Inference of Probability Distributions for Trust and Security applications

Vladimiro Sassone
Based on joint work with
Mogens Nielsen & Catuscia Palamidessi
Outline
Outline

• Motivations
Outline

• Motivations

• Bayesian vs Frequentist approach
Outline

• Motivations
• Bayesian vs Frequentist approach
• A class of functions to estimate the distribution
Outline

• Motivations
• Bayesian vs Frequentist approach
• A class of functions to estimate the distribution
• Measuring the precision of an estimation function
Motivations

• Inferring the probability distribution of a random variable

• Examples of applications in Trust & Security
  • How much we can trust an individual or a set of individuals
  • Input distribution in a noisy channel to compute the Bayes risk
  • Application of the Bayesian approach to hypothesis testing (anonymity, information flow)
  • ...

Nielsen, Palamidessi, Sassone
Setting and assumptions

- For simplicity we consider only binary random variables
  - honest/dishonest, secure/insecure, ...
- Goal: infer (an approximation of) the probability of success
- Means: Sequence of $n$ trials.
  Observation (Evidence): $s, f$

\[
X = \{ \text{succ}, \text{fail} \}
\]

\[
Pr(\text{succ}) = \theta
\]

\[
s = \#\text{succ}
\]

\[
f = \#\text{fail} = n - s
\]
Using the evidence to infer $\theta$

- The Frequentist method:
  \[ F(n, s) = \frac{s}{n} \]

- The Bayesian method:
  Assume an \textit{a priori} probability distribution for $\theta$ (representing your partial knowledge about $\theta$, whatever the source may be) and combine it with the evidence, using Bayes’ theorem, to obtain the \textit{a posteriori} distribution.
Bayesian vs Frequentist

- Criticisms to the frequentist approach
  - **Limited applicability:** sometimes it is not possible to measure the frequencies (in this talk we consider the case in which this is possible)
  - Eg: what is the probability that my submitted paper will be accepted?
  - **Misleading evidence:** For small samples (small n) we can be unlucky, i.e. get unlikely results
  - This is less dramatic for the Bayesian approach because the a priori distribution reduces the effect of a misleading evidence, provided it is close enough to the real distribution
Bayesian vs Frequentist

- Criticisms to the Bayesian approach
  - We need to assume an a priori probability distribution; as we usually do not know the real distribution, the assumption can be somehow arbitrary and differ significantly from reality

- Observe that the two approaches give the same result as $n$ tends to infinity: the “true” distribution
  - Frequentist approach: because of the law of large numbers
  - Bayes approach: because the a priori “washes out” for large values of $n$. 
The surprising thing is that the Frequentist approach can be worse than the Bayesian approach even when the trials give a “good” result, or when we consider the average difference (from the “true” $\theta$) wrt all possible results.

Example: “true $\theta$” = 1/2, $n = 1$

$$F(n, s) = \frac{s}{n} = \begin{cases} 0 & s = 0 \\ 1 & s = 1 \end{cases}$$

The difference from the true distribution is 1/2.
Bayesian vs Frequentist

The surprising thing is that the Frequentist approach can be worse than the Bayesian approach even when the trials give a “good” result, or when we consider the average difference (from the “true” $\theta$) wrt all possible results.

Example: “true $\theta” = 1/2, $n = 1

\[
F(n, s) = \frac{s}{n} = \begin{cases} 
0 & s = 0 \\
1 & s = 1 
\end{cases}
\]

The difference from the true distribution is $1/2$

A better function would be

\[
F_c(n, s) = \frac{s + 1}{n + 2} = \begin{cases} 
\frac{1}{3} & s = 0 \\
\frac{2}{3} & s = 1 
\end{cases}
\]

The difference from the true distribution is $1/6$
Bayesian vs Frequentist

The surprising thing is that the Frequentist approach can be worse than the Bayesian approach even when the trials give a “good” result, or when we consider the average difference (from the “true” $\theta$) wrt all possible results.

Example: “true $\theta$” = 1/2, $n = 2$

$$F(n, s) = \frac{s}{n} = \begin{cases} 
0 & s = 0 \\
\frac{1}{2} & s = 1 \\
1 & s = 2 \end{cases}$$

The average difference from the true distribution is 1/4.
Bayesian vs Frequentist

The surprising thing is that the Frequentist approach can be worse than the Bayesian approach even when the trials give a “good” result, or when we consider the average difference (from the “true” $\theta$) w.r.t. all possible results.

Example: “true $\theta$” = 1/2, $n = 2$

$$\Pr(s)$$

<table>
<thead>
<tr>
<th>$s$</th>
<th>$Pr(s)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1/2</td>
</tr>
<tr>
<td>1/4</td>
<td>1/4</td>
</tr>
<tr>
<td>1/2</td>
<td>1/2</td>
</tr>
<tr>
<td>3/4</td>
<td>1/4</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

$$F(n, s) = \frac{s}{n} = \begin{cases} 0 & s = 0 \\ \frac{1}{2} & s = 1 \\ 1 & s = 2 \end{cases}$$

The average distance from the true distribution is $1/4$.

Again, a better function would be

$$F_c(n, s) = \frac{s + 1}{n + 2} = \begin{cases} \frac{1}{4} & s = 0 \\ \frac{1}{2} & s = 1 \\ \frac{3}{4} & s = 2 \end{cases}$$

The average distance from the true distribution is $1/8$. 
Bayesian vs Frequentist
Bayesian vs Frequentist

- We will see that $F_c(s,n) = \frac{(s+1)}{(n+2)}$ corresponds to one of the possible Bayesian approaches.
Bayesian vs Frequentist

• We will see that $F_c(s,n) = (s+1)/(n+2)$ corresponds to one of the possible Bayesian approaches.

• Of course, if the “true” $\theta$ is different from 1/2 then $F_c$ can be worse than $F$. 
Bayesian vs Frequentist

- We will see that $F_c(s,n) = (s+1)/(n+2)$ corresponds to one of the possible Bayesian approaches.

- Of course, if the “true” $\theta$ is different from $1/2$ then $F_c$ can be worse than $F$.

- And, of course, the problem is that we don’t know what $\theta$ is (the value $\theta$ is exactly what we are trying to find out!).
Bayesian vs Frequentist

• We will see that \( F_c(s,n) = (s+1)/(n+2) \) corresponds to one of the possible Bayesian approaches.

• Of course, if the “true” \( \theta \) is different from \( 1/2 \) then \( F_c \) can be worse than \( F \).

• And, of course, the problem is that we don’t know what \( \theta \) is (the value \( \theta \) is exactly what we are trying to find out!).

• However, \( F_c \) is still better than \( F \) if we consider the average distance wrt all possible \( \theta \in [0,1] \), assuming that they are all equally likely (i.e. that \( \theta \) has a uniform distribution).
Bayesian vs Frequentist

• We will see that $F_c(s,n) = (s+1)/(n+2)$ corresponds to one of the possible Bayesian approaches.

• Of course, if the “true” $\theta$ is different from $1/2$ then $F_c$ can be worse than $F$.

• And, of course, the problem is that we don’t know what $\theta$ is (the value $\theta$ is exactly what we are trying to find out!).

• However, $F_c$ is still better than $F$ if we consider the average distance wrt all possible $\theta \in [0,1]$, assuming that they are all equally likely (i.e. that $\theta$ has a uniform distribution).

• In fact we can prove that, under a suitable notion of “difference”, and for $\theta$ uniformly distributed, $F_c$ is the best function of the kind $G(s,n) = (s+t)/(n+m)$. 
A Bayesian approach

- **Assumption**: $\theta$ is the generic value of a continuous random variable $\Theta$ whose probability density is a *Beta distribution* with (unknown) parameters $\sigma, \varphi$

\[
B(\sigma, \varphi)(\theta) = \frac{\Gamma(\sigma+\varphi)}{\Gamma(\sigma)\Gamma(\varphi)} \theta^{\sigma-1}(1 - \theta)^{\varphi-1}
\]

where $\Gamma$ is the extension of the factorial function i.e. $\Gamma(n) = (n - 1)!$ for $n$ natural number

- Note that the uniform distribution is a particular case of Beta distribution, with $\sigma = 1, \varphi = 1$

- $B(\sigma, \varphi)$ can be seen as the a posteriori probability density of $\Theta$ given by a uniform a priori (principle of maximum entropy) and a trial sequence resulting in $\sigma - 1$ successes and $\varphi - 1$ failures.
Examples of Beta Distribution
Examples of Beta Distribution

\( \sigma = \varphi = 1 \ldots 6 \)
Examples of Beta Distribution

\[ \sigma = \varphi = 1 \ldots 6 \]

\[ \sigma = 1 \ldots 6 \quad \varphi = 2 \sigma \]
Other examples of Beta Distribution
The Bayesian Approach

• Assume an *a priori* probability distribution for $\Theta$ (representing our partial knowledge about $\Theta$, whatever the source may be) and combine it with the *evidence*, using Bayes’ theorem, to obtain the *a posteriori* probability distribution

$$P_{d}(\theta \mid s) = \frac{Pr(s \mid \theta) \cdot P_{d}(\theta)}{Pr(s)}$$

• One possible definition for the estimation function (*algorithm*) is the mean of the *a posteriori* distribution

$$A(n, s) = E_{P_{d}(\theta \mid s)}(\Theta) = \int_{0}^{1} \theta \cdot P_{d}(\theta \mid s) \, d\theta$$
The Bayesian Approach

- Since the distribution of $\Theta$ is assumed to be a beta distribution $B(\sigma, \varphi)$, it is natural to take as a priori a function of the same class, i.e. $B(\alpha, \beta)$.

- In general we don’t know the “real parameters” $\sigma, \varphi$, hence $\alpha, \beta$ may be different from $\sigma, \varphi$.

- The likelihood $Pr(\ s \ | \ \theta)$ is a binomial, i.e.

$$Pr(s \ | \ \theta) = \binom{s + f}{s} \quad \theta^s \ (1 - \theta)^f$$

- The Beta distribution is a conjugate of the binomial, which means that the application of Bayes theorem gives as a posteriori a function of the same class, and more precisely

$$Pd(\theta \ | \ s) = B(\alpha + s, \beta + f)$$
The Bayesian Approach

- Summarizing, we are considering three probability density functions for $\Theta$:
  - $B(\sigma, \varphi)$: the “real” distribution of $\Theta$
  - $B(\alpha, \beta)$: the a priori (the distribution of $\Theta$ up to our best knowledge)
  - $B(s + \alpha, f + \beta)$: the a posteriori

- The result of the mean-based algorithm is:

$$A_{\alpha,\beta}(n, s) = E_{B(s+\alpha, f+\beta)}(\Theta) = \frac{s + \alpha}{s + f + \alpha + \beta} = \frac{s + \alpha}{n + \alpha + \beta}$$
The Bayesian Approach

- The frequentist method can be seen as the limit of the Bayesian mean-based algorithms, for \( \alpha, \beta \to 0 \)

- Intuitively, the Bayesian mean-based algorithms give the best result for \( \alpha / (\alpha + \beta) = \theta \) and \( \alpha, \beta \to \infty \)

- How can we compare two Bayesian algorithms in general, i.e. independently of \( \theta \)?
Measuring the precision of Bayesian algorithms

- Define a “difference” $D(A(n,s), \theta)$ (possibly a distance, but not necessarily. It does not need to be symmetric)
  - non-negative
  - zero iff $A(n,s) = \theta$
  - what else?

- Consider the expected value $D_E(A,n,\theta)$ of $D(A(n,s), \theta)$ with respect to the likelihood (the conditional probability of $s$ given $\theta$)

$$D_E(A, n, \theta) = \sum_{s=0}^{n} \Pr(s \mid \theta) \cdot D(A(n, s), \theta)$$

- **Risk of $A$**: the expected value $R(A,n)$ of $D_E(A,n,\theta)$ with respect to the “true” distribution of $\Theta$

$$R(A, n) = \int_0^1 Pd(\theta) \cdot D_E(A, n, \theta) \, d\theta$$
Measuring the precision of Bayesian Algorithms

- Note that the definition of “Risk of A” is general, i.e. it is a natural definition for any estimation algorithm (not necessarily Bayesian or mean-based)

- What other conditions should $D$ satisfy?

- It seems natural to require that $D$ be such that $R(A,n)$ has a minimum (for all $n$’s) when the a priori distribution coincides with the “true” distribution

- It is not obvious that such $D$ exists
Measuring the precision of Bayesian Algorithms

We have considered the following candidates for $D(x,y)$ (all of which can be extended to the n-ary case):

- The norms:
  - $|x - y|$  
  - $|x - y|^2$  
  - $...$
  - $|x - y|^k$
  - $...$

- The Kullback-Leibler divergence

$$D_{KL}((y, 1 - y) \parallel (x, 1 - x)) = y \log_2 \frac{y}{x} + (1 - y) \log_2 \frac{1 - y}{1 - x}$$
Measuring the precision of Bayesian algorithms

- **Theorem.** For the mean-based Bayesian algorithms, with a priori $B(\alpha, \beta)$, we have that the condition is satisfied (i.e. the Risk is minimum when $\alpha, \beta$ coincide with the parameters $\sigma, \varphi$ of the “true” distribution), by the following functions:
  - The 2nd norm $(x - y)^2$
  - The Kullback-Leibler divergence

- We find it very surprising that the condition is satisfied by these two very different functions, and not by any of the other norms $|x - y|^k$ for $k \neq 2$
Inference of Probability Distributions for trust and security

\[ D(x, y) = (x - y)^2 \]
\[ \sigma = 1, \varphi = 1 \]

\[ D_E(A_{\alpha, \beta}, 5, 1/2) \]
\[ n = 5, \theta = 1/2 \]

\[ R(A_{\alpha, \beta}, 5) \]
\[ n = 5 \]

For the Kullback-Leibler divergence the plots are similar, but much more steep, and they diverge for \( \alpha \to 0 \) or \( \beta \to 0 \)
Work in progress

• Note that for the 2nd norm $D(x,y) = (x-y)^2$ the average $D_E$ is a distance. This contrasts with the case of $D(x,y) = D_{KL}(y||x)$ and makes the first more appealing.

• How robust is the theorem that “certifies” that the 2nd-norm-based $D_E$ is a “good” distance? In particular:
  • Does it extend to the case of multi-valued random variables?
  • Note that in the multi-valued case the likelihood is a multinomial, the conjugate a priori is a Dirichelet and the $D$ is the Euclidian distance (squared)

• What are the possible applications?
Possible applications (work in progress)

- We can use $D_E$ to compare two different estimation algorithms.
  - Mean-based vs other ways of selecting a $\theta$
  - Bayesian vs non-Bayesian
- In more complicated scenarios there may be different Bayesian mean-based algorithms. Example: noisy channel.
- $D_E$ induces a metric on distributions. Bayes’ equations define transformations on this metric space from the a priori to the a posteriori. We intend to study the properties of such transformations in the hope that they will reveal interesting properties of the corresponding Bayesian methods, independent of the a priori.