Intelligent Systems: Reasoning and Recognition

James L. Crowley

ENSIMAG 2 / MoSIG M1 Second Semester 2013/2014

Lesson 13 2 April 2014

# Bayesian Recognition and Reasoning



Bibliographical sources:
"Pattern Recognition and Machine Learning", C. M. Bishop, Springer Verlag, 2006.
"Pattern Recognition and Scene Analysis", R. E. Duda and P. E. Hart, Wiley, 1973.

# Notation

| | |
|---|---|
| x | a variable |
| X | a random variable (unpredictable value) |
| N | The number of possible values for x (Can be infinite). |
| $\vec{x}$ | A vector of D variables. |
| $\vec{X}$ | A vector of D random variables. |
| D | The number of dimensions for the vector $\vec{x}$ or $\vec{X}$ |
| E | An observation. An event. |
| $C_k$ | The class k |
| k | Class index |
| K | Total number of classes |
| $\omega_k$ | The statement (assertion) that $E \in C_k$ |
| $M_k$ | Number of examples for the class k. (think M = Mass) |
| M | Total number of examples. $M = \sum_{k=1}^{K} M_k$ |
| $\{\vec{x}_m^k\}$ | A set of $M_k$ examples for the class k. $\{\vec{x}_m\} = \bigcup_{k=1,K} \{\vec{x}_m^k\}$ |

# Bayesian Recognition and Reasoning

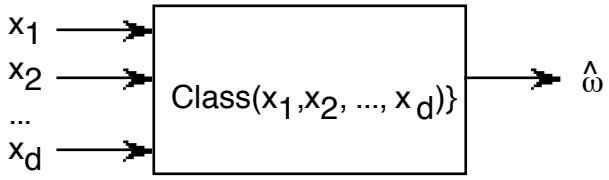
For two proposition A and B, Bayes Rule tells us that

$$P(A, B) = P(A \mid B)\, P(B) = P(B \mid A)\, P(B)$$

This can be used for Recognition or for Reasoning

### Bayesian Recognition

Our problem is to build a box that maps a set of features $\vec{X}$ from an Observation, E into a class $C_k$ from a set of K possible Classes.




Let $\omega_k$ be the proposition that the event E belongs to class k:

$$\omega_k = E \in C_k$$

In order to minimize the number of mistakes, we will maximize the probability that that the event E $\in$ the class k

$$\hat{\omega}_k = \arg\max_k \left\{ \Pr(\omega_k \mid \vec{X}) \right\}$$

A fundamental tool for this is Baye's rule.

$$p(\omega_k \mid \vec{X}) = \frac{p(\vec{X} \mid \omega_k) P(\omega_k)}{p(\vec{X})}$$

**Bayesian Reasoning**

Bayes Rule also gives a technique for using evidence to support a hypothesis.

Let H a hypothesis and let E be evidence for the hypotheses.

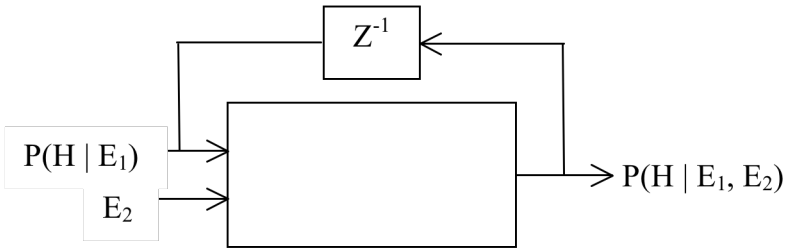
Then Bayes rule tells us

$$P(H, E) = P(H \mid E)\, P(E) = P(E \mid H)\, P(H)$$

so that

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}$$

The problem is how to incrementally accumulate evidence




$Z^{-1}$ is a memory buffer

We will see this in the second half of today's lesson.

**Supervised Learning**

We will use a set of labeled "training set" of samples to estimate the probabilities $p(\vec{X})$, $p(\vec{X} \mid \omega_k)$, and $P(\omega_k)$. This is referred to as "supervised learning".

Assume that we have $K$ classes.
For each class we have a set of $M_k$ sample events $S_k = \{\vec{x}_m^k\}$.

The union of the training samples for each class gives us our training set:

$$S = \{\vec{x}_m\} = \bigcup_{k=1,K} \{\vec{x}_m^k\} \text{ composed of } M = \sum_{k=1}^{K} M_k \text{ samples (think } M = \text{Mass)}$$

In the simplest cases, we can use histogram (tables of frequencies) to represent the probabilities. Alternatively, we can present $p(\vec{X})$, $p(\vec{X} \mid \omega_k)$ as Probability Density Functions.

**Example: Grades in Two Courses**

Suppose we have a set of events described by a pair of properties.
For example, consider the your grade in 2 classes $x_1$ and $x_2$.

Assume your grade is a letter grade from the set {A, B, C, D, F}.

We can build a 2 dimensional hash table, where each letter grade acts as a key into the table $h(x_1, x_2)$.

This hash table has Q= 5 x 5 = 25 cells.

Each student is an observation with a pair of grades $(x_1, x_2)$.

$$\forall m=1, M \text{ : if } h(x_1, x_2) := h(x_1, x_2) + 1;$$

Question: How many students are needed to fill this table?
Answer $M \geq 8Q = 200$.

An example, consider the table as follows:

| | | $x_1$ | | | | | |
|---|---|---|---|---|---|---|---|
| | $h(x_1,x_2)$ | A | B | C | D | F | $r(x_2)$ |
| $x_2$ | A | 2 | 5 | 3 | 1 | | 11 |
| | B | 5 | 16 | 8 | 1 | | 30 |
| | C | 2 | 12 | 20 | 3 | 1 | 38 |
| | D | | 2 | 6 | 2 | 2 | 12 |
| | F | | | 4 | 4 | 1 | 9 |
| | $c(x_1)$ | 9 | 35 | 41 | 11 | 4 | 100 |

Any cell, $(x_1, x_2)$ represents the probability that a student got grade $X_1$ for course $C_1$ and grade $X_2$ for course $C_2$.

$$p(X_1 = x_1 \wedge X_2 = x_2) = \frac{1}{M} h(x_1,x_2)$$

Let us note the sum of column $x_1$ as $c(x_1)$ and sum of row $x_2$ as $r(x_2)$ and the value of cell $x_1, x_2$ as $h(x_1,x_2)$

$$c(x_1) = \sum_{x_2=\{A,B,...F\}} h(x_1,x_2) \qquad r(x_2) = \sum_{x_1=\{A,B,...F\}} h(x_1,x_2)$$

for example $r(x_1=B) = 30$, $C(x_2=B) = 35$, $h(x_1,x_2) = 16$

From this table we can easily see three fundamental laws of probability:

**Sum Rule**

$$p(X_1 = x_1) = \sum_{x_2=\{A,B,...,F\}} p(X_1 = x_1, X_2 = x_2) = \frac{1}{M} \sum_{x_2=\{A,B,...,F\}} h(x_1,x_2) = \frac{1}{M} c(x_1)$$

example: $$p(x_1 = B) = \sum_{x_2=A,B,...,F} p(x_1 = B, x_2) = \frac{1}{M} \sum_{x_2=A,B,...,F} h(B,x_2) = \frac{c(B)}{M} = \frac{35}{100}$$

from which we derive the sum rule: $$p(X_1 = x_1) = \sum_{X_2} p(X_1 = x_1, X_2 = x_2)$$

or more simply $$p(X_1) = \sum_{X_2} p(X_1, X_2)$$

This is sometimes called the "marginal" probability, obtained by "summing out" the other probabilities.

**Conditional probability**

We can define a "conditional" probability as the fraction of one probability given another.

$$p(X_1 = x_1 \mid X_2 = x_2) = \frac{h(x_1,x_2)}{r(x_2)} = \frac{h(x_1,x_2)}{\sum_{x_1} h(x_1,x_2)}$$

For example.

$$p(X_1 = B \mid X_2 = C) = \frac{h(B,C)}{\sum_{x_1} h(x_1,C)} = \frac{12}{38} \text{ and } p(X_2 = C \mid X_1 = B) = \frac{h(B,C)}{\sum_{x_2} h(B,x_2)} = \frac{12}{35}$$

From this, we can derive Bayes rule :

$$p(X_1 \mid X_2)\cdot p(X_2) = \frac{h(X_1,X_2)}{\sum_{X_1} h(X_1,X_2)} \cdot \sum_{X_1} h(X_1,X_2) = h(X_1,X_2) = \frac{h(X_1,X_2)}{\sum_{X_2} h(X_1,X_2)} \cdot \sum_{X_2} h(X_1,X_2) = p(X_2 \mid X_1)\cdot p(X_1)$$

or more simply

$$p(X_1 \mid X_2)\cdot p(X_2) = p(X_2 \mid X_1)\cdot p(X_1)$$

or more commonly written:

$$p(X_1 \mid X_2) = \frac{p(X_2 \mid X_1)\cdot p(X_1)}{p(X_2)}$$

**Product Rule**

We can also use the histogram to derive the product rule.

Note that $p(X_1 = i, X_2 = j) = h(i,j)$

$$p(X_1 = i \mid X_2 = j) = \frac{h(i,j)}{\sum_i h(i,j)}$$

and $p(X_1,X_2) = p(X_1 \mid X_2)\cdot p(X_2)$

These rules show up frequently in machine learning and Bayesian estimation.

Note that we did not need to use numerical values for $x_1$ or $x_2$.

**Symbolic Features**

If the features are symbolic, h(x) is addressed using a hash table, and the feature and feature values act as a hash key. As before h(x) counts the number of examples of each symbol. When symbolic x has N possible symbols then

$p(X = x) = \frac{1}{M} h(x)$ as before

"Bag of Features" methods are increasingly used for learning and recognition.
The only difference is that there is no "order" relation between the feature values.

# Bayesian Reasoning as Evidence Accumulation

Bayesian Reasoning is a widely used technique to validate or invalidate hypothesis using uncertain or unreliable information. With this approach, a hypothesis statement, H, is formulated and assigned a probability, P(H). As new evidence, E, for or against the hypothesis is obtained it is also assigned a probability P(E) as well as a probability that it confirms the hypothesis, P(E|H). Baye's rule can then used to update the probability of the hypothesis:

$$P(H \mid E) \leftarrow \frac{P(E \mid H)P(H)}{P(E)}$$

In Bayesian reasoning, this rule is applied recursively as new evidence is obtained.

Assume that we have two independent evidences, $E_1$ and $E_2$.

$$P(H \mid E_1, E_2) = \frac{P(E_1, E_2 \mid H)P(H)}{P(E_1, E_2)}$$

Assume that we have *K* hypotheses, $H_k$ and we seek to accumulate evidence to select the most likely hypothesis.

Let us define $S=\{E_n\}$ as a body of previous evidence composed of *N* observations, and *E* as a new observation.

Assuming that the new evidence E is independent of the previous evidence S, Baye's Rule tells us:

$$P(H_k \mid E, S) = \frac{P(E, S \mid H_k)P(H_k)}{P(E, S)}$$

Formally, evidence accumulation poses the problem of how to represent the joint probability of the new evidence, E, and all of the past evidence *S*. This is solved by assuming conditional independence for the evidence which may not be strictly true. To solve this we may note that Bayes rule also gives

$$P(H_k \mid E, S)P(E, S) = P(E, S \mid H_k)P(H_k)$$

Because *P(E, S)* is the same for all hypotheses, it can be dropped.
This gives a likelihood for each hypothesis which can be expressed as:

$$L(H_k \mid E,S) = P(E,S \mid H_k)P(H_k)$$

and that

$$P(E,S \mid H_k) = P(E \mid H_k)P(S \mid H_k)$$

so that

$$L(H_k \mid E,S) = P(E \mid H_k)P(S \mid H_k)P(H_k)$$

For $L(H_k \mid E,S)$ to be a probability, we normalize by the sum of all likelihoods.

Which gives $$P(H_k \mid E,S) \leftarrow \frac{P(E \mid H_k)P(S \mid H_k)P(H_k)}{\sum_{j=1}^{K} P(E \mid H_j)P(S \mid H_j)P(H_j)}$$

Thus we can accumulate evidence recursively and arrive at a probability by renormalizing.

Note that the values for the cumulative evidence $P(S \mid H_k)$ are in fact a product of probabilities the N individual evidences, $E_n$.

$$P(H_k \mid S) = \frac{\prod_{n=1}^{N} P(E_n \mid H_k)P(H_k)}{\sum_{j=1}^{K}\prod_{n=1}^{N} P(E_n \mid H_j)P(H_j)}$$

We accumulate this product recursively as we update the probabilities for each hypothesis with new evidence.