Forecasting Tools in Practical Applications: Selection and Evaluation Methodology

Shlomi Dolev[1], Sergey Frenkel[2], Julie Cwikel[1] and Victor Zakharov[2],

[1]Ben-Gurion University of the Negev, Beer-Sheva, Israel

[2]FRC "Computer Science and Control" Russian Academy of Sc., Moscow, Russia,

Goal

- **- to show how the specified properties of sequences and probability estimates affect the quality of the choice of predictors,**
- **- a rule for choosing a predictor based on the results of previous (potential) predictions is formulated**
-

Motivation

Designers of forecasting system, for example, a traffic management and/or network security system, is interested to asses expected quality and possible costs of forecasting at every required moment when the forecasting subsystem is running.

Example: tuning of acquired predictors (e.g., MS AWS Azure Machine Learning, Google Cloud Machine Learning etc.), parameters, which affects the total computational costs of the entire network management scheme.

So, the designer needs some characteristic of the predicted data, which determines the a priori difficulty of their prediction.

Contribution

- Conceptual model for assessing the predictability of a dataset, which makes it possible to separate, when assessing the possibility of using certain prediction tools, the internal (intrinsic) predictability of data, and "instrumental" predictability, which evaluates the predictability of data by prediction tools

- the properties of both the data sets and the predictors affecting the forecast by choosing predictor based on the performance of previous predictions are described.

- It is shown how certain properties (not necessarily statistical and probabilistic) of sequences can be used for a priori assessment of the prediction quality

Predictability

- Let $x_1, x_2, ..., x_t \ldots$ be the specified ordered set. The observer consistently observes the values $x_1, x_2, ..., x_{t-1}$ of known types *A*.
- At time t, having received the values $x_1, x_2, ..., x_{t-1}$, the observer predicts the next value x_t, that is, calculates according to some rule or algorithm *f* the value b_t (in the given alphabet), which will be received by the element of the sequence x_t.
- A random variable is *predictable* if its posterior probability distribution after the chosen (given) number of observations *differs significantly (in the sense of some measure of difference) from the prior probability distribution.*
- Predictability of a data set is the ability to predict the appearance at the next moment of a certain value of a random variable associated with this set, based on its previously known values.
- From a practical point of view, it depends both on the properties of the data itself and on the prediction algorithm used. *However, a set of data can also be attributed to its own predictability*

Mathematical aspects

- the conditional probabilities $\gamma(x_{t+1}|x_1, x_2, ..., x_t)$, defined on all words of a *certain sequence* $X=x_1, x_2, ..., x_t...$ *serve* to gain the best prediction, because they contain all information about the future behavior of the stochastic process

The natural conditions are satisfied $\sum_{a\in A} \gamma(x_{t+1}=a|x_1, x_2, ..., x_t) = 1$ and $\gamma(x_{t+1}|x_t) \geq 0$, for all $x_{t+1} \in A$, $t \geq 0$.

- **Example: Bernoulli sequence, Prob(1)= p_1 :**
- **$X=x1$, $\gamma(x_2=1|x_{1=}1)=\gamma(x_2=1|x_{1=}0)= p_1$, $\gamma(x_2=0|x_{1=}1)=\gamma(x_2=1|x_{1=}0)=1- p_1$**
- **$X= x_1, x_2, ..., x_t,$**
- **If $\sum_{1=1,10} x_i = k$, than $\gamma(x_{t+1}=1|x_1, x_2, ..., x_t)=k/t$**

- We know probability of correct prediction, IF we know the data probabilistic model!.

Math aspects: Universal Predictions

- Let the probabilistic model of the sequence $x_1, x_2, ..., x_t$ is not known.

Then, we need a method that:

whatever the actual data-generating probabilistic model, this method could work not much worse than one could if we knew as model at work.

That is, we should have a predictor independent of the probabilistic data model.

Then we should consider *the universal prediction*, in which predictions are without explicit probabilistic models of performed sequences, although with certain assumptions about the probabilistic measures of hypothetical sources generating these sequences, for example, their ergodicity!

Universal Prediction

- Let $\gamma(x_{t+1}|x_1 \ldots x_t)$ is unknown conditional probability distribution
- The discrepancy between the probability distribution $P(x_{t+1} = a| x_1 \ldots x_t)$ and the predictor $\gamma(a| x_1 \ldots x_t)$ for a given $x_1 \ldots \ldots x_t$ is given by the equality:

$$KL(P, \gamma, x_1,\ldots x_t) = \sum_{a \in A} P(x_{t+1} = a / x_1 \ldots x_t) \log(\gamma(x_{t+1} = a / x_1 \ldots x_t) / P(x_{t+1} = a / x_1 \ldots x_t))$$

KL is Kulback-Leibler divergence between a distribution P and its estimation.

- $P(a|x_1,\ldots,x_t)$ is the conditional distribution for the universal predictor.

Examples:

- Laplace predictor
- $\gamma(1/x_{t+1}|x_1,\ldots x_t)=(n_{x1,\ldots xt}+1)/(|t|+|A|))$
- where $n_{x1,\ldots xt}$ is the number of subsequences $x_1,\ldots x_t$
- *If A={0,1}:*
- $\gamma(1/x_{t+1}|x_1,\ldots x_t)=(n_1+1)/(M+2)$,
- $\gamma(0/x_{t+1}|x_1,\ldots x_t)=(n_0+1)/(M+2)$,
- n_1, n_0 are the number of ones and zeros in the sample of the size M.
- Laplace predictor is universal as its considers prediction as a set of estimations of unknown (conditional) probabilities, and the average error of the Laplace predictor (estimated by the KL divergence) goes to zero for any unknown i.i.d. source, when the sample size t grows.

- Lempel-Ziv(LZ) compressor as an Universal Predictor:
- x_{t+1} is the corresponding leaf in the partial matches tree with the conditional probability induced by the incremental parsing algorithm.

Internal Predictability

- Based on the concept of "universal predictor", we can construct a characterization of the predictability of a particular data sequence without using its probabilistic model. Let us call this characteristic of the internal predictability of an individual sequence "Intrinsic predictability", since it does not depend on any assumptions about the mathematical model of the source of the sequence, or on the predictor.
- *Internal predictability* of a sequence $x_1,x_2,\ldots,x_t$ in the moment t is:
- $$IP_t=D(P,P_M)$$
- $P=Prob(x_{t+1}=b_{t+1}|x_1,..,x_i,\ldots x_t)$, $x_i \in \{0,1\}^n$ is the probabilistic distribution of the source, generating this sequence. This is what we try to approximate using the universal prediction scheme, b_{t+1} is predicted by an universal predictor which is polynomial in complexity, e.g.. Laplace predictor,
- P_M is a “model” distribution, corresponding to a known model of random sequence, for example, to Bernoulli distribution,
- D is a measure of discrepancy between two distributions, e.g. Kulback-Leibler, or a metric, which allow to compare the compression rate by a LZ compression algorithm with the distribution of a partial matches (say, 010 in a sequence 00101110010100111010011).
-

Instrumental predictability

- **Instrumental predictability (InP) characterizes the minimum possible cost (loss) of misprediction that a given set of predictors can provide for a given sequence.**

$$InP = inf_f(lim_{n\to\infty} sup\ (L_f))$$

f is the set of available predictors, θ is some probability measure on the data predicted (X), over which averaging (E_θ) is performed (for example-Bernoulli (in this case θ is just the parameter of Bernoulli distribution)),

- L_f is average loss function, which, in the framework of the widely used decision theory

- **That is, based on the predicted b_t value, the user decides, the error of which may have a certain price.**
- **This is a characteristic of predictability on a given set of available predictors e.g., XGB, SGD, etc.**

Optimal predictor for Bernoulli sequences

- Let $x_1,..,x_i,\ldots x_{t-1}$ be the Bernoulli binary sequence with the parameter p_1.

Optimal according to the Hamming loss criterion ("incorrectly predicted the next bit-lost everything“) predictor of the Bernoulli sequence is :

If Bernoulli distribution parameter $p_1 \geq 1/2$, predicts the future value $x_{t+1} = 1$,

and $x_{t+1} = 0$ if $p_1 \leq 1/2$.

It is rejection of any forecast if Prob(0|xt)= Prob(1|x t)=1/2, that is the forecast is “skip”, where “skip” means the absence of an optimal forecast.

Application

- Let's x= x_1,..,x_n, be considered as "individual sequence", as existing in a single copy, for which nothing is known about (probabilistic or deterministic) laws of its generation.

- Suppose we use a predictor *p*, to predict the value of a binary time series at time *t*+1, knowing the predicted values at times *{t − M,...t}*, where *M>>m* is the window size, used for prediction at the moments within (t-M, t), m is the length of the predictor training sequence.
- Then, $N_p \backslash M$ is an empirical estimate of the conditional probability $\gamma()$.
- where N_p is the number of correct predictions by the predictor in the window of size *M*.
- If M is large enough, the success rate SR_p can be considered as Laplace estimator, that is an universal predictor, that is we may estimate the internal predictability of the sequence x in the moment t. IP_t .

As we use the predictor p to compute the $\gamma()$., we can, for example estimate if the sequence differs from the principally unpredictable Bernoulli sequence with p1=1/2 with the predictor, as well as the predictor Instrumental predictability.

- If the IP_t is close to randomness (Bernoulli with p1=1/2) and InP is close to Bernoulli optimal predictor, this predictor is rejected for this subsequence x, in order to test another.

- .

Implementation

- Let us apply to the sequence in the M-window the Bernoulli optimal predictor (BO) for prediction.

- We can talk about the difference between some predictor of a binary sequence p from such a simple predictor (BO), both in the calculated value success rate and in the structure of errors (i.e., in the error ratio “0 instead of 1” (“0→1”) and (“1 instead of 0” (“1→0”))

- If there is no significant difference, then we can talk about the inefficiency of the predictor f_p, since a similar solution is obtained by a much simpler Bernoulli predictor.

Implementation

- *If*
- $p_1 > ½$
- $p_1 <$ SR and $p_c(M, 0) > (1-p_1)/\tau$,

- $p_1 < ½$
  - $(1-p_1) <$ SR and $p_{V=1}(M, 0) > p_1/\tau$,

τ is between 2- 3

where $p_c(M_M, 0)$ is the probability of correct prediction of $x_t=0$.

then the predictor p is better then the Bernoulli optimal one, and can be used for the next time.

An Example

- Comparison XGB vs.SGD predictors for a section of size 100 of a sequence:
- 1111011010111011100010111011111111011111000100010001000110011 1100101101001011111011010011101001001101 1

- Result of work the XGB:
- 1111111111111111111111111111011111111111111111111111011111111110111 1111111111 1111111111111111111

- If we used the BO predictor, then we would predict all members of the considered section of the sequence as single ones, that is, the error would be 39%.

- The array of predicted values contains 97 ones and 3 zeros. Correctly predicted ones - 59, correctly predicted zeros - 1, incorrectly predicted ones - 2, incorrectly predicted zeros - 38.

- Both conditions of inequality (5) are not met. The predictor is not good, it gives worse results than the Bernoulli PB predictor.

Example Contd.

- Therefore, the SGD predictor was checked, it gave the following array of predicted values:
- 11111111101111111111111101111111111111101111101111111 01111111011111111111 11111100111111111111111111
-
- This array contains 92 ones and 8 zeros. Correctly predicted ones - 59, correctly predicted zeros - 6, incorrectly predicted ones - 2, incorrectly predicted zeros - 33. There are 65 correct predictions, 35 incorrect predictions. Choosing τ=3s, both conditions of formula (5) are fulfilled, the predictor obtained results that differ from the PB.

- Therefore, in this case, we can talk about high sequence predictability for the next time point for the SGD predictor, this predictor is preferable to the XGB predictor for predicting the next value.

Conclusion

- Conceptual data models and their consistency with prediction algorithms are most important aspects of characterizing of on-line prediction process .

- A promising direction of research is heterogeneous prediction models that include both stochastic and ontological models of systems and data.

Thank You!