Intelligent Systems: Reasoning and Recognition

James L. Crowley

ENSIMAG 2 / MoSIG M1 Second Semester 2009/2010

Lesson 12 26 March 2010

Introduction to Bayesian Recognition

Sources Bibliographiques :
"Neural Networks for Pattern Recognition", C. M. Bishop, Oxford Univ. Press, 1995.
"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.
T_k	The class (tribe) k
k	Class index
K	Total number of classes
ω_k	The statement (assertion) that $E \in T_k$
M_k	Number of examples for the class k. (think M = Mass)
M	Total number of examples. $M = \sum_{k=1}^{K} M_k$
$\{X_m^k\}$	A set of M_k examples for the class k. $\{X_m\} = \bigcup_{k=1,K} \{X_m^k\}$

Bayesian Recognition

Recognition is a fundamental ability for intelligence, and indeed for all life.
To survive, any creature must be able to recognize food, enemies and friends.

Recognize: To identify object or entity as known:

Two forms of recognition: Indentify and Classify

Identify: To recognize an object or entity as an individual
Classify: To recognize an object or entity as a member of a class.

Categorize is sometimes used in place of classify.

A class is a form of set, defined by a membership test.

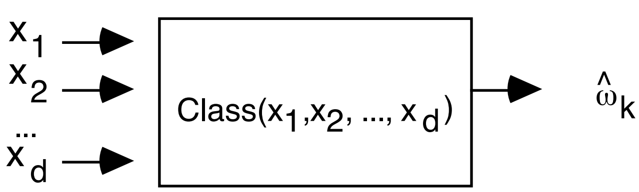
Classification is a process of associated an entity (or an event) as a member of a class. The entity is described by a vector of features, provided by an observation.
The assignment of an entity to a class provided by a test made on a vector of observed properties called features.

observation or event: E. A transduction of the state of the universe, provided by a sensor. Observations are described by a vector of features, $\vec{X}$

Features: observable properties that permit discrimination between classes.
A set of D features, x_d, are assembled into a feature vector $\vec{X}$

$$\vec{X} = \begin{pmatrix} x_1 \\ x_2 \\ \dots \\ x_D \end{pmatrix}$$

A classifier is a process that maps an obervation (or event) to a class based on a vector of observed features.
The result is the proposition $\omega_k = E \in \text{Class } C_k$



The techniques from Pattern recognition and statistics provide a variety of methods to construct membership tests for classification of observations. The most appropriate technique depends on the number and nature of the classes and the features.

There are two families of technique: Generative and discriminative.
These correspond to the two methods to define a set:

Extension: Provide a list of members.
Intension: Provide a conjunction of predicates.

Generative methods compare the features from an observation to a set of prototype examples, using a similarity function, Sim()

$$\hat{\omega}_k = \arg\text{-}\max_k \left\{ Sim(\vec{X}, \vec{X}_m^k) \right\}$$

Discriminative methods apply a set of tests, one for each class.

$$\hat{\omega}_k = \arg\text{-}\max_k \left\{ Test-for-k(\vec{X}) \right\}$$

For a Generative method, we enumerate the M examples of the K classes. The estimate is the most similar, as provided by some simarity function.
Thus for an observed unknown even X, the "estimated" class, $\omega^\wedge{}_k$ is given by :

$$\hat{\omega}_k = \forall_k \forall_m : \arg\text{-}\max_k \left\{ Sim(\vec{X}, \vec{X}_m^k) \right\}$$

Similarity to a prototype

Simple Euclidean distance is often used to measure similarity.

$$\hat{\omega}_k = \forall_k \forall_m : \underset{k}{\arg-\max}\left\{ \| \vec{X}, \vec{X}_m^k \| \right\}$$

A more intelligent method is to normalize the distance by a Metric Λ_k

$$\hat{\omega}_k = \forall_k \forall_m : \underset{k}{\arg-\max}\left\{ (\vec{X} - \vec{X}_m^k)^T \Lambda_k (\vec{X} - \vec{X}_m^k) \right\}$$

We can avoid having to scan all samples by replacing samples of the same class with the average of the samples.

$$\vec{\mu}_k = E\{\vec{X}_m^k\} = \frac{1}{M_k}\sum_{m=1}^{M_k} \vec{X}_m^k$$

Then:

$$\hat{\omega}_k = \forall_k : \underset{k}{\arg-\max}\left\{ (\vec{X} - \vec{\mu}_k)^T \Lambda_k (\vec{X} - \vec{\mu}_k) \right\}$$

where the metric Λ_k is provide by the inverse of the class covariance :$\Lambda_k = C_k^{-1}$

$$C_k = E\{(\vec{X}_m^k - \vec{\mu}_k)^2\}$$

Discriminative tests avoid iterating through the M examples of each class by compiling a series of simple tests. These can be combined in a variety of ways.

A classical (and effective) means is by vote over as large set of simple linear classification functions. We will see more of this later.

In either case, our objective is to minimize the probability of error.

$$\hat{\omega}_k = \underset{k}{\arg-\max}\left\{ \Pr(E \in T_k \mid \vec{X} \right\}$$

The operator "|" is called to as "given" or "provided that". It is the Bayesian conditional operator.

Bayesian Classification

With a Baysian approach, the tests are designed to minimize the number of errors. False positives and false negatives count equally as errors.

An alternative would be to include the cost of error, which may not be the same for a false positive and a false negative. This is an easy extension.

Let ω_k be the proposition that the event belongs to class k: $\omega_k = E \in T_k$

ω_k Proposition that event E $\in$ the class k
$p(\omega_k) = p(E \in T_k)$ Probability that E is a member of class k

Given an observation $\vec{X}$, the decision criteria is

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

where $\omega_k \equiv E \in T_k$

The meaning of "given" is provided by Bayes Rule:

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

Applying Bayes rule for classification will require us to define probability.