Evidential Reasoning

Riccardo Pucella

Bellairs Workshop on Mathemathical Methods for Security

March 23, 2007

We consider the problem of inference on the basis of observations.

Suppose Alice has two coins, F (with probability $1/2$ of landing heads) and B (with probability $3/4$ of landing heads).

She chooses a coin, tosses it, and it lands heads.

How likely was the coin to be F? How likely was it to be B?

Suppose she tosses the coin 100 times, and it lands heads every time. How likely was it to be F? How likely was it to be B?

Note that in the second case, the likelihood of the coin being B has increased—we have gained some evidence for the coin being B. How can we quantify this evidence?

There are various ways of quantifying evidence, the subject of *confirmation theory*.

Idea: we have an experiment (tossing a coin) whose outcomes (landing heads, tails) provide evidenc for an hypothesis (F, B).

Most approaches rely on likelihood functions. The relevant information can be captured using an evidence space

$$\mathcal{E} = (\mathcal{H}, \mathcal{O}, \boldsymbol{\mu})$$

where $\mathcal{H}$ is a finite set of hypotheses, $\mathcal{O}$ is a finite set of observations (or outcomes), and $\boldsymbol{\mu}$ is a iikelihood vector, assigning to every hypothesis h a likelihood function $\boldsymbol{\mu}_h$ which is just a probability distribution on $\mathcal{O}$. (Basically, this is the probability of the observations when the true hypothesis is h.)

The plan of this talk is to (I) review the classical framework, (II) generalize the framework to uncertain likelihoods, and (III) deal with complex experiments.

I: The classical framework

Given an evidence space $\mathcal{E}$, we can quantify the weight of evidence of an observation o towards an hypothesis h in various ways:

The most obvious approach is simply to take

$$w_{\mathcal{E}}(o, h) = \boldsymbol{\mu}_h(o).$$

A generalization of a measure commonly used in statistics is to define

$$w^M_{\mathcal{E}}(o,h) = \frac{\boldsymbol{\mu}_h(o)}{\max_{h' \in \mathcal{H}} \boldsymbol{\mu}_{h'}(o)}.$$

Alternatively, Shafer defines

$$w^S_{\mathcal{E}}(o,h) = \frac{\boldsymbol{\mu}_h(o)}{\sum_{h' \in \mathcal{H}} \boldsymbol{\mu}_{h'}(o)}.$$

Note that each of these measures are of the form

$$w_{\mathcal{E}} = c(o)\boldsymbol{\mu}_h(o) \tag{1}$$

for some c. They have different properties, some more useful than others. An exact assessment of which is more useful is still pending.

In what sense do the above measures capture what we called evidence? Well, suppose we do have a prior probability distribution on $\mathcal{H}$. Clearly, given a prior probability distribution and the likelihoods, we can apply Bayes' theorem to obtain the posterior probability of hypotheses given a particular observation. We can show that weights of evidence as defined above—as long as they satisfy (1)—can be used to update a prior probability directly.

First, define an operation $\oplus$ that takes two functions $f, g : X \to [0,1]$ for some finite set X, and produces

$$f \oplus g = \lambda x. \left(\frac{f(x)g(x)}{\sum_{y \in X} f(y)g(y)} \right).$$

Suppose I am given a prior probability μ_0 on the hypotheses, and I observe $o \in \mathcal{O}$. I claim that μ_o, the posterior probability on hypotheses after observing o, can be computed as

$$\mu_0 \oplus w_{\mathcal{E}}(o, \cdot),$$

where $w_{\mathcal{E}}(o, \cdot)$ is the weight of evidence for a fixed o viewed as a function of h, satisfying (1).

Theorem: *if P is a probability distribution on $\mathcal{H} \times \mathcal{O}$ satisfying $P(\{h\} \times \mathcal{O}) = \mu_o(h)$ and $P(h, o) = \boldsymbol{\mu}_h(o)$ for all h, o, then $\mu_o(h) = P(\{h\} \times \mathcal{O} \mid \mathcal{H} \times \{o\})$.*

Thus, weights of evidence capture all the information required by Bayes' theorem.

The question becomes: to what extent can you make decisions based only on weights of evidence?

Statistics have had a long debate about this, Bayesians versus frequentists, but in Computer Science, we are loathe to posit priors, as opposed to what happens in Economics or in Epidemiology, where priors are often determined experimentally.

Two examples should highlight this claim.

Probabilistic Analysis of Algorithms. Consider Rabin's primality test, which relies on a predicate $P(n, a)$ (for $0 \leq a \leq n-1$) that is equal to 1 for at least $n/2$ choices of a when n is composite, and equal to 0 for all choices of a when n is prime.

Rabin's primality test is used to check if n is prime or composite, and works as follows. Given n, choose a uniformly between 0 and $n-1$, and check $P(n,a)$: if it is 1, answer "composite", and if it is 0, answer "prime". The weights of evidence w^S for this scenario can be computed to be:

w^S	prime	composite
"prime"	$\geq 2/3$	$\leq 1/3$
"composite"	0	1

Intrusion Detection. An intrusion detector is a monitor that observes sequences of actions performed by users of a system, and attempts to determine when those actions are in fact an intrusion as opposed to a reasonable use of the system.

Thus, we can view the intrusion detector as attempting to gather evidence for one of two hypotheses: the current user is a real user, versus the current user is an intruder. The detector has to make a decision, at some point, when to announce an intrusion—this then points back to the question of how to make decisions in the presence of evidence.

II: Uncertain likelihoods

Consider the following more general scenario, that is difficult to capture in the above framework: Alice has two coins as before, one fair (F) and one biased (B); Bob has one coin, double-headed (D). Alice chooses a coin and hands it Zoe, while Bob hands his coin to Zoe. Zoe chooses one of the coins, and tosses it; it lands heads. What is the likelihood that the coin was Alice, versus the likelihood that the coin was Bob's?

The problem here is that the likelihoods are not uniquely given by the hypotheses. For hypothesis A (the coin chosen by Zoe is Alice's), there are two likelihoods possible, one for each possible choice that Alice makes.

A natural approach is to refine the hypotheses. Suppose that we consider instead the hypotheses A_F (the coin chosen by Zoe is Alice's, and it is the fair coin), A_B (the coin chosen by Zoe is Alice's, and it is the biased coin), and B (the coin chosen by Zoe is Bob's). Now the likelihoods are uniquely given by the hypotheses, so we can then apply the above framework to obtain an evidence space. Suppose, however, that we are really interested in the hypothesis A—how do we obtain the weight of evidence of an observation for A from the weights of evidence of the observation for A_F and A_B? In general, we cannot. In fact, as we now show, there are some scenarios for which refinement is not even a possibility.

Here is a direct approach to deal with the scenario above. We consider all possible evidence spaces over $\mathcal{H}$ and $\mathcal{O}$ that are consistent with the information in the scenario; in particular, for every possible choice made in the scenario, we have a different evidence space corresponding to the scenario and those choices. We then compute weights for all those evidence spaces, and put them together into a set of weights of evidence.

A generalized evidence space is a tuple

$$\mathcal{G} = (\mathcal{H}, \mathcal{O}, \Delta)$$

where Δ is a set of likelihood vectors.

Let $S(\mathcal{G}) = \{(\mathcal{H}, \mathcal{O}, \boldsymbol{\mu}) \mid \boldsymbol{\mu} \in \Delta\}$.

We can then compute the set of weights $w_{\mathcal{G}}(o, h)$ for observation o and hypothesis h by taking $\{w_{\mathcal{E}}(o, h) \mid \mathcal{E} \in S(\mathcal{G})\}$.

Just like standard evidence spaces, weights of evidence capture all the information required to compute posterior probabilities. Given a prior probability μ_0 on the hypotheses, let $\mathcal{P}_o$ be the set of posterior probability distributions after observing o, where each posterior probability distribution in the set is obtained from a different likelihood vector in Δ. If we write $\mathcal{P}_o(h)$ for $\{\mu(h) \mid \mu \in \mathcal{P}_o\}$, it is not hard to see that

$$\mathcal{P}_o(h) = \{\mu_0 \oplus w_{\mathcal{E}}(o, \cdot) \mid \mathcal{E} \in S(\mathcal{G})\}.$$

How do generalize evidence space relate to the refinement approach described earlier, which has the advantage of being intuitive?

First, what is a refinement? We say that $(\mathcal{H}', \mathcal{O}, \boldsymbol{\mu}')$ refines $(\mathcal{H}, \mathcal{O}, \Delta)$ if (1) every hypothesis in $\mathcal{H}$ is refined by a set of hypotheses in $\mathcal{H}'$, that is, there is a surjective function $g : \mathcal{H}' \to \mathcal{H}$, and (2) for every $\boldsymbol{\mu} \in \Delta$ and every $h \in \mathcal{H}$, $\boldsymbol{\mu}_h = \boldsymbol{\mu}'_{h'}$ for some $h' \in g^{-1}(h)$, and vice versa.

Let $\mathcal{E}$ be a refinement of $\mathcal{G}$. Let μ_0 be a prior probability on $\mathcal{H}$. Let $Ext(\mu_0)$ be the set of probability distribution on $\mathcal{H}'$ that extend μ_0, in the following sense: $\mu'_0 \in Ext(\mu_0)$ if and only if $\mu_0(h) = \mu'_0(g^{-1}(h))$. It is possible to show that:

$$(\mathcal{P}_o)^*(h) = \{\mu'_0 \oplus w_{\mathcal{E}}(o, \cdot) \mid \mu'_0 \in Ext(\mu_0)\}^*(g^{-1}(h)),$$

where $\mathcal{P}^*(h) = sup\{\mu(h) \mid \mu \in \mathcal{P}\}$ for a set $\mathcal{P}$ of probability distributions. (A similar results holds for the inf.)

So, if a generalized evidence can be refined, then for a given prior probability on the hypotheses, the bounds on the posterior probabilities after seeing an observation as given by the generalized evidence space and as given by the refinement approach agree. So while we cannot in general get the weight of evidence of a compound hypothesis from the weights of evidence of the refinements, we can still use refinement to obtain useful bounds on probabilities.

However, when can we use refinement? Call Δ uncorrelated if there exists sets of probability distributions $\mathcal{P}_h$ for every h such that $\boldsymbol{\mu} \in \Delta$ if and only if $\boldsymbol{\mu}_h \in \mathcal{P}_h$. Call $(\mathcal{H}, \mathcal{O}, \Delta)$ uncorrelated when Δ is uncorrelated.

Theorem: *There exists an evidence space refining $\mathcal{G}$ if and only if $\mathcal{G}$ is uncorrelated.*

The point is that there are scenarios where $\mathcal{G}$ is not uncorrelated. For instance, suppose that Alice and Bob both have two coins, call them a_1, a_2 and b_1, b_2. They each give a coin to Zoe, who tosses a coin of her chosing. Suppose that Alice and Bob agree beforehand that they are to give a_1, b_1 or a_2, b_2 to Zoe. The resulting generalized evidence space capturing this scenario is easily seen to not be uncorrelated.

III: Complex experiments

The above framework gives a general theory for dealing with experiments where likelihoods are uncertain. Following a discussion with Catuscia last year, I started to examine exactly how to derive a generalized evidence space from a description of a scenario. The literature tends to focus on simple scenarios that are easy to analyze. Is there a mechanical way of deriving an evidence space from a description of the scenario, however complicated it may be?

The answer is to define a language in which to express scenarios, and give a semantics to that language in such a way that a generalized evidence space corresponding to the scenario can be easily derived.

To give a sense for the language, consider the first scenario used to illustrate generalized evidence spaces above, where Alice has two coins and Bob has one. It can be writen as follows:

$$
\begin{array}{l}
A \leftarrow \mathsf{choose}[\{H:0.5, T:0.5\}, \{H:0.75, T:0.25\}]; \\
B \leftarrow \{H:1, T:0\}; \\
\mathsf{if}\ (\mathsf{hyp} = A)\ \mathsf{then}\ Z \leftarrow A\ \mathsf{else}\ Z \leftarrow B; \\
\mathsf{observe}\ Z
\end{array}
$$

More complex scenarios can also be captured:

$$
\begin{array}{l}
A \leftarrow \mathsf{choose}[\{H:0.5, T:0.5\}, \{H:0.75, T:0.25\}]; \\
B \leftarrow \{H:1, T:0\}; \\
\mathsf{if}\ (\mathsf{hyp} = A)\ \mathsf{then}\ Z \leftarrow A\ \mathsf{else}\ Z \leftarrow B; \\
\mathsf{repeat}\ 100\ \mathsf{times} \\
\quad \mathsf{observe}\ Z
\end{array}
$$

or

$$
\begin{array}{l}
\mathsf{repeat}\ 100\ \mathsf{times} \\
\quad A \leftarrow \mathsf{choose}[\{H:0.5, T:0.5\}, \{H:0.75, T:0.25\}]; \\
\quad B \leftarrow \{H:1, T:0\}; \\
\quad \mathsf{if}\ (\mathsf{hyp} = A)\ \mathsf{then}\ Z \leftarrow A\ \mathsf{else}\ Z \leftarrow B; \\
\quad \mathsf{observe}\ Z
\end{array}
$$

which have subtly different behavior.

While I will not go into the details, it is straightforward to give a semantics of the following form to this statements in this language:

$$\mathfrak{Ev}[\![S]\!] : \Sigma \times PD(\mathcal{O}^*) \longrightarrow \wp(\Sigma \times PD(\mathcal{O}^*)),$$

where Σ is a set of states that capture the values of all variables in the program, and $PD(\mathcal{O}^*)$ is the set of probability distributions over $\mathcal{O}^*$. Intuitively, the semantics takes a state and a current likelihood for the hypothesis held in the state (the hypothesis is held in a variable hyp), and returns the set of possible states and likelihoods obtained by running the experiment.

The semantics tells us how to associate to every experiment $(\mathcal{H}, \mathcal{O}, S)$ an evidence space $\mathcal{E} = (\mathcal{H}, \mathcal{O}^*, \Delta)$ by taking

$$\Delta = \{\boldsymbol{\mu} \mid \exists \sigma.\sigma(\mathsf{hyp}) = h \wedge (\sigma, \boldsymbol{\mu}_h) \in \mathfrak{Ev}[\![S]\!](\sigma_0[\mathsf{hyp} \mapsto h], \mu_0)\}$$

where σ_0 assigns a default value to every variable, and μ_0 is the probability distribution that assigns 1 to the empty sequence $\langle\rangle$ and 0 to every other sequence.

It is possible to justify this semantics by consider a standard probabilistic semantics for such a language, given by

$$\mathfrak{Pr}[\![S]\!] : \Sigma \times \mathcal{O}^* \longrightarrow \wp(PD(\Sigma \times \mathcal{O}^*)),$$

which can be lifted to

$$\mathfrak{Tr}[\![S]\!] : PD(\Sigma \times \mathcal{O}^*) \longrightarrow \wp(PD(\Sigma \times \mathcal{O}^*)).$$

One can check that the posterior probability distributions obtained by updating a prior distribution via the generalized weights of evidence for an observation are the same as obtained by conditioning the results of the $\mathfrak{Tr}[\![-]\!]$ semantics on the lifted prior distribution (lifted to the initial states) on the observation made, and projecting the probability down to $\mathcal{H}$. The details, and in fact even the statement of this result, are left for a future presentation.