

Machine Learning and Pattern Recognition

Ole-Christoffer Granmo

University of Agder

E-mail: ole.granmo@uia.no

August 30, 2007

Bio

Name: Ole-Christoffer Granmo

Born: Skien, Norway, 1974

Education: Cand. Scient., University of Oslo, Norway, 1999; Dr. Scient., University of Oslo, Norway, 2004

Research Interests: Intelligent Systems, Stochastic Modelling and Inference, Machine Learning, Pattern Recognition, Learning Automata, Distributed Computing, and Surveillance and Monitoring

First Computer: ZX Spectrum, 1984

- Zilog Z80A CPU running at 3.5 MHz with 48 KB RAM



Outline

- Overview of Machine Learning
 - Definition
 - Examples
 - Where is This Headed?
- Lecture Goals
 - Knowledge and Skills to be Obtained
 - Job Opportunities
- Pattern Recognition
 - What is a Pattern Recognition System?
 - Bayesian Inference
 - Probability Theory
 - Naive Bayesian Classifier
 - Text Classification

Part I:

Overview of Machine Learning

Machine Learning: Defining Question

- The field of Machine Learning seeks to answer the question:
“How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?”

Machine Learning Tasks

- The question covers a broad range of learning tasks, such as:
 - How to design autonomous mobile robots that learn to navigate from their own experience?
 - How to data mine historical medical records to learn which future patients will respond best to which treatments?
 - How to build search engines that automatically customize to their user's interests?

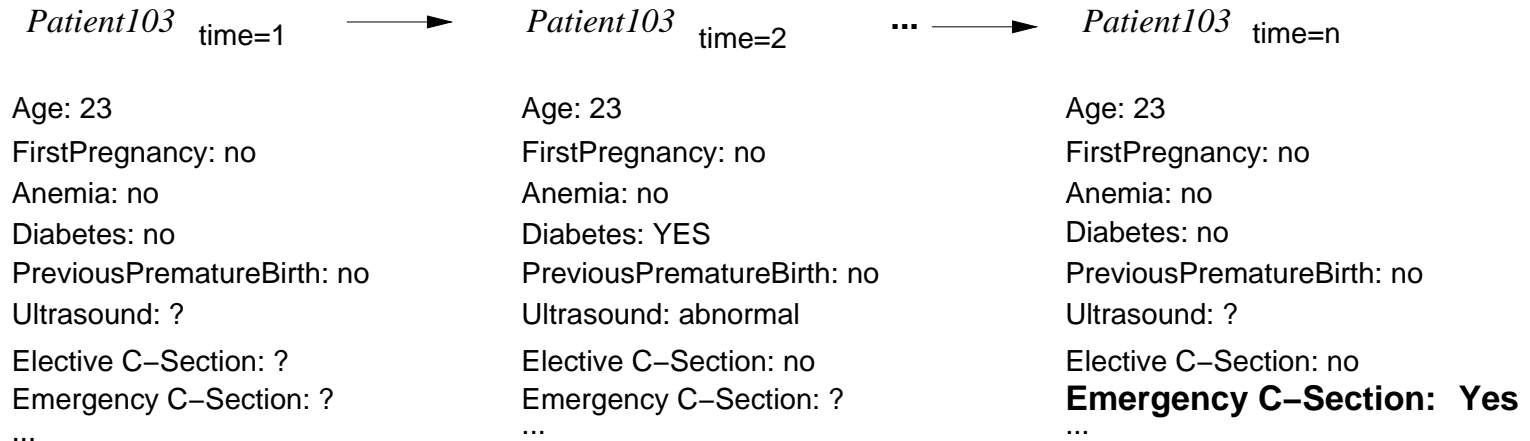
What Is a Learning Problem?

- A *Learning Problem* is defined in terms of a:
 - Task T
 - Performance metric P
 - Type of experience E
- We say that a machine learns if the system reliably improves its performance P at task T , following experience E
- Depending on how we specify T , P , and E , the learning task might also be called by names such as data mining, autonomous discovery, database updating, programming by example, etc.

Place of Machine Learning within Computer Science

- *The application is too complex for people to manually design the algorithm*
 - Software for sensor-base perception tasks, such as speech recognition and computer vision
 - All of us can easily label which photographs contain a picture of our mother, but none of us can write down an algorithm to perform this task
- *The application requires that the software customize to its operational environment after it is fielded*
 - One example of this is speech recognition systems that customize to the user who purchases the software
 - Machine learning here provides the mechanism for adaptation
- The machine learning niche within the software world is growing rapidly:
 - Bookstores that customize to your purchasing preferences
 - Email readers that customize to your particular definition of spam
 - ...

Example Learning Problem I



- Given:
 - 9714 patient records, each describing a pregnancy and birth
 - Each patient record contains 215 features
- Learn to predict:
 - Classes of future patients at high risk for *Emergency Cesarean Section*

Example Learning Results I

<i>Patient103</i> time=1	→	<i>Patient103</i> time=2	...	→	<i>Patient103</i> time=n
Age: 23		Age: 23			Age: 23
FirstPregnancy: no		FirstPregnancy: no			FirstPregnancy: no
Anemia: no		Anemia: no			Anemia: no
Diabetes: no		Diabetes: YES			Diabetes: no
PreviousPrematureBirth: no		PreviousPrematureBirth: no			PreviousPrematureBirth: no
Ultrasound: ?		Ultrasound: abnormal			Ultrasound: ?
Elective C-Section: ?		Elective C-Section: no			Elective C-Section: no
Emergency C-Section: ?		Emergency C-Section: ?			Emergency C-Section: Yes
...	

- One of 18 learned rules:

If No previous vaginal delivery, and
 Abnormal 2nd Trimester Ultrasound, and
 Malpresentation at admission

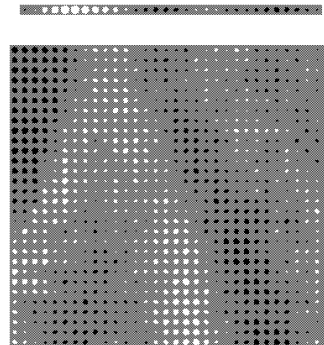
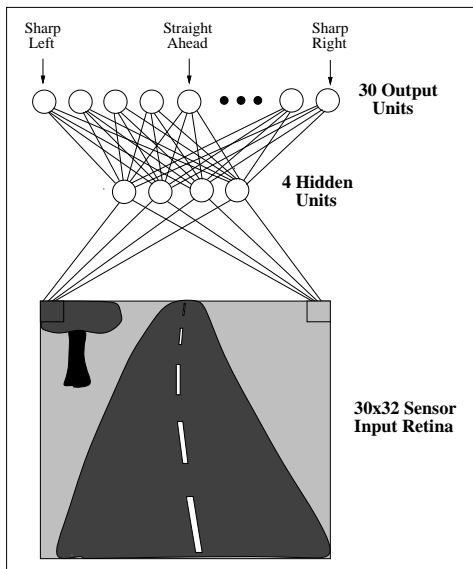
Then Probability of Emergency C-Section is 0.6

Over training data: $26/41 = .63$,

Over test data: $12/20 = .60$

Example Learning Problem II

ALVINN [Pomerleau] drives 70 mph on highways



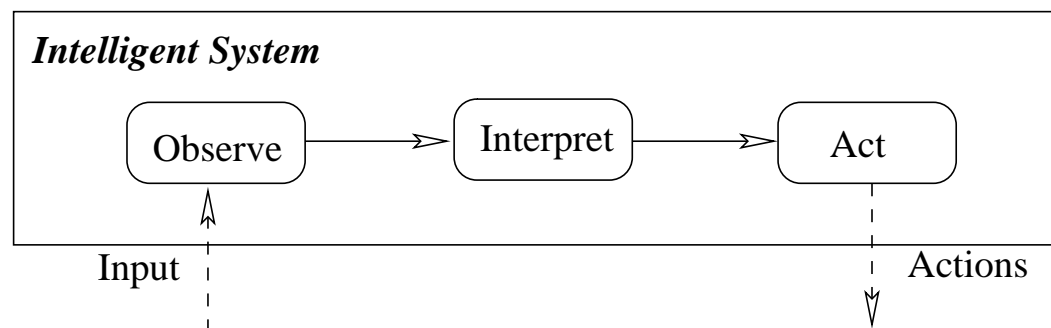
Where is This Headed?

- **Today:** *tip of the iceberg*
 - First-generation algorithms: neural nets, decision trees, regression ...
 - Applied to well-formated database
 - Budding industry
- **Opportunity for tomorrow:** *enormous impact*
 - Learn across full mixed-media data
 - Learn across multiple internal databases, plus the web and news-feeds
 - Learn by active experimentation
 - Learn decisions rather than predictions
 - Cumulative, lifelong learning
 - Programming languages with learning embedded?

Part II:

Lecture Goals

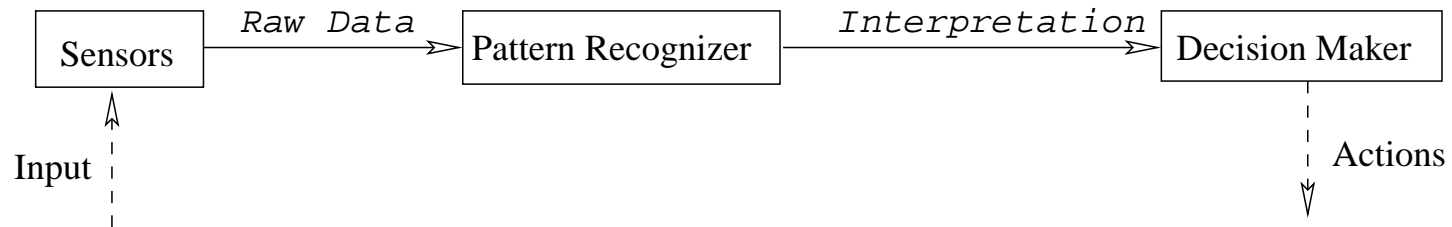
Knowledge and Skills to be Obtained



- *Realize how a real-world system can (1) observe, (2) interpret, and (3) act in an intelligent manner*
- Example systems
 - **Chess Playing:** Observe Board → Interpret Game Situation → Perform Best Move
 - **Spam Detection:** Parse E-mail → Interpret Words and Symbols → Delete or Pass On E-mail
 - **Face Recognition:** Capture Picture → Interpret Pixels → Identify Person

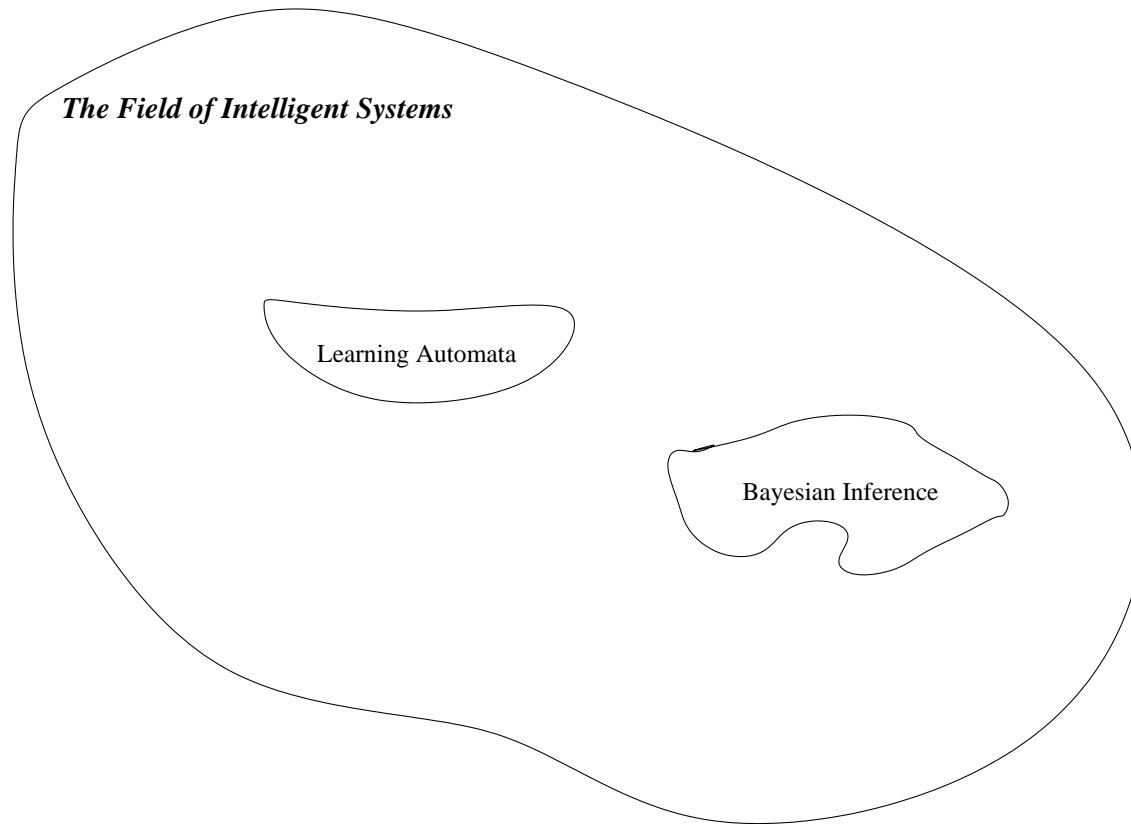
Knowledge and Skills to be Obtained

Basic Building Blocks



- *Appreciate the generic building blocks of an intelligent system*
- **Example building blocks:**
 - *Sensors:* Camera, Microphone, Web Page Parser, Network Packet Sniffer, etc.
 - *Pattern Recognizers:* Neural Networks, Decision Trees, Bayesian Networks, etc.
 - *Decision Makers:* Bayesian Approaches, Multi-Armed Bandit Approaches, Fuzzy Logic Approaches, etc.

Knowledge and Skills to be Obtained



- *Learn how to implement and use a fundamental technique for pattern recognition and a fundamental technique for decision making*
- By concentrating on using a few fundamental techniques, you will learn the techniques more thoroughly, and eventually develop your own understanding and intuition, allowing you to solve problems on your own

Job Opportunities

- Skills in Pattern Recognition and Machine Learning are *interdisciplinary*
- Accordingly, people educated in Pattern Recognition and Machine Learning typically work in a multitude of fields
 - Communications
 - Image Analysis
 - Intrusion Detection
 - Monitoring of Web
 - Environmental Monitoring
 - Robot Control
 - Military Applications
- E.g.: Video Surveillance, Network Intrusion Detection, Monitoring of Word of Mouth on the Web, Telecommunications, Games, etc.

Part III:

Pattern Recognition

What is a Pattern Recognition System?

Pattern Recognition System



- A complete pattern recognition system consists of:
 - A sensor that gathers the observations to be classified
 - A feature extraction mechanism that computes numeric or symbolic information from the observations
 - A classification scheme that does the actual job of classifying observations, relying on the extracted features

Example Systems

Picture Classifier: Pixels \rightarrow Color Histogram \rightarrow Picture Category

Network Packet Anomaly Detector: Network Packet \rightarrow 48 First Bytes \rightarrow
Anomalous?

Locating Discussion Board Users: Posting Time Stamps \rightarrow $\{0, \dots, 23\}$
(Central European Time) \rightarrow User Location (Time Zone)

Categorization of Articles: ? \rightarrow ? \rightarrow ?

Bayesian Inference I

- **Bayesian inference:** Statistical inference in which *observations* are used to update the probability that a *hypothesis* may be *true*
- *Remark:* We are interested in using Bayesian inference for Pattern Recognition

Bayesian Inference II

- **In the courtroom**

- Hypothesis: *The defendant is guilty*
- Observations: *E.g., DNA evidence*

- **Text analysis**

- Hypothesis: *The article discusses mobile phones*
- Observations: *Words occurring in article*

- **Medical diagnosis**

- Hypothesis: ?
- Observations: ?

Motivating a Bayesian Approach I

- **Cox's Theorem:** Any system for *plausible reasoning* intended to ensure
 - consistency with classical deductive logic
 - correspondence with commonsense reasoningis *isomorphic* to probability theory
- **Question:** Why consider anything else?

Motivating a Bayesian Approach II

“I spent about six months writing software that looked for individual spam features before I tried the statistical approach. What I found was that recognizing that last few percent of spams got very hard, and that as I made the filters stricter I got more false positives. [...] I don’t know why I avoided trying the statistical approach for so long. I think it was because I got addicted to trying to identify spam features myself, as if I were playing some kind of competitive game with the spammers. [...] When I did try statistical analysis, I found immediately that it was much cleverer than I had been.”

Paul Graham — Author of a Plan for Spam

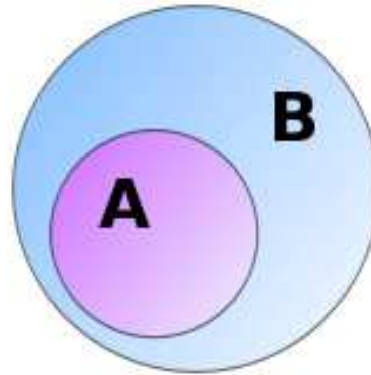
Motivating a Bayesian Approach to Pattern Recognition

- Bayesian inference provides a basis for practical pattern recognition and inference algorithms:
 - Naive Bayes classifier
 - Bayesian belief networks
- Bayesian inference provides a useful conceptual framework
 - Provides “gold standard” for evaluating other pattern recognition and learning algorithms

Probability Theory

- Probability is the likelihood that something *is the case* or *will happen*
- Probability theory is used extensively in areas such as statistics, mathematics, science and philosophy
- The purpose is to draw conclusions about the likelihood of potential events and the underlying mechanics of complex systems

Sample Space and Events



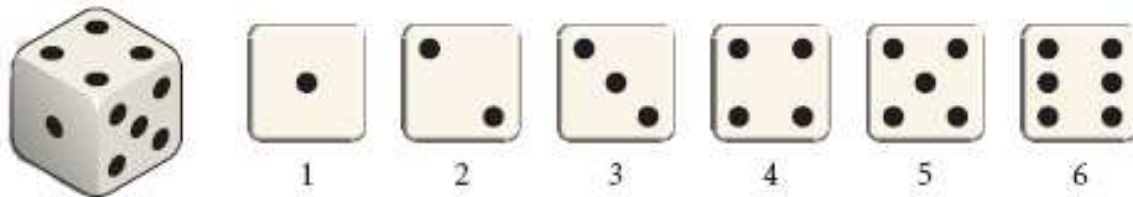
- A sample space is a set of all possible outcomes for an activity or experiment
 - **Rolling a Die:** $\{1, 2, 3, 4, 5, 6\}$
 - **Tossing a Coin:** $\{Heads, Tails\}$
 - **Randomly Selecting a Word from a Document:**
 $\{A, An, Able, Ability, Abler, Ablest, Ably, \dots\}$
- Any *subset* of the sample space is usually called an event
 - **Example of event:** Rolling an even number with a die, i.e., $\{2, 4, 6\}$
- **Question:** What is the sample space when two words are selected at random from a document?

Classical Definition of Probability

If A random experiment can result in N mutually exclusive and equally likely outcomes;

And N_A of these outcomes result in the occurrence of the event A

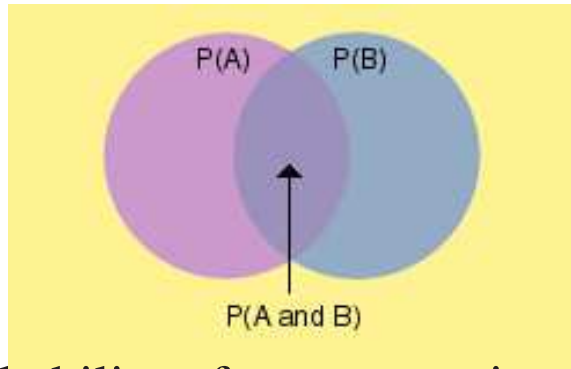
Then The probability of A is defined by $P(A) = \frac{N_A}{N}$



Example:

$$P(\text{"Rolling an even number with a die"}) = P(\{2, 4, 6\}) = \frac{3}{6} = 0.5$$

Joint Probability



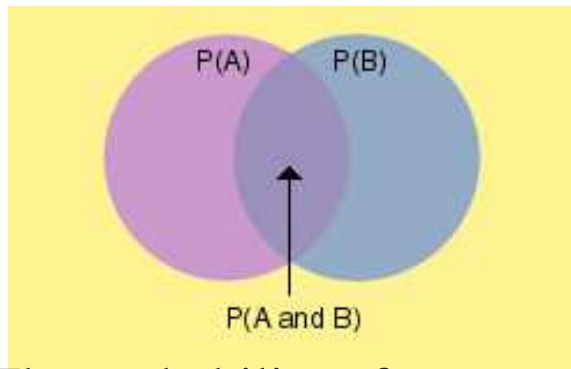
Joint Probability: The probability of two events in conjunction (both events together)

The joint probability of A and B is written $P(A \wedge B)$ or $P(A, B)$

Example:

$$P(\text{"Rolling an even number"} \wedge \text{"Rolling 2"}) = P(\{2, 4, 6\} \wedge \{2\}) = ?$$

Conditional Probability



Conditional Probability: The probability of some event A , given the occurrence of some other event B

Conditional probability is written $P(A|B)$, and is read "*the probability of A, given B*"

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

Question: What is the probability of getting a 2 when tossing a die, given that the outcome of the toss is even?

$$P(\{2\}|\{2, 4, 6\}) = \frac{P(\{2\} \wedge \{2, 4, 6\})}{P(\{2, 4, 6\})} = ?$$

Conditional Independence

- Two events A and B are independent if and only if $P(A \wedge B) = P(A)P(B)$
- Two events A and B are conditionally independent given a third event C precisely if A and B are independent events given C :

$$P(A \wedge B|C) = P(A|C)P(B|C)$$

Bayes' Theorem

Bayes' theorem tells how to update or revise beliefs in light of new evidence

$$P(h|o) = \frac{P(o|h)P(h)}{P(o)}$$

- $P(h)$ = prior probability of hypothesis h
- $P(o)$ = prior probability of observations o
- $P(h|o)$ = probability of h given o
- $P(o|h)$ = probability of o given h

Choosing Hypotheses

$$P(h|o) = \frac{P(o|h)P(h)}{P(o)}$$

We generally want to identify the most probable hypothesis, which we call the *maximum a posteriori* hypothesis h_{MAP} :

$$\begin{aligned} h_{MAP} &= \arg \max_{h \in H} P(h|o) \\ &= \arg \max_{h \in H} \frac{P(o|h)P(h)}{P(o)} \\ &= \arg \max_{h \in H} P(o|h)P(h) \end{aligned}$$

Example: Bayesian Inference in the Courtroom I

- Let h be the event that the defendant is guilty and $\neg h$ be the event that he is innocent
- Let o be the event that the defendant's DNA matches DNA found at the crime scene
- Let $P(o|h) = 1.0$ be the probability of seeing event o assuming that the defendant is guilty
- Let $P(h) = 0.3$ be the juror's personal estimate of the probability that the defendant is guilty, based on the evidence other than the DNA match
- Let $P(o|\neg h) = 10^{-6}$ be the probability that an innocent person chosen at random would have DNA that matched that at the crime scene

Bayes' Theorem tells us that we can calculate $P(h|o)$ — the probability that the defendant is guilty assuming the DNA match event o :

$$P(h|o) = \frac{P(h)P(o|h)}{P(o)} = \frac{P(h)P(o|h)}{P(o, h) + P(o, \neg h)} = \frac{P(h)P(o|h)}{P(h)P(o|h) + P(\neg h)P(o|\neg h)}$$

Example: Bayesian Inference in the Courtroom II

$$P(h|o) = \frac{0.3 \times 1.0}{0.3 \times 1.0 + 0.7 \times 10^{-6}} = 0.99999766667$$

Naive Bayes Classifier I

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
{Sunny, Rainy}	{Warm, Cold}	{Normal, High}	{Weak, Strong}	{Cool, Warm}	{Change, Same}	{Yes, No}

- Let $H = h_j \in \{h_1, h_2, \dots, h_m\}$ be the hypotheses under consideration [†]
- Let $\langle O_1 = o_1, O_2 = o_2, \dots, O_n = o_n \rangle$ be the different kinds of observations that have been made
- Then the most probable hypothesis is:

$$h_{MAP} = \operatorname{argmax}_{h_j \in H} P(h_j | o_1, o_2 \dots o_n)$$

$$h_{MAP} = \operatorname{argmax}_{h_j \in H} \frac{P(o_1, o_2 \dots o_n | h_j) P(h_j)}{P(o_1, o_2 \dots o_n)}$$

$$= \operatorname{argmax}_{h_j \in H} P(o_1, o_2 \dots o_n | h_j) P(h_j)$$

[†] We assume that the hypotheses are *mutually exclusive* (cannot occur together) and *exhaustive* (covers all cases)

Naive Bayes Classifier II

Naive Bayes assumption:

$$P(o_1, o_2 \dots o_n | h_j) = \prod_i P(o_i | h_j)$$

which gives

$$\text{Naive Bayes classifier: } h_{NB} = \operatorname{argmax}_{h_j \in H} P(h_j) \prod_i P(o_i | h_j)$$

How to Find the Probabilities?

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

- **Solution 1:** Fix the probabilities based on expert knowledge
- **Solution 2:** Estimate the probabilities using a set of example data (training set)

— $\hat{P}(\text{EnjoySpt} = \text{Yes}) = \frac{3}{4} = 0.75$

— $\hat{P}(\text{Temp} = \text{Warm} | \text{EnjoySpt} = \text{Yes}) = \frac{3}{3} = 1.0$

— $\hat{P}(\text{Sky} = \text{Sunny} | \text{EnjoySpt} = \text{No}) = ?$

Naive Bayes Algorithm

Naive_Bayes_Learn(*examples*)

For each target value h_j

$$\hat{P}(h_j) \leftarrow \text{estimate } P(h_j)$$

For each observation value o_i of observation O_i

$$\hat{P}(o_i|h_j) \leftarrow \text{estimate } P(o_i|h_j)$$

Classify_New_Instance(x)

$$h_{NB} = \operatorname{argmax}_{h_j \in H} \hat{P}(h_j) \prod_i \hat{P}(o_i|h_j)$$

Naive Bayes: Example

Consider *PlayTennis*, and new instance

$\langle \text{Outlk} = \text{sun}, \text{Temp} = \text{cool}, \text{Humid} = \text{high}, \text{Wind} = \text{strong} \rangle$

Want to compute:

$$h_{NB} = \operatorname{argmax}_{h_j \in H} P(h_j) \prod_i P(o_i | h_j)$$

$$P(y) P(\text{sun}|y) P(\text{cool}|y) P(\text{high}|y) P(\text{strong}|y) = .005$$

$$P(n) P(\text{sun}|n) P(\text{cool}|n) P(\text{high}|n) P(\text{strong}|n) = .021$$

$$\rightarrow h_{NB} = n$$

Naive Bayes: Subtleties

- Conditional independence assumption is often violated

$$P(o_1, o_2 \dots o_n | h_j) = \prod_i P(o_i | h_j)$$

- ...but it works surprisingly well anyway. Note don't need estimated posteriors $\hat{P}(h_j|x)$ to be correct; need only that

$$\operatorname{argmax}_{h_j \in H} \hat{P}(h_j) \prod_i \hat{P}(o_i | h_j) = \operatorname{argmax}_{h_j \in H} P(h_j) P(o_1 \dots, o_n | h_j)$$

- However, note that Naive Bayes posteriors often are unrealistically close to 1 or 0

Learning to Classify Text

Why?

- Learn which news articles are of interest
- Learn to classify web pages by topic

Naive Bayes is among most effective algorithms

What attributes shall we use to represent text documents??

Learning to Classify Text

Target concept *Interesting?* : *Document* $\rightarrow \{+, -\}$

1. Represent each document by vector of words
 - one attribute per word position in document
2. Learning: Use training examples to estimate
 - $P(+)$
 - $P(-)$
 - $P(doc|+)$
 - $P(doc|-)$

Naive Bayes conditional independence assumption

$$P(doc|h_j) = \prod_{i=1}^{length(doc)} P(o_i = w_k|h_j)$$

where $P(o_i = w_k|h_j)$ is probability that word in position i is w_k , given h_j

one more assumption: $P(o_i = w_k|h_j) = P(o_m = w_k|h_j), \forall i, m$

Learning to Classify Text

LEARN_NAIVE_BAYES_TEXT(*Examples*, *H*)

1. Collect all words and other tokens that occur in *Examples*:

Vocabulary \leftarrow all distinct words and other tokens in *Examples*

2. Calculate the required $P(h_j)$ and $P(w_k|h_j)$ probability terms:

For each target value h_j in *H* do:

— $docs_j \leftarrow$ subset of *Examples* for which the target value is h_j

— $P(h_j) \leftarrow \frac{|docs_j|}{|Examples|}$

— $Text_j \leftarrow$ a single document created by concatenating all members of $docs_j$

— $n \leftarrow$ total number of words in $Text_j$ (counting duplicate words multiple times)

— for each word w_k in *Vocabulary*

* $n_k \leftarrow$ number of times word w_k occurs in $Text_j$

* $P(w_k|h_j) \leftarrow \frac{n_k+1}{n+|Vocabulary|}$

Learning to Classify Text

CLASSIFY_NAIVE_BAYES_TEXT(*Doc*)

- *positions* ← all word positions in *Doc* that contain tokens found in *Vocabulary*
- Return h_{NB} , where

$$h_{NB} = \operatorname{argmax}_{h_j \in H} P(h_j) \prod_{i \in \text{positions}} P(o_i | h_j)$$

Twenty NewsGroups

Given 1000 training documents from each group

Learn to classify new documents according to which newsgroup it came from

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

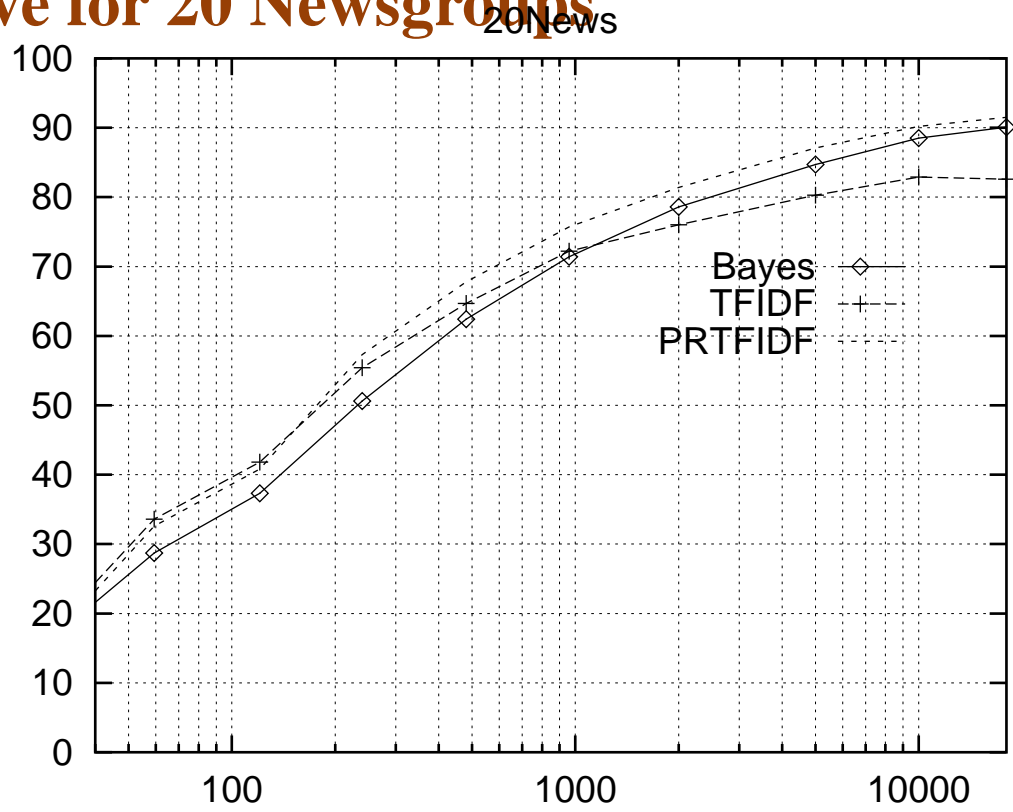
Naive Bayes: 89% classification accuracy

Article from rec.sport.hockey

Path: cantaloupe.srv.cs.cmu.edu!das-news.harvard.edu!ogics
From: xxx@yyy.zzz.edu (John Doe)
Subject: Re: This year's biggest and worst (opinion)...
Date: 5 Apr 93 09:53:39 GMT

I can only comment on the Kings, but the most obvious candidate for pleasant surprise is Alex Zhitnik. He came highly touted as a defensive defenseman, but he's clearly much more than that. Great skater and hard shot (though wish he were more accurate). In fact, he pretty much allowed the Kings to trade away that huge defensive liability Paul Coffey. Kelly Hrudey is only the biggest disappointment if you thought he was any good to begin with. But, at best, he's only a mediocre goaltender. A better choice would be Tomas Sandstrom, though not through any fault of his own, but because some thugs in Toronto decided

Learning Curve for 20 Newsgroups



Accuracy vs. Training set size (1/3 withheld for test)

Part III:

Reinforcement Learning (Next Week)