04 - Bayesian paradigm

Master in Foundations of Data Science

Bayesian Statistics and Probabilistic Programming

Fall 2018-2019

Josep Fortiana

Monday, October 15, 2018

Universitat de Barcelona

04 - Bayesian paradigm

What is a *paradigm?*

What is a Statistical model?

Classical (frequentist) treatment

Bayesian perspective

04 - Bayesian paradigm

What is a *paradigm?*

What is a Statistical model?

Classical (frequentist) treatment

Bayesian perspective

The word, in general

From Wikipedia: Paradigm comes from Greek παράδειγμα (paradeigma), "pattern, example, model".

Derives from a verb meaning to place an object beside another one to compare them.

In Thomas Kuhn's *The Structure of Scientific Revolutions*

Set of concepts and practices that define a scientific discipline at any particular period:

- What is to be observed
- The kind of questions that are supposed to be asked
- How these questions are to be structured
- What predictions are or can be made
- How the results of scientific investigations should be interpreted

04 - Bayesian paradigm

What is a *paradigm?*

What is a Statistical model?

Classical (frequentist) treatment

Bayesian perspective

A Quality Control problem with data

Assume we are in charge of Quality Control for a production line of an electronic component.

We want to assess the proportion of defective units so that, hopefully, technicians will keep it under a given, very small value.

A Quality Control problem with data

Assume we are in charge of Quality Control for a production line of an electronic component.

We want to assess the proportion of defective units so that, hopefully, technicians will keep it under a given, very small value.

Uncertainty

Dealing with uncertainty

Statistical model: a mathematical construct including randomness.

Dealing with uncertainty

Statistical model: a mathematical construct including randomness.

Observations x are values of a r.v. X

Dealing with uncertainty

Statistical model: a mathematical construct including randomness.

Observations x are values of a r.v. X

What do we know about X (pmf/pdf, cdf)?

Repertoire

Usually a model consists of several pdf's. Either:

- Two pdf's (e.g., fair *vs* loaded die)
- Many pdf's, indexed by parameters (parametric statistics)
- "Many-many" pdf's (nonparametric statistics)

A model for the Quality Control problem

Observed data: either $x = 0$, meaning correct, or $x = 1$, meaning defective component.

Statistical model: x is a binary variable, the value of a Bernoulli r.v. X.

This is a parametric model: $X \sim \mathsf{Ber}(\theta)$. $\theta \in (0, 1)$, the unknown proportion of defective components, is the parameter.

All models are wrong, some are useful

Since all models are wrong the scientist cannot obtain a "correct" one by excessive elaboration.

On the contrary following William of Occam he should seek an economical description of natural phenomena.

Just as the ability to devise simple but evocative models is the signature of the great scientist so overelaboration and overparameterization is often the mark of mediocrity.

George E. P. Box (1976)

Models are not the reality

In general, when building statistical models, we must not forget that the aim is to understand something about the real world.

Or predict, choose an action, make a decision, summarize evidence, and so on, but always about the real world, not an abstract mathematical world: our models are not the reality—a point well made by George Box in his oft-cited remark that "all models are wrong, but some are useful".

David Hand (2014)

04 - Bayesian paradigm

What is a *paradigm?*

What is a Statistical model?

Classical (frequentist) treatment

Bayesian perspective

The classical framework in Statistics

The *classical, Neyman-Pearson, or frequentist* paradigm assumes there are:

1. Observable r.v. X with values on a *sample space* $\mathscr{X}$.

2. Parameters θ, constant in the sense they have a fixed, non-random value.
   Their value is always unknown, but belongs to a given *parameter space* Θ.

Classical parameter estimation

Finding and studying *statistics*, functions of the observed X's which "contain some information" about the unknown parameters θ.

Such statistics are called *estimators* of the unknown parameters.

Frequentist interpretation

Theoretical properties of parameter estimators:

- Bias
- Risk
- Efficiency,

are translated into statistical properties of a collection of estimations, the result of a potential large number of realizations of the experiment.

Example: empirical mean

The empirical mean $\overline{X}_n$ of a random sample:

$$X_1, \ldots, X_n \quad \text{i.i.d.} \sim \mathrm{N}(\mu, \sigma^2),$$

is never equal to the unknown parameter μ.

Example: empirical mean

The empirical mean $\overline{X}_n$ of a random sample:

$$X_1, \ldots, X_n \quad \text{i.i.d.} \sim \mathrm{N}(\mu, \sigma^2),$$

is never equal to the unknown parameter μ.

More precisely, we will never know if these two quantities coincide.

Example: empirical mean

The empirical mean $\overline{X}_n$ of a random sample:

$$X_1, \ldots, X_n \quad \text{i.i.d.} \sim \mathrm{N}(\mu, \sigma^2),$$

is never equal to the unknown parameter μ.

More precisely, we will never know if these two quantities coincide.

Still more precisely, the preceding statement is empty.

Frequentist perspective for the empirical mean

A large number N of realizations in equal conditions.

We cannot say how close to μ is a given realization $\overline{X}_n[1]$ of the sample empirical mean.

The *variance* of $\overline{X}_n$ is a measure of the dispersion of the N values $\overline{X}_n[i]$, meaning how likely is that a given $\overline{X}_n[1]$, the one actually observed, lies far apart of μ.

Back to the Quality Control problem (frequentist)

We take an n-sample, finding n_1 defective units.

The relative frequency $f_1 = \dfrac{n_1}{n}$ is "close" to the probability $\theta = \mathrm{P}\{X = 1\}$, (from the LLN).

The observed n_1 is (the value of) a r.v. , $n_1 \sim \mathrm{B}(n, \theta)$.

θ, the parameter, has a fixed value.

f_1 is an estimator of θ

The likelihood function (frequentist)

For a sample $\boldsymbol{x} = (x_1, \ldots x_n)$, where each x_i is a binary 0/1 value, assumed values of n i.i.d.$\sim$ Ber(θ) r.v. , the *likelihood function* is:

$$\mathscr{L}(\boldsymbol{x}\,;\theta) = \theta^{n_1} \cdot (1-\theta)^{n-n_1},$$

the result of multiplying n copies of the Bernoulli pmf.

In Bayesian statistics it has a slightly modified meaning and notation.

Maximum likelihood (ML) in frequentist statistics

Once $\boldsymbol{x}$ is known, $\mathscr{L}(\boldsymbol{x}\,;\theta)$ is a function of θ alone.

The ML estimate of θ is the "most likely" value, namely $\hat{\theta}$ such that, for the observed $\boldsymbol{x}$, $\mathscr{L}(\boldsymbol{x}\,;\hat{\theta})$ is maximum.

Problem is, θ is fixed. Statements about: "Probability of θ ..." , or similar, are meaningless.

Using "likelihood" as a synonym –sort-of– is a sheepish avowal of wrong-doing. This is the most blatant inconsistency in classical statistics logics

Classical (frequentist) hypotheses testing

Two hypotheses, the *null hypothesis* H_0, that the pdf of the observed data is a given f_0, *vs* another possible one, f_1, the *alternative hypothesis* H_1.

In a parametric model, $H_0 : \theta$ belongs to a given subset $\Theta_0 \subset \Theta$ *vs* $H_1 : \theta$ belongs to $\Theta \setminus \Theta_0$.

Decision mechanism

We find a (random, depending on the observed X) partition of the sample space $\mathscr{X}$, into two disjoint regions, $\mathscr{X}_0$, the *acceptance region,* and its complement $\mathscr{X}_1 = \mathscr{X} \setminus \mathscr{X}_0$.

When we observe $x \in \mathscr{X}_0$ then we accept H_0.

Significance level: Frequentist interpretation

A hypothesis test at a given *significance level,* $\alpha = 0.05$, say, is thought of as meaning that:

In a large number of repetitions under equal external conditions,

The proportion of wrong decisions rejecting H_0 when H_0 is actually true *(Type I Error)* will be around 5%.

Discussion of the frequentist paradigm

It is generally impossible to translate this *unreal statistics* into hard, checkable or falsifiable, experimental facts.

Often in real life a single realization of an experiment is possible and that's it.

04 - Bayesian paradigm

What is a *paradigm?*

What is a Statistical model?

Classical (frequentist) treatment

Bayesian perspective

First difference Classical/Bayesian

[C] θ is unknown, but it has some fixed value. We can elicit some (always uncertain) information about it from data.

[B] θ is unknown. We can describe its behavior as that of any other uncertain quantity, i.e., by means of a r.v.

Bayesian parameters

θ is unknown. Described by a r.v.

Two relevant pdf's:

- *Prior* pdf, $h(\theta)$
- *Posterior* pdf, given the observed data $\boldsymbol{x}$, $h(\theta|\boldsymbol{x})$.

Transition $h(\theta) \rightarrow h(\theta|\boldsymbol{x})$ by Bayes' rule for pdf's.

Bayes' rule for pdf's

The transformation:

$$h(\theta|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta) \cdot h(\theta)}{f(\boldsymbol{x})}.$$

$f(\boldsymbol{x})$ is the marginal pdf.

Bayesian likelihood

In Bayesian parlance, the *likelihood* function is a perfectly decent object:

It is just the conditional pdf of the r.v. $\boldsymbol{x}$, given a value of the r.v. θ.

In the QC example:

$$f(\boldsymbol{x}\,|\theta) \equiv \mathscr{L}(\boldsymbol{x}\,;\theta) = \theta^{n_1} \cdot (1-\theta)^{n-n_1}.$$

The prior pdf

Prior design is an essential part of Bayesian modelling.

Two main approaches:

- Reference priors: constructed using a set of objective rules.
- Subjective priors: trying to include whatever information is available about θ, or lack thereof.

Summary: The Bayesian paradigm

The Bayesian paradigm closely mimics the dynamic process of knowledge acquisition in the experimental sciences:

- We have an *initial* or *a priori* knowledge about a certain situation or phenomenon (a knowledge which can be null).
- We perform an experiment, resulting in some evidence X.
- We combine the information extracted from X with our previous knowledge, obtaining as a result an *a posteriori* o *final* description of the given situation.

Steps in implementing the Bayesian paradigm

1. Both observable quantities X and parameters θ are r.v. , the only constants being the actually observed quantities x.

2. Initial knowledge about θ is a *prior pdf or pmf* $p(\theta)$.

3. The *likelihood function* $p(x|\theta)$ is the pdf or pmf of X for a given parameter value.

4. Bayes' formula quantifies transition from prior to posterior.

5. Final knowledge on parameters is the *posterior pdf or pmf* $p(\theta|x)$, conditional to x.

A prior for the QC problem

Assuming no previous knowledge about the probability θ of a given unit being defective.

A possible *uninformative or vague prior* distribution is $\theta \sim \text{Unif}(0, 1)$,

$$h(\theta) = \begin{cases} 1, & \text{if } \theta \in (0, 1), \\ 0, & \text{otherwise.} \end{cases}$$

Applying Bayes rule

The numerator is the joint pdf of $\boldsymbol{x}$ and θ, the product of likelihood and prior:

$$h(\theta|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta) \cdot h(\theta)}{f(\boldsymbol{x})} = \frac{\theta^{n_1} \cdot (1-\theta)^{n-n_1}}{f(\boldsymbol{x})}.$$

The denominator is the integral, with respect to θ, of the numerator.

Marginal pdf of x

The denominator, the marginal pdf of $\boldsymbol{x}$, is:

$$f(\boldsymbol{x}) = \int_0^1 \theta^{n_1} \cdot (1-\theta)^{n-n_1}\, d\theta = B(n_1+1, n-n_1+1),$$

we know (do we?) this is a Beta function.

Posterior distribution of the parameter

Now our FIRST BAYESIAN RESULT:

The posterior pdf is:

$$h(\theta|\boldsymbol{x}) = \frac{1}{B(n_1+1, n-n_1+1)}\,\theta^{n_1}\cdot(1-\theta)^{n-n_1},$$

We recognize (do we?) a Beta$(n_1+1, n-n_1+1)$ pdf.

Recovering the classical ML estimate

The mode of this posterior distribution (the maximum of $h(\theta|\boldsymbol{x})$) is:

$$\frac{n_1}{n}.$$

Thus, in a sense, we recover or, rather, give a Bayesian justification for, the classical ML point estimation procedure.

Richer information

This hybrid Bayesian-frequentist result is called Maximum-A-Posteriori (MAP) estimation.

The whole posterior pdf is a much more thorough description of all the information about θ gathered from the observed data.