Report on learning practice # 3

Sampling of multivariate random variables

Performed by:

Darya Murova

Kate Deviatiarova

Valeria Sakovskaya

Saint-Petersburg

2020

**Table of contents:**

1. Substantiation of chosen sampling.

2. Sampling of chosen target variables using univariate parametric distributions (from practice #2) with 2 different sampling methods.

3. Estimation of relations between predictors and chosen target variables.

4. Bayesian network

5. Quality analysis.

***Brief theoretical part***

***Task 2***

***Inverse transform sampling*** is a method for generating random numbers from any probability distribution by using its inverse cumulative distribution F−1(x). The cumulative distribution for a random variable X is  In what follows, we generate independent realizations of a random variable U uniformly distributed on

***Algorithm for continuous distributions:***

Assume we want to generate a random variable X with cumulative distribution function (CDF) FX. The inverse transform sampling algorithm is simple:

1. Generate   
2. Let -1X

Then, X will follow the distribution governed by the CDF FX, which was our desired result.

Note that, inverting FX is easy if X is an exponential random variable, but its harder if X is Normal random variable.

**Task 4-5**

Bayesian models are probabilistic models can define relationships between variables and be used to calculate probabilities.

A [probabilistic graphical model](https://en.wikipedia.org/wiki/Graphical_model) (PGM), or simply “*graphical model*” for short, is a way of representing a probabilistic model with a graph structure.

The nodes in the graph represent random variables and the edges that connect the nodes represent the relationships between the random variables.

A graph comprises nodes (also called vertices) connected by links (also known as edges or arcs). In a probabilistic graphical model, each node represents a random variable (or group of random variables), and the links express probabilistic relationships between these variables.

In the model we have:

* **Nodes**: Random variables in a graphical model.
* **Edges**: Relationships between random variables in a graphical model.

There are many different types of graphical models, although the two most commonly described are the Hidden Markov Model and the Bayesian Network.

The [Hidden Markov Model](https://en.wikipedia.org/wiki/Hidden_Markov_model) (HMM) is a graphical model where the edges of the graph are undirected, meaning the graph contains cycles. Bayesian Networks are more restrictive, where the edges of the graph are directed, meaning they can only be navigated in one direction. This means that cycles are not possible, and the structure can be more generally referred to as a directed acyclic graph (DAG).

Directed graphs are useful for expressing causal relationships between random variables, whereas undirected graphs are better suited to expressing soft constraints between random variables. Central to the Bayesian network is the notion of [conditional independence](https://en.wikipedia.org/wiki/Conditional_independence).

Sampling from a Bayesian Network:

* Sample nodes in some order so that by the time we sample a node we have values for all of its parents
* We can then sample from the distribution specified by the CPD
* Need ability to sample from the distributions underlying the CPD

Make forward sampling:

1. Split [0,1] interval into bins whose sizes are determined by the probabilities P(xi), i=1,..,k
2. Generate a sample s uniformly from the interval
3. If s is in the ith interval then sampled value is

Then we should compute expectations:

We can estimate it as using basic convergence bounds, we know that from a set of particles D ={ξ[1]),..,(ξ[M]} generated via sampling.

**Metrics for the model:**

**The K2 metric.** The probability metric derived by Cooper and Herskovits scores belief-network structures, thereby distinguishing among alternative networks, given a database of cases. Their result is summarized as theorem 1, and is based on the four assumptions that follow.

1. The database variables are discrete.
2. Cases occur independently, given a belief-network model.
3. There are no cases that have variables with missing values.
4. Before observing the database, we are indifferent regarding which numerical probabilities to assign to the belief-network structure.

**BIC score.** In the discrete case, the BIC score can only be negative.

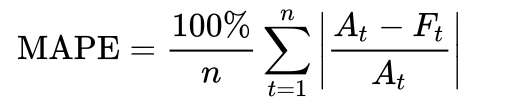
Text

Description automatically generated with medium confidence

Models with lower BIC are generally preferred

**Metrics for analysis and prediction’s quality:**

**The mean absolute percentage error (MAPE**), also known as mean absolute percentage deviation (MAPD), is a measure of prediction accuracy of a forecasting method in [statistics](https://en.wikipedia.org/wiki/Statistics" \o "Statistics). It usually expresses the accuracy as a ratio defined by the formula:



where  is the actual value and Ft is the forecast value. Their difference is divided by the actual value . The absolute value of this ratio is summed for every forecasted point in time and divided by the number of fitted points .

**:** whose main purpose is either the [prediction](https://en.wikipedia.org/wiki/Prediction" \l "Statistics" \o "Prediction) of future outcomes or the testing of [hypotheses](https://en.wikipedia.org/wiki/Hypotheses" \o "Hypotheses), on the basis of other related information. It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model.

The **RMSE** is the square root of the variance of the residuals and indicates the absolute fit of the model to the data (difference between observed data to model's predicted values). It can be interpreted as the standard deviation of the unexplained variance, and is in the same units as the response variable. Lower values indicate better model fit.

***Results***

**Task 1.**

Our target variables:

* Amount of precipitation in millimetres: prcp
* Solar radiation KJ/m2: gbrd
* Air temperature (instant) in celsius degrees: temp

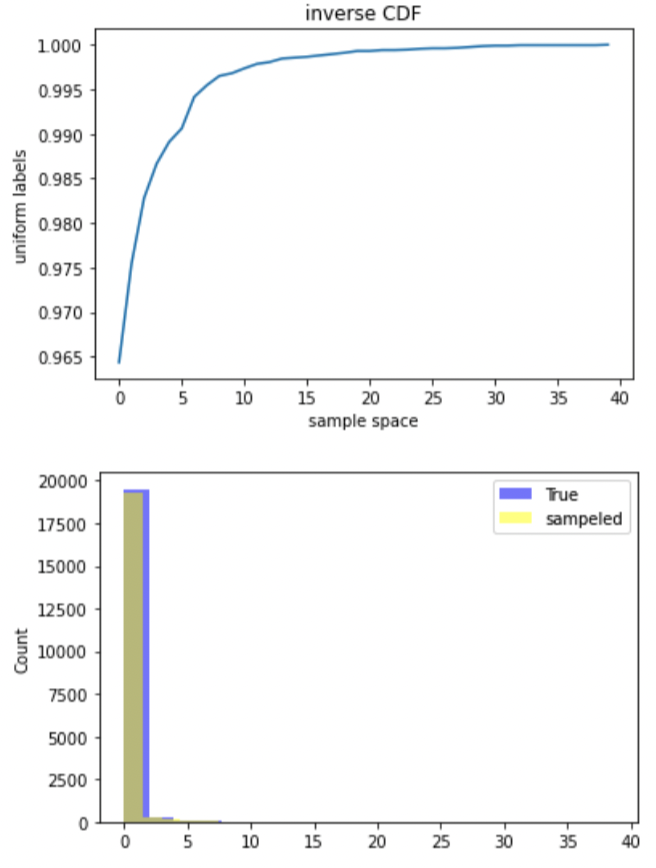
Predictors:

* Relative humid in % (instant)
* Latitude
* Wind gust in metres per second
* Wind speed in metres per second
* Atmospheric pressure at station level (mb)
* Wind direction in radius degrees

**Task 2.**

**Sampling of chosen target variables using univariate parametric distributions (from practice #2) with 2 different sampling methods.**

We’ve obtained the following results:

Chart, line chart

Description automatically generated Chart, histogram

Description automatically generated

*Figure 1. inverse CDF and comparison of sampled and real data for prcp, gbrd, temp respectively*

In order to estimate the quality of sampling, we decided we compare mean and standard deviation in sampled and in real data. Because we have different distributions, those are the only uniform metrics for the comparison.

*Table 1. Comparison of mean and standard deviation of real and sampled data (inverse-transform sampling method)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Mean real | Mean sample | Stdev real | Stdev sample |
| Amount of precipitation in millimetres (***prcp****)* | 0.1777855 | 0.617202 | 1.200110 | 1.056723 |
| Solar radiation KJ/m2 **(*gbrd)*** | 727.408054 | 733.061158 | 1048.117440 | 1037.919304 |
| Air temperature (instant) in Celsius degrees**(temp)** | 20.079315 | 20.031197 | 5.882493 | 5.909129 |

We can see that sampling worked nicely for ***temp,*** because it’s close to normal distribution. The rest have not very good results, but still close to the real.

Compare it to the random sampling method.

Here we’ve got the following results after sampling:

Chart, shape

Description automatically generated

In the table below we’ve highlighted green metrics which are better in random sampling method.

We can see that both methods worked 50/50. But inverse – transform method seems to be a bit better.

*Table 2. Comparison of mean and standard deviation of real and sampled data (random sampling method)*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Mean real | Mean sample | Stdev real | Stdev sample |
| Amount of precipitation in millimetres (***prcp****)* | 0.177785 | 0.182088 | 1.200110 | 1.196926 |
| Solar radiation KJ/m2 **(*gbrd)*** | 727.408054 | 721.069041 | 1048.117440 | 1040.86357 |
| Air temperature (instant) in Celsius degrees**(temp)** | 20.079315 | 20.115356 | 5.882493 | 5.8851358 |

**Task 3.**

**Estimate relations between predictors and chosen target variables. At least, they should have significant correlation coefficients.**

A picture containing chart

Description automatically generated

Figure 2. Correlation analysis of targets and predictors

For the analysis of relationship, we used Pierson correlation coefficient. Below you can see the results. And chosen connections for later Bayesian model building.

*Table 3. Analysis of significant correlation relations between predictors and chosen target variables.*

|  |  |  |
| --- | --- | --- |
| **Target** | **Predictor** | **Correlation** |
| Solar radiation KJ/m2 | Wind speed in meters per second **(wdsp)** | Medium positive |
|  | Relative humid in % (instant) **(hmdy)** | Significant negative |
|  | Wind gust in meters per second **(gust)** | Medium positive |
| Amount of precipitation in millimetres (prcp) | Relative humid in % (instant) **(hmdy)** | Medium positive |
|  | Wind gust in meters per second **(gust)** | Light positive |
| Air temperature (instant) in Celsius degrees(temp) | Relative humid in % (instant) **(hmdy)** | Strong negative |
|  | Latitude **(lat)** | Light positive |
|  | Wind gust in meters per second **(gust)** | Light positive |
|  | Atmospheric pressure at station level (mb) **(stp** | Light negative |

**Task 4. Build a Bayesian network for a chosen set of variables. Choose its structure on the basis of multivariate analysis and train distributions in nodes using the chosen algorithm.**

Based on the correlation analysis we built the following Bayesian model, using NetworkX package, train distributions in nodes using the chosen algorithm.

Diagram

Description automatically generated

Figure 3. Structure of nodes and links in the model based on correlation analysis

We got directed graph.

Chart, line chart

Description automatically generated

Figure 4. Bayesian Network for our chosen variables

**Task 5. Build a Bayesian network for the same set of variables but using 2 chosen algorithms for structural learning.**

In the discretization process (need to compute K2 and BIC) we chose number of bins based on the results in task 6. We’ve got lowest RMSE and MAPE and highest R^2 starting from 30 bins (In order to get scores, we need to transform the data back into the continuous).

For our model, based on correlation analysis we obtained:

|  |  |
| --- | --- |
| K2 score | -187521 |
| BIC | -191857 |

Good results as our aim to minimize those scores.

In order to get scores, we need to transform the data back into the continuous.

**Hill Climb search.** Hill climbing is a [mathematical optimization](https://en.wikipedia.org/wiki/Optimization_(mathematics)) technique which belongs to the family of [local search](https://en.wikipedia.org/wiki/Local_search_(optimization)). It is an [iterative algorithm](https://en.wikipedia.org/wiki/Iterative_algorithm) that starts with an arbitrary solution to a problem, then attempts to find a better solution by making an [incremental](https://en.wikipedia.org/wiki/Incremental_heuristic_search) change to the solution. If the change produces a better solution, another incremental change is made to the new solution, and so on until no further improvements can be found.

Based on K2 score, which equals -156259.83, we get the following structure:

Diagram

Description automatically generated with low confidence

Figure 5. K2 Hill Climb Structure

Also, we used tree search algorithm that finds structure that fits best to the given data set without parameterization. Based on K2 score (-163195.47) we obtained following structure:

A picture containing sky

Description automatically generated

Figure 6. K2. Tree Search structure

**Task 6. Analyze a quality of sampled target variables from the point of view of problem statement (e.g. prediction, gap filling, synthetic generation).**

Finally, to analyze a quality of sampled target variables we have chosen MAPE, RMSE, R^2 scores.

In the table below the results presented:

|  |  |  |  |
| --- | --- | --- | --- |
|  | R^2 | RMSE | MAPE |
| Manual model | 0.8099 | 1.548 | 1.4177 |
| Tree search | 0.57983 | 1.1077 | 1.1077 |
| Hill Climb | 0.740301 | 1.613 | 1.4815 |

**Conclusion**

All the models worked well, but if we compare RMSE and MAPE, we can see that tree search seems to be the best.

**Sourcecode**