effective (depending on the application)

Naive Bayes

- Hey wait, maybe we can apply Bayes Theorem concept to machine learning!
- One way to doing this is the called Naive Bayesian Classifier
  - It's simple
  - It's computationally efficient
  - It can be remarkably effective (depending on the application)

Naive Bayes

- Consider the following setting:

Naive Bayes

- Consider the following setting:
  - The task, T, is a concept learning task (boolean valued target function $f : D \rightarrow \{0, 1\}$)

Naive Bayes

- Consider the following setting:
  - The task, T, is a concept learning task (boolean valued target function $f : D \rightarrow \{0, 1\}$)
  - Training experience, E, is input/output examples of f

Naive Bayes

- Consider the following setting:
  - The task, T, is a concept learning task (boolean valued target function $f : D \rightarrow \{0, 1\}$)
  - Training experience, E, is input/output examples of f
  - Each piece of training data has attributes a_1 through a_n

Naive Bayes

- Consider the following setting:
  - The task, T, is a concept learning task (boolean valued target function $f : D \rightarrow \{0, 1\}$)
  - Training experience, E, is input/output examples of f
  - Each piece of training data has attributes a_1 through a_n
  - We want to find the most likely output, v_{MAP} , of f, given $(a_1, a_2, \ldots, a_n)$

Naive Bayes

- Consider the following setting:
  - The task, T, is a concept learning task (boolean valued target function $f : D \rightarrow \{0, 1\}$)
  - Training experience, E, is input/output examples of f
  - Each piece of training data has attributes a_1 through a_n
  - We want to find the most likely output, v_{MAP} , of f, given $(a_1, a_2, \ldots, a_n)$
    - Sidebar: *MAP* stands for Maximum A Posteriori

Naive Bayes

- Consider the following setting:
  - The task, T, is a concept learning task (boolean valued target function $f : D \rightarrow \{0, 1\}$)
  - Training experience, E, is input/output examples of f
  - Each piece of training data has attributes a_1 through a_n
  - We want to find the most likely output, v_{MAP} , of f, given $(a_1, a_2, \ldots, a_n)$
    - Sidebar: *MAP* stands for Maximum A Posteriori
- Then (where $B = \{0, 1\}$):

Naive Bayes

- Consider the following setting:
  - The task, T, is a concept learning task (boolean valued target function $f : D \rightarrow \{0, 1\}$)
  - Training experience, E, is input/output examples of f
  - Each piece of training data has attributes a_1 through a_n
  - We want to find the most likely output, v_{MAP} , of f, given $(a_1, a_2, \ldots, a_n)$
    - Sidebar: *MAP* stands for Maximum A Posteriori
- Then (where $B = \{0, 1\}$):
  - $v_{MAP} = \mathrm{argmax}_{b \in B} \Pr(b \mid a_1, a_2, \ldots, a_n)$

Naive Bayes

- Consider the following setting:
  - The task, T, is a concept learning task (boolean valued target function $f : D \rightarrow \{0, 1\}$)
  - Training experience, E, is input/output examples of f
  - Each piece of training data has attributes a_1 through a_n
  - We want to find the most likely output, v_{MAP} , of f, given $(a_1, a_2, \ldots, a_n)$
    - Sidebar: *MAP* stands for Maximum A Posteriori
- Then (where $B = \{0, 1\}$):
  - $v_{MAP} = \operatorname{argmax}_{b \in B} \Pr(b \mid a_1, a_2, \ldots, a_n)$
  - $v_{MAP} = \operatorname{argmax}_{b \in B} \frac{\Pr(a_1, a_2, \ldots, a_n | b) \Pr(b)}{\Pr(a_1, a_2, \ldots, a_n)}$ (by Bayes Theorem)

Naive Bayes

- Consider the following setting:
  - The task, T, is a concept learning task (boolean valued target function $f : D \rightarrow \{0,1\}$)
  - Training experience, E, is input/output examples of f
  - Each piece of training data has attributes a_1 through a_n
  - We want to find the most likely output, v_{MAP} , of f, given $(a_1, a_2, \ldots, a_n)$
    - Sidebar: *MAP* stands for Maximum A Posteriori
- Then (where $B = \{0,1\}$):
  - $v_{MAP} = \operatorname{argmax}_{b \in B} \Pr(b \mid a_1, a_2, \ldots, a_n)$
  - $v_{MAP} = \operatorname{argmax}_{b \in B} \frac{\Pr(a_1, a_2, \ldots, a_n \mid b) \Pr(b)}{\Pr(a_1, a_2, \ldots, a_n)}$ (by Bayes Theorem)
  - $\quad = \operatorname{argmax}_{b \in B} \Pr(a_1, a_2, \ldots, a_n \mid b) \Pr(b)$

Naive Bayes

- Consider the following setting:
  - The task, T, is a concept learning task (boolean valued target function $f : D \rightarrow \{0, 1\}$)
  - Training experience, E, is input/output examples of f
  - Each piece of training data has attributes a_1 through a_n
  - We want to find the most likely output, v_{MAP} , of f, given $(a_1, a_2, \ldots, a_n)$
    - Sidebar: *MAP* stands for Maximum A Posteriori
- Then (where $B = \{0, 1\}$):
  - $v_{MAP} = \operatorname{argmax}_{b \in B} \Pr(b \mid a_1, a_2, \ldots, a_n)$
  - $v_{MAP} = \operatorname{argmax}_{b \in B} \frac{\Pr(a_1, a_2, \ldots, a_n \mid b) \Pr(b)}{\Pr(a_1, a_2, \ldots, a_n)}$ (by Bayes Theorem)
  - $\quad = \operatorname{argmax}_{b \in B} \Pr(a_1, a_2, \ldots, a_n \mid b) \Pr(b)$
  - $\quad = \operatorname{argmax}_{b \in B} \Pr(b) \Pr(a_1 \mid b) \Pr(a_2 \mid b) \cdots \Pr(a_n \mid b)$

Naive Bayes

- Consider the following setting:
  - The task, T, is a concept learning task (boolean valued target function $f : D \rightarrow \{0, 1\}$)
  - Training experience, E, is input/output examples of f
  - Each piece of training data has attributes a_1 through a_n
  - We want to find the most likely output, v_{MAP} , of f, given $(a_1, a_2, \ldots, a_n)$
    - Sidebar: *MAP* stands for Maximum A Posteriori
- Then (where $B = \{0, 1\}$):
  - $v_{MAP} = \mathrm{argmax}_{b \in B} \Pr(b \mid a_1, a_2, \ldots, a_n)$
  - $v_{MAP} = \mathrm{argmax}_{b \in B} \frac{\Pr(a_1, a_2, \ldots, a_n \mid b) \Pr(b)}{\Pr(a_1, a_2, \ldots, a_n)}$ (by Bayes Theorem)
  - $= \mathrm{argmax}_{b \in B} \Pr(a_1, a_2, \ldots, a_n \mid b) \Pr(b)$
  - $= \mathrm{argmax}_{b \in B} \Pr(b) \Pr(a_1 \mid b) \Pr(a_2 \mid b) \cdots \Pr(a_n \mid b)$
    - By simplifying assumption that $\Pr(a_1, a_2, \ldots, a_n \mid b) = \Pr(a_1 \mid b) \Pr(a_2 \mid b) \cdots \Pr(a_n \mid b)$

Naive Bayes

- Consider the following setting:
  - The task, T, is a concept learning task (boolean valued target function $f : D \rightarrow \{0, 1\}$)
  - Training experience, E, is input/output examples of f
  - Each piece of training data has attributes a_1 through a_n
  - We want to find the most likely output, v_{MAP} , of f, given $(a_1, a_2, \ldots, a_n)$
    - Sidebar: *MAP* stands for Maximum A Posteriori
- Then (where $B = \{0, 1\}$):
  - $v_{MAP} = \mathrm{argmax}_{b \in B} \Pr(b \mid a_1, a_2, \ldots, a_n)$
  - $v_{MAP} = \mathrm{argmax}_{b \in B} \frac{\Pr(a_1, a_2, \ldots, a_n \mid b) \Pr(b)}{\Pr(a_1, a_2, \ldots, a_n)}$ (by Bayes Theorem)
  - $= \mathrm{argmax}_{b \in B} \Pr(a_1, a_2, \ldots, a_n \mid b) \Pr(b)$
  - $= \mathrm{argmax}_{b \in B} \Pr(b) \Pr(a_1 \mid b) \Pr(a_2 \mid b) \cdots \Pr(a_n \mid b)$
    - By simplifying assumption that $\Pr(a_1, a_2, \ldots, a_n \mid b) = \Pr(a_1 \mid b) \Pr(a_2 \mid b) \cdots \Pr(a_n \mid b)$
  - Sidebar: $\prod_i \Pr(a_i \mid b) = \Pr(a_1 \mid b) \Pr(a_2 \mid b) \cdots \Pr(a_n \mid b)$

Naive Bayes

- Consider the following setting:
  - The task, T, is a concept learning task (boolean valued target function $f : D \rightarrow \{0, 1\}$)
  - Training experience, E, is input/output examples of f
  - Each piece of training data has attributes a_1 through a_n
  - We want to find the most likely output, v_{MAP} , of f, given $(a_1, a_2, \ldots, a_n)$
    - Sidebar: *MAP* stands for Maximum A Posteriori
- Then (where $B = \{0, 1\}$):
  - $v_{MAP} = \mathrm{argmax}_{b \in B} \Pr(b \mid a_1, a_2, \ldots, a_n)$
  - $v_{MAP} = \mathrm{argmax}_{b \in B} \frac{\Pr(a_1, a_2, \ldots, a_n \mid b) \Pr(b)}{\Pr(a_1, a_2, \ldots, a_n)}$ (by Bayes Theorem)
  - $\quad = \mathrm{argmax}_{b \in B} \Pr(a_1, a_2, \ldots, a_n \mid b) \Pr(b)$
  - $\quad = \mathrm{argmax}_{b \in B} \Pr(b) \Pr(a_1 \mid b) \Pr(a_2 \mid b) \cdots \Pr(a_n \mid b)$
    - By simplifying assumption that $\Pr(a_1, a_2, \ldots, a_n \mid b) = \Pr(a_1 \mid b) \Pr(a_2 \mid b) \cdots \Pr(a_n \mid b)$
  - Sidebar: $\prod_i \Pr(a_i \mid b) = \Pr(a_1 \mid b) \Pr(a_2 \mid b) \cdots \Pr(a_n \mid b)$

Naive Bayes

- Bottom line: If we just estimate the $\Pr(b)$ and $\Pr(a_i \mid b)$ probabilities based on the training data, we have the most likely output (if our simplifying assumption holds, anyway)

Naive Bayes

- Bottom line: If we just estimate the $\Pr(b)$ and $\Pr(a_i \mid b)$ probabilities based on the training data, we have the most likely output (if our simplifying assumption holds, anyway)
- Won't zeros cause problems for this?

Naive Bayes

- Bottom line: If we just estimate the $\Pr(b)$ and $\Pr(a_i \mid b)$ probabilities based on the training data, we have the most likely output (if our simplifying assumption holds, anyway)
- Won't zeros cause problems for this?
- One way to handle that is to "invent" some faux training examples to start out with (called *m*-estimates)

Naive Bayes

- Bottom line: If we just estimate the $\Pr(b)$ and $\Pr(a_i \mid b)$ probabilities based on the training data, we have the most likely output (if our simplifying assumption holds, anyway)
- Won't zeros cause problems for this?
- One way to handle that is to "invent" some faux training examples to start out with (called m-estimates)

Some terminology

- Target concept: the function we are trying to learn

Some terminology

- Target concept: the function we are trying to learn
- Hypothesis: one possibility under consideraion for the target concept

Some terminology

- Target concept: the function we are trying to learn
- Hypothesis: one possibility under consideraion for the target concept
- Hypothesis space: set of all possibilities for the target concept

Some terminology

- Target concept: the function we are trying to learn
- Hypothesis: one possibility under consideraion for the target concept
- Hypothesis space: set of all possibilities for the target concept
- Conjunction: an “and” (in the logical sense)

Some terminology

- Target concept: the function we are trying to learn
- Hypothesis: one possibility under consideraion for the target concept
- Hypothesis space: set of all possibilities for the target concept
- Conjunction: an "and" (in the logical sense)
- Disjunction: an "or" (in the logical sense)

Some terminology

- Target concept: the function we are trying to learn
- Hypothesis: one possibility under consideraion for the target concept
- Hypothesis space: set of all possibilities for the target concept
- Conjunction: an "and" (in the logical sense)
- Disjunction: an "or" (in the logical sense)
- Instance: If $f : X \rightarrow \{0, 1\}$ is a target concept, X is the set of instances

Some terminology

- Target concept: the function we are trying to learn
- Hypothesis: one possibility under consideraion for the target concept
- Hypothesis space: set of all possibilities for the target concept
- Conjunction: an "and" (in the logical sense)
- Disjunction: an "or" (in the logical sense)
- Instance: If $f : X \to \{0, 1\}$ is a target concept, X is the set of instances
- A posteriori: based on experience (empirical)

Some terminology

- Target concept: the function we are trying to learn
- Hypothesis: one possibility under consideraion for the target concept
- Hypothesis space: set of all possibilities for the target concept
- Conjunction: an "and" (in the logical sense)
- Disjunction: an "or" (in the logical sense)
- Instance: If $f : X \to \{0, 1\}$ is a target concept, X is the set of instances
- A posteriori: based on experience (empirical)
- A priori: independent of experience (logical)

Some terminology

- Target concept: the function we are trying to learn
- Hypothesis: one possibility under consideraion for the target concept
- Hypothesis space: set of all possibilities for the target concept
- Conjunction: an “and” (in the logical sense)
- Disjunction: an “or” (in the logical sense)
- Instance: If $f : X \to \{0, 1\}$ is a target concept, X is the set of instances
- A posteriori: based on experience (empirical)
- A priori: independent of experience (logical)