PREDICTION OF PATHOGENIC SNV

Prof. Giorgio Valentini 6 CFU

Luca Cappelletti

Lecture Notes Year 2017/2018



IT Master Degree Universiy of Milan Italy 27 giugno 2018

Indice

I	Da	taset	4
1	Data	a points	5
	1.1	Retrieving the dataset	5
	1.2	Composition	5
		1.2.1 Training dataset	5
		1.2.2 Testing dataset	5
2	Met		6
	2.1	How the graphs are realized	
		2.1.1 Metric sample distribution	
		2.1.2 Plot graphs	6
		2.1.3 Normalized plot graphs	6
	2.2	-r r	7
		2.2.1 Metric sample distribution	7
		2.2.2 Metric values	7
	2.3	CpGperCpG	8
		2.3.1 Metric sample distribution	8
		2.3.2 Metric values	8
	2.4	CpGperGC	9
		2.4.1 Metric sample distribution	
		2.4.2 Metric values	
	2.5	DGVCount	
		2.5.1 Metric sample distribution	
		2.5.2 Metric values	
	26	DnaseClusteredHyp	
	2.0	2.6.1 Metric sample distribution	
		2.6.2 Metric values	
	2.7	DnaseClusteredScore	
	2.1	2.7.1 Metric sample distribution	
		2.7.2 Metric values	
	2.0	EncH3K27Ac	
	2.8		
		2.8.1 Metric sample distribution	
	0.0	2.8.2 Metric values	
	2.9	EncH3K4Me1	
		2.9.1 Metric sample distribution	
	0.10	2.9.2 Metric values	
	2.10	EncH3K4Me3	
		2.10.1 Metric sample distribution	
		2.10.2 Metric values	
	2.11	GCContent	
		2.11.1 Metric sample distribution	
		2.11.2 Metric values	
	2.12	GerpRS	
		2.12.1 Metric sample distribution	
		2.12.2 Metric values	17
	2.13	GerpRSpv	18
		2.13.1 Metric sample distribution	18

INDICE INDICE

	2.13.2 Metric values	
	2.14 ISCApath	
	2.14.1 Metric sample distribution	
	2.14.2 Metric values	
	2.15.1 Metric sample distribution	
	2.15.2 Metric values	
	2.16 dbVARCount	
	2.16.1 Metric sample distribution	
	2.16.2 Metric values	
	2.17 fantom5Perm	
	2.17.1 Metric sample distribution	 22
	2.17.2 Metric values	 22
	2.18 fantom5Robust	
	2.18.1 Metric sample distribution	
	2.18.2 Metric values	
	2.19 fracRareCommon	
	2.19.1 Metric sample distribution	
	2.19.2 Metric values	
	2.20 mamPhastCons46way	
	2.20.2 Metric values	
	2.20.2 Metric values	
	2.21.1 Metric sample distribution	
	2.21.2 Metric values	
	2.22 numTFBSConserved	
	2.22.1 Metric sample distribution	 27
	2.22.2 Metric values	 27
	2.23 priPhastCons46way	
	2.23.1 Metric sample distribution	
	2.23.2 Metric values	
	2.24 priPhyloP46way	
	2.24.1 Metric sample distribution	
	2.24.2 Metric values	
	2.25.1 Metric sample distribution	
	2.25.2 Metric values	
	2.26 verPhastCons46way	
	2.26.1 Metric sample distribution	
	2.26.2 Metric values	 31
	2.27 verPhyloP46way	
	2.27.1 Metric sample distribution	
	2.27.2 Metric values	 32
3	Metric distribution summary	33
Ŭ	, mound distribution summary	00
4	Scatter plot	34
	4.1 Scatter plot	
	4.2 Identified data correlations	
	4.2.1 dbVARCount and DGVCount	
	4.2.2 mamPhyloP46way and verPhyloP46way	 37
II	I Theory	39
_	Innut modelling	40
Э	Input modelling 5.1 Input values	40
	5.1.1 Normalized metric	
	5.1.2 Rarity	
	5.1.3 Entropy	
	5.2 Feet	
	5.3 Oversampling of positives	

 INDICE INDICE

	5.5	Undersampling of negatives	4		
6	Out	tput modelling	42		
7	7.1	ight initialization Weight distribution based on input distribution			
8	Loc 8.1	Activation function	4 4		
9	9.1 9.2	Activation function	45		
10	10 References 46				

Parte I

Dataset

Data points

First we begin looking at the dataset, the distributions of the given metrics and the statistical analysis of these data points.

1.1 Retrieving the dataset

The dataset can be downloaded from https://homes.di.unimi.it/valentini/ProgettoBioinformatica1718/data/.

1.2 Composition

1.2.1 Training dataset

In the training dataset there are 981389 data points, each one comprised of 26 metrics. The first 356 are pathogenic and all the others are negative.

1.2.2 Testing dataset

In the test dataset there are 190189 data points, still each one comprised of 26 metrics. The first 40 are pathogenic and the following are negative.

Metrics

2.1 How the graphs are realized

All the graphs are in triples: positives, negatives and mixed. All the zeros are removed as in most metrics *seemed* to indicate an unknown value.

2.1.1 Metric sample distribution

Are realized by calculating the frequencies and estimating the density distributions parameters via MLE.

2.1.2 Plot graphs

Are realized by sorting the values of the metric.

2.1.3 Normalized plot graphs

Are realized by sorting the values of the metric, with the domain and codomain normalized.

2.2. CPGOBSEXP CAPITOLO 2. METRICS

2.2 CpGobsExp

2.2.1 Metric sample distribution

The data points seem to follow a **Beta** distribution with the following parameters:

 $\alpha = 7.6689746880295795$ $\beta = 6778383.524935903$ loc = -0.09826818916997124 scale = 306278.3184506849

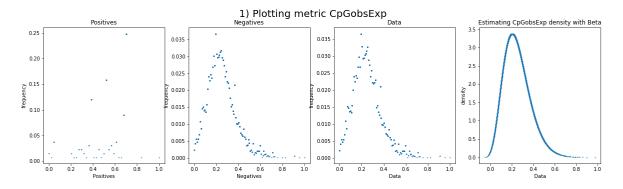


Figura 2.1: Sampling distribution of metric CpGobsExp

2.2.2 Metric values

1) Plotting metric CpGobsExp

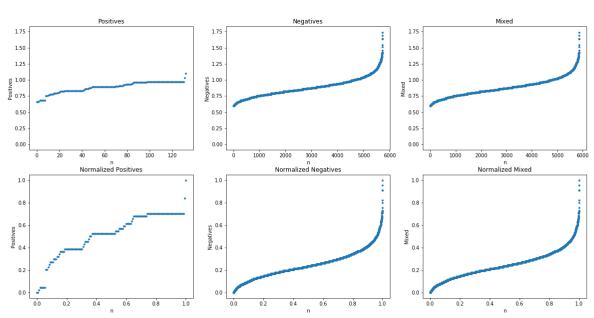


Figura 2.2: Values of metric CpGobsExp

2.3. CPGPERCPG CAPITOLO 2. METRICS

2.3 CpGperCpG

2.3.1 Metric sample distribution

The data points seem to follow a **Beta** distribution with the following parameters:

 $\alpha = 6.402175341881067 \qquad \beta = 97129163.31117742$ $loc = -0.05698922703576313 \qquad scale = 4337764.42876015$

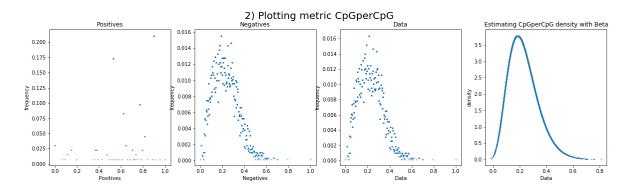


Figura 2.3: Sampling distribution of metric CpGperCpG

2.3.2 Metric values

2) Plotting metric CpGperCpG 35 30 Mixed 50 15 10 80 2000 3000 4000 5000 Normalized Negatives Normalized Positives 1.0 0.8 0.2

Figura 2.4: Values of metric CpGperCpG

2.4. CPGPERGC CAPITOLO 2. METRICS

2.4 CpGperGC

2.4.1 Metric sample distribution

The data points seem to follow a **Gaussian** distribution with the following parameters:

 $\mathbb{E}\left(X\right) = 0.4602356242601636 \qquad \text{Var}\left(X\right) = 0.15949294643574352$

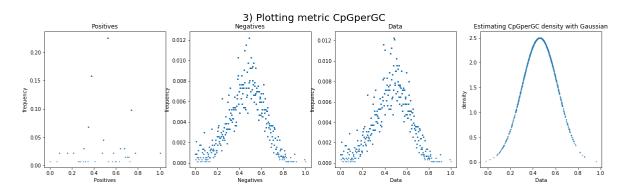


Figura 2.5: Sampling distribution of metric CpGperGC

2.4.2 Metric values

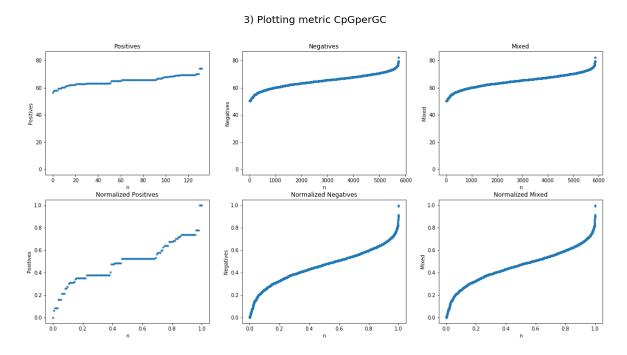


Figura 2.6: Values of metric CpGperGC

2.5. DGVCOUNT CAPITOLO 2. METRICS

2.5 DGVCount

2.5.1 Metric sample distribution

 $\alpha = 0.20940038672579409$

The data points seem to follow a **Gamma** distribution with the following parameters:

4) Plotting metric DGVCount Positives 0.35 0.35 0.30 0.25 0.25 0.20 ≥ 0.20 를 0.15 0.15 0.10 0.10 0.10 0.05 0.05 0.05

loc = -1.1962983066939984e - 30

scale = 1.2347090894162929

Figura 2.7: Sampling distribution of metric DGVCount

0.6

0.4 0.6 Negatives

2.5.2 Metric values

4) Plotting metric DGVCount

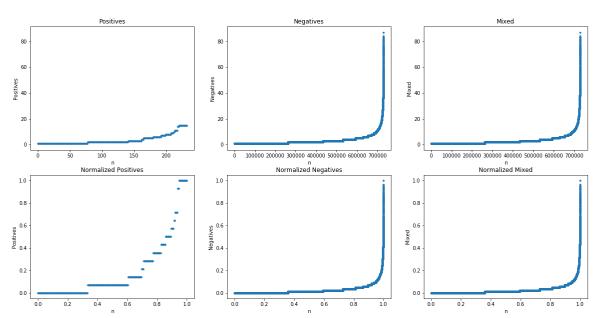


Figura 2.8: Values of metric DGVCount

2.6. DNASECLUSTEREDHYP CAPITOLO 2. METRICS

2.6 DnaseClusteredHyp

2.6.1 Metric sample distribution

The data points seem to follow a **Gamma** distribution with the following parameters:

 $\alpha = 0.4176887081406805 \qquad \text{loc} = -3.362626207862299 e - 29 \qquad \text{scale} = 0.3676310948709975$

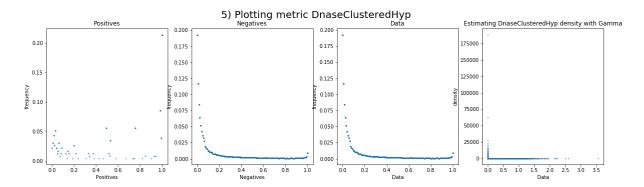


Figura 2.9: Sampling distribution of metric DnaseClusteredHyp

2.6.2 Metric values

5) Plotting metric DnaseClusteredHyp

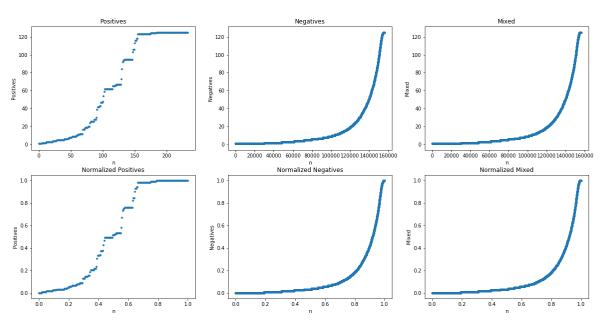


Figura 2.10: Values of metric DnaseClusteredHyp

2.7. DNASECLUSTEREDSCORE CAPITOLO 2. METRICS

2.7 DnaseClusteredScore

2.7.1 Metric sample distribution

The data points seem to follow slightly a Beta distribution with the following parameters:

 $\alpha = 0.2709657632937803 \qquad \beta = 0.44530002562349713$ $loc = -0.09309893086089688 \qquad scale = 1.0930989308608972$

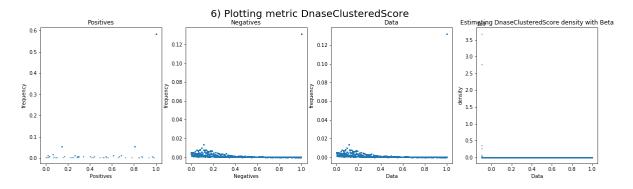


Figura 2.11: Sampling distribution of metric DnaseClusteredScore

2.7.2 Metric values

6) Plotting metric DnaseClusteredScore

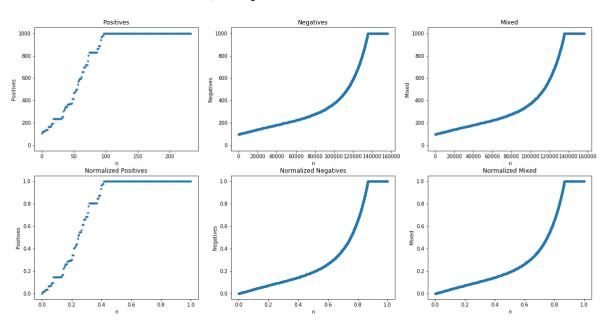


Figura 2.12: Values of metric DnaseClusteredScore

2.8. ENCH3K27AC CAPITOLO 2. METRICS

2.8 EncH3K27Ac

2.8.1 Metric sample distribution

The data points seem to follow a family of **Gamma** distributions (a speculation for this distribution could be the different groups from which the data are extracted), we will approximate them to one with a linear combination of the parameters:

 $\alpha = 0.0004042086221537893$ loc = -2.859398162696207e - 24 scale = 0.03076944787133299

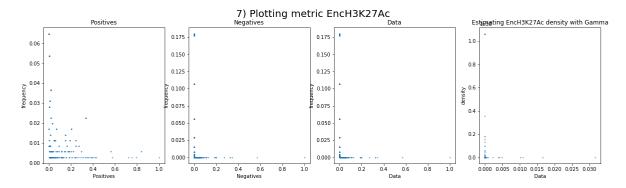


Figura 2.13: Sampling distribution of metric EncH3K27Ac

2.8.2 Metric values

7) Plotting metric EncH3K27Ac

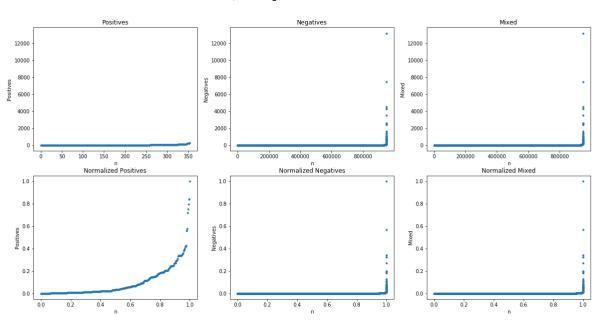


Figura 2.14: Values of metric EncH3K27Ac

2.9. ENCH3K4ME1 CAPITOLO 2. METRICS

2.9 EncH3K4Me1

2.9.1 Metric sample distribution

The data points seem to follow a family of **Gamma** distributions (a speculation for this distribution could be the different groups from which the data are extracted), we will approximate them to one with a linear combination of the parameters:

 $\alpha = 0.22566387737236238 \qquad \text{loc} = -6.619765504581537e - 27 \qquad \text{scale} = 1.396157055181753$

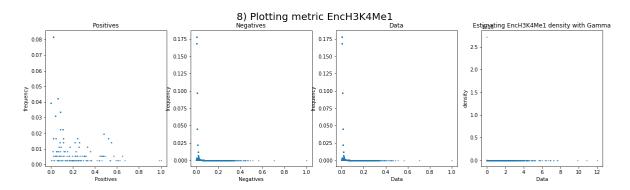


Figura 2.15: Sampling distribution of metric EncH3K4Me1

2.9.2 Metric values

8) Plotting metric EncH3K4Me1

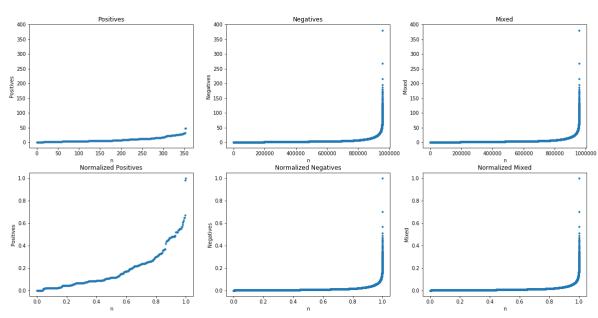


Figura 2.16: Values of metric EncH3K4Me1

2.10. ENCH3K4ME3 CAPITOLO 2. METRICS

2.10 EncH3K4Me3

2.10.1 Metric sample distribution

The data points seem to follow a family of **Gamma** distributions (a speculation for this distribution could be the different groups from which the data are extracted), we will approximate them to one with a linear combination of the parameters:

 $\alpha = 0.007502428717446465$ loc = -3.469650119186857e - 25 scale = 0.04125297431971783

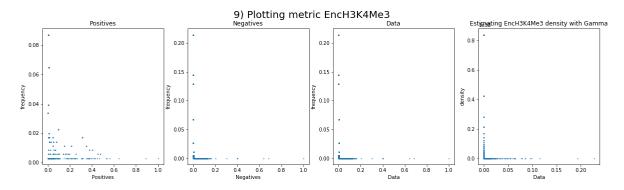


Figura 2.17: Sampling distribution of metric EncH3K4Me3

2.10.2 Metric values

9) Plotting metric EncH3K4Me3

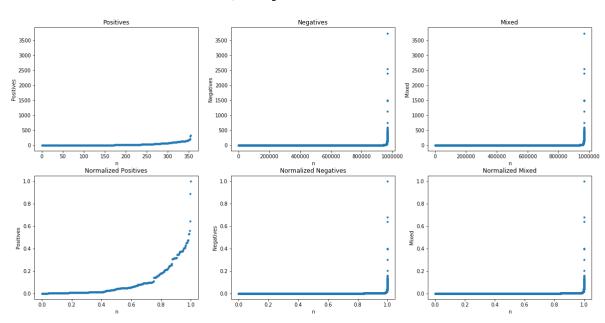


Figura 2.18: Values of metric EncH3K4Me3

2.11. GCCONTENT CAPITOLO 2. METRICS

2.11 GCContent

2.11.1 Metric sample distribution

The data points seem to be a combination of two **Gaussian** distributions. This will be approximated to one with the following parameters:

 $\mathbb{E}(X) = 0.4482813176478024$ Var(X) = 0.1097424869360011

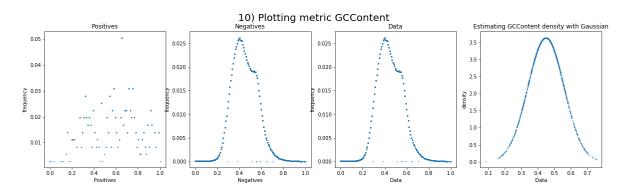


Figura 2.19: Sampling distribution of metric GCContent

2.11.2 Metric values

10) Plotting metric GCContent 0.8 0.2 0.0 0.0 0.0 400000 600000 800000 1000000 1000000 200 600000 Normalized Negatives Normalized Positives Normalized Mixed 1.0 0.8 0.2 0.2

Figura 2.20: Values of metric GCContent

2.12. GERPRS CAPITOLO 2. METRICS

2.12 GerpRS

2.12.1 Metric sample distribution

The data points seem to follow a family of **Gamma** distributions (a speculation for this distribution could be the different groups from which the data are extracted), we will approximate them to one with a linear combination of the parameters:

 $\alpha = 0.8688332877203315 \qquad \text{loc} = -1.7081810436826354e - 28 \qquad \text{scale} = 0.11512094125204281$

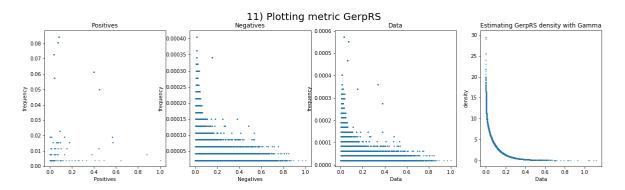


Figura 2.21: Sampling distribution of metric GerpRS

2.12.2 Metric values

11) Plotting metric GerpRS

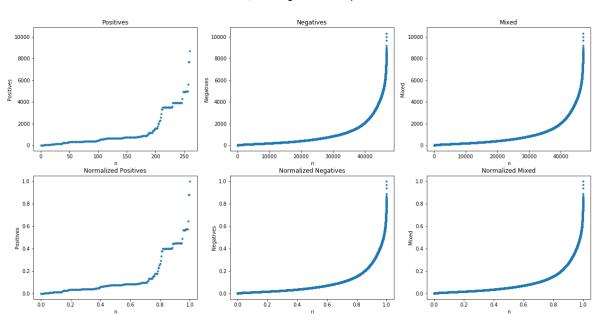


Figura 2.22: Values of metric GerpRS

2.13. GERPRSPV CAPITOLO 2. METRICS

2.13 GerpRSpv

2.13.1 Metric sample distribution

The data points seem to follow a family of **Gamma** distributions (a speculation for this distribution could be the different groups from which the data are extracted), we will approximate them to one with a linear combination of the parameters:

 $\alpha = 0.5165290213220888 \qquad \text{loc} = -6.952792177974854 \\ e - 30 \qquad \text{scale} = 0.2530358950266992$

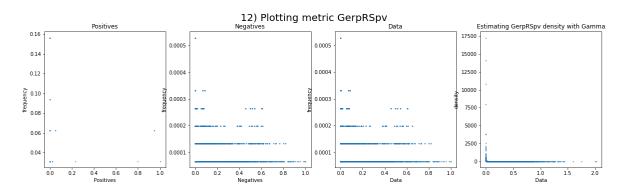


Figura 2.23: Sampling distribution of metric GerpRSpv

2.13.2 Metric values

12) Plotting metric GerpRSpv

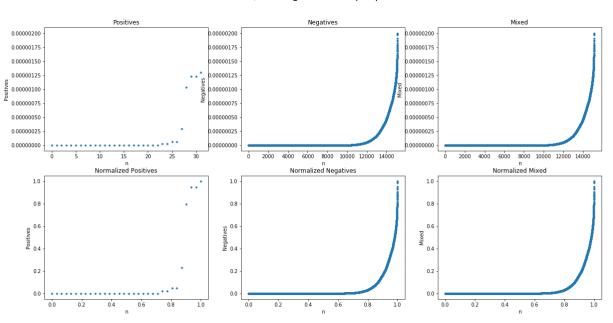


Figura 2.24: Values of metric GerpRSpv

2.14. ISCAPATH CAPITOLO 2. METRICS

2.14 ISCApath

2.14.1 Metric sample distribution

The data points seem to follow a **Gamma** distribution with the following parameters:

 $\alpha = 0.08318618903703257 \qquad \text{loc} = -1.9358902729364646e - 30 \qquad \text{scale} = 1.2606790181148981$

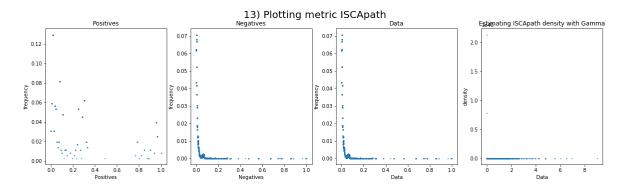


Figura 2.25: Sampling distribution of metric ISCApath

2.14.2 Metric values

13) Plotting metric ISCApath

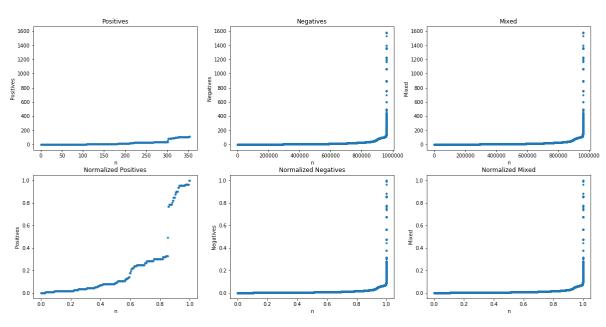


Figura 2.26: Values of metric ISCApath

2.15. COMMONVAR CAPITOLO 2. METRICS

2.15 commonVar

2.15.1 Metric sample distribution

The data points seem to follow an Exponential Weibull distribution with the following parameters:

 $\alpha = 5.038707296051438 \qquad \beta = 1.0160276119461702$ $loc = -0.012528678364149837 \qquad scale = 0.025052745155722922$

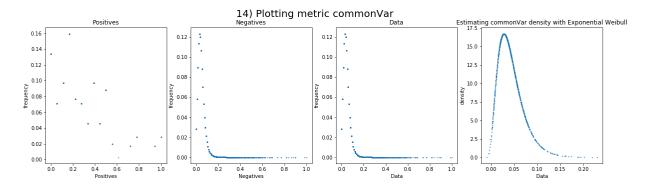


Figura 2.27: Sampling distribution of metric commonVar

2.15.2 Metric values

14) Plotting metric commonVar

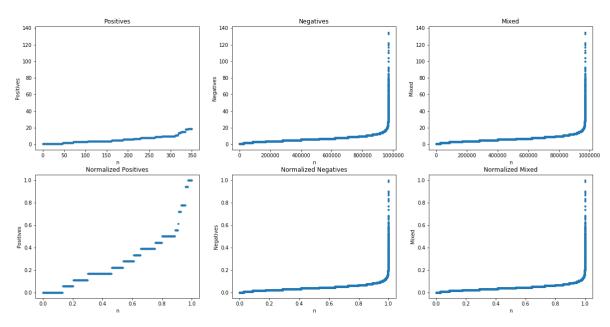


Figura 2.28: Values of metric commonVar

2.16. DBVARCOUNT CAPITOLO 2. METRICS

2.16 dbVARCount

2.16.1 Metric sample distribution

The data points seem to follow a **Gamma** distribution with the following parameters:

 $\alpha = 0.20940038672579409 \qquad \text{loc} = -1.1962983066939984e - 30 \qquad \text{scale} = 1.2347090894162929$

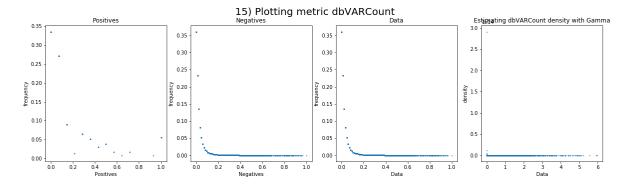


Figura 2.29: Sampling distribution of metric dbVARCount

2.16.2 Metric values

15) Plotting metric dbVARCount

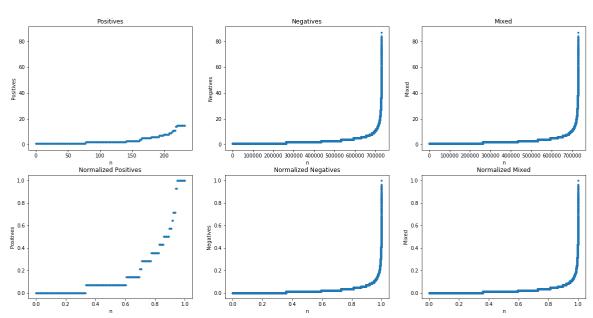


Figura 2.30: Values of metric dbVARCount

2.17. FANTOM5PERM CAPITOLO 2. METRICS

2.17 fantom5Perm

2.17.1 Metric sample distribution

The data points seem to follow a **Gamma** distribution with the following parameters:

 $\alpha = 0.06895533706017208$ loc = -3.220296247423778e - 30 scale = 1.2605014923175824

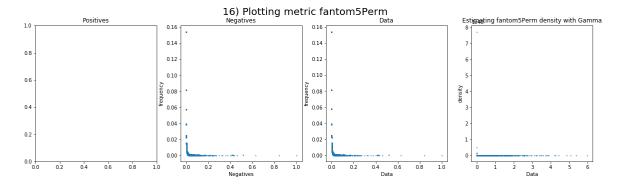


Figura 2.31: Sampling distribution of metric fantom5Perm

2.17.2 Metric values

16) Plotting metric fantom5Perm

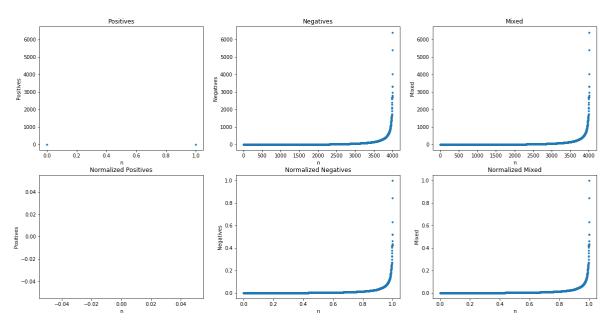


Figura 2.32: Values of metric fantom5Perm

2.18. FANTOM5ROBUST CAPITOLO 2. METRICS

2.18 fantom5Robust

2.18.1 Metric sample distribution

 $\alpha = 0.08983952110680529$

The data points seem to follow a **Gamma** distribution with the following parameters:

17) Plotting metric fantom5Robust Positives Estingating fantom5Robust density with Gamma 0.8 0.12 0.12 0.10 0.10 0.6 80.0 lledneucy 0.08 0.06 0.06 0.4 0.02 0.02

loc = -3.220296247423778e - 30

scale = 1.2605014923175824

Figura 2.33: Sampling distribution of metric fantom5Robust

2.18.2 Metric values

17) Plotting metric fantom5Robust

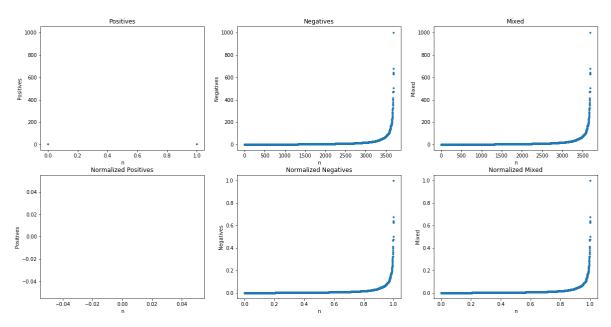


Figura 2.34: Values of metric fantom5Robust

2.19. FRACRARECOMMON CAPITOLO 2. METRICS

2.19 fracRareCommon

2.19.1 Metric sample distribution

The data points seem to follow an **Beta** distribution with the following parameters:

 $\alpha = 2772.739504773501 \qquad \beta = 14.986077009876375$ $loc = -69.93503912437342 \qquad scale = 71.09741090721741$

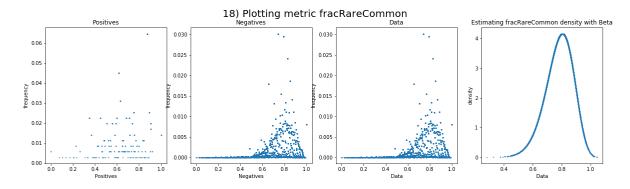


Figura 2.35: Sampling distribution of metric fracRareCommon

2.19.2 Metric values

18) Plotting metric fracRareCommon

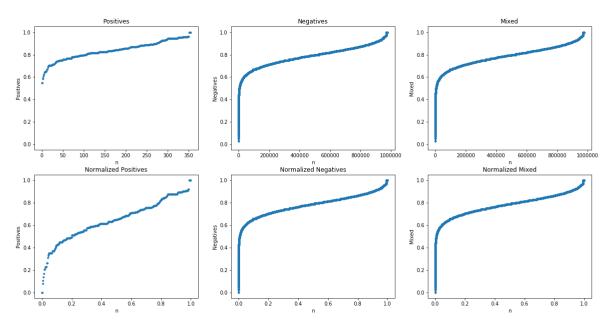


Figura 2.36: Values of metric fracRareCommon

scale = 0.45230902834164866

mamPhastCons46way 2.20

2.20.1 Metric sample distribution

 $\alpha = 0.3215099801387991$

The data points seem to follow a **Gamma** distribution with the following parameters:

19) Plotting metric mamPhastCons46way 0.15 0.8

loc = -6.260887365023215e - 31

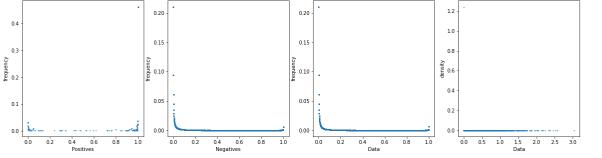


Figura 2.37: Sampling distribution of metric mamPhastCons46way

2.20.2 Metric values

19) Plotting metric mamPhastCons46way

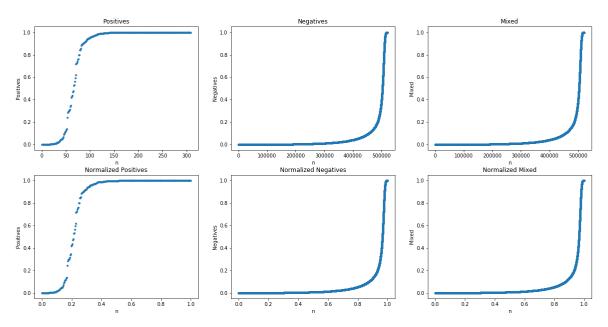


Figura 2.38: Values of metric mamPhastCons46way

2.21. MAMPHYLOP46WAY CAPITOLO 2. METRICS

2.21 mamPhyloP46way

2.21.1 Metric sample distribution

The data points seem to follow a **Gaussian** distribution with the following parameters:

 $\mathbb{E}(X) = 0.7032457913828309$ Var(X) = 0.07627203289198752

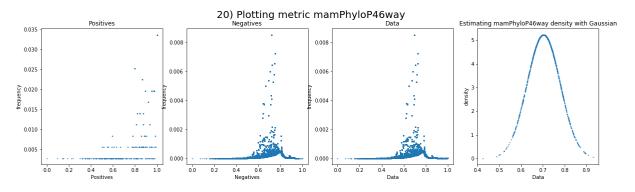


Figura 2.39: Sampling distribution of metric mamPhyloP46way

2.21.2 Metric values



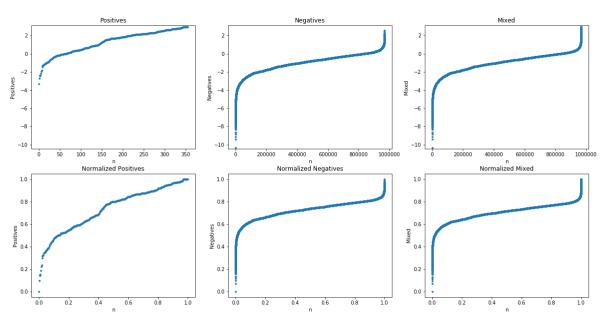


Figura 2.40: Values of metric mamPhyloP46way

2.22. NUMTFBSCONSERVED CAPITOLO 2. METRICS

2.22 numTFBSConserved

2.22.1 Metric sample distribution

The data points seem to follow a **exponential** distribution with the following parameters:

$$\mathbb{E}(X) = -4.600037873301623e - 12$$
 $Var(X) = 0.033419421646804975$

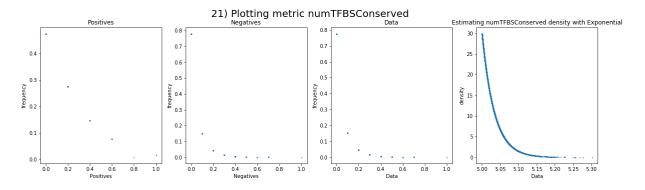


Figura 2.41: Sampling distribution of metric numTFBSConserved

2.22.2 Metric values

21) Plotting metric numTFBSConserved

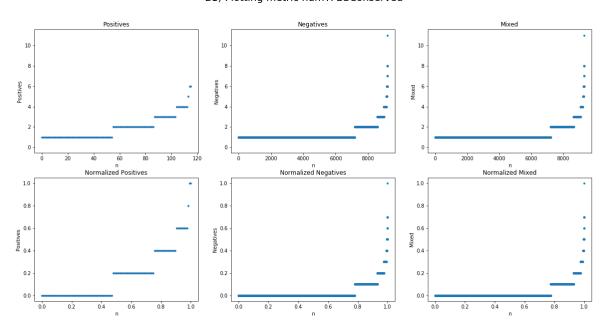


Figura 2.42: Values of metric numTFBSConserved

2.23. PRIPHASTCONS46WAY CAPITOLO 2. METRICS

2.23 priPhastCons46way

2.23.1 Metric sample distribution

The data points seem to follow a **Gamma** distribution with the following parameters:

 $\alpha = 0.2836383862597563 \qquad \text{loc} = -1.8643137904859329 \\ e - 31 \qquad \text{scale} = 0.37399746075497264$

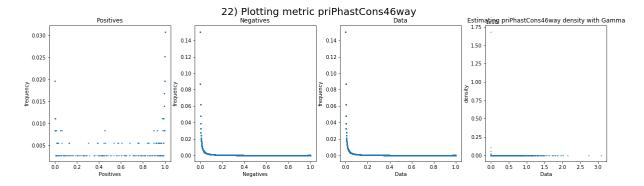


Figura 2.43: Sampling distribution of metric priPhastCons46way

2.23.2 Metric values

22) Plotting metric priPhastCons46way

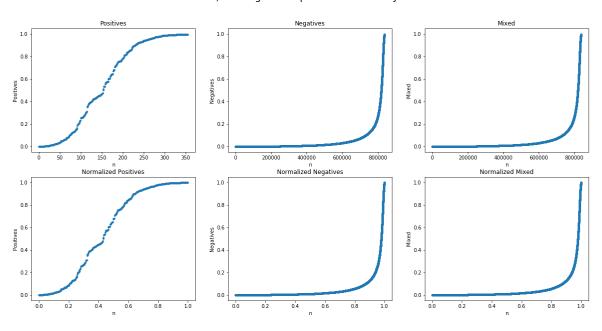


Figura 2.44: Values of metric priPhastCons46way

2.24. PRIPHYLOP46WAY CAPITOLO 2. METRICS

2.24 priPhyloP46way

2.24.1 Metric sample distribution

The data points seem to follow an **Beta** distribution with the following parameters:

 $\alpha = 2095270.7440875275 \qquad \beta = 4.199025269606416$ $loc = -103376.03746996864 \qquad scale = 103377.03863437689$

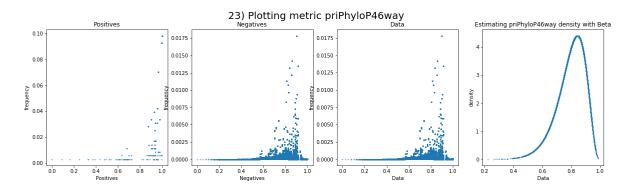


Figura 2.45: Sampling distribution of metric priPhyloP46way

2.24.2 Metric values

23) Plotting metric priPhyloP46way 150 200 250 400000 600000 800000 1000000 400000 600000 800000 1000000 n Normalized Negatives Normalized Positives Normalized Mixed 1.0 0.8 0.2

Figura 2.46: Values of metric priPhyloP46way

2.25. RAREVAR CAPITOLO 2. METRICS

2.25 rareVar

2.25.1 Metric sample distribution

The data points seem to follow an **Beta** distribution with the following parameters:

 $\alpha = 14.148202647100376$ $\beta = 7669045.025220526$ loc = -0.008523116473417407 scale = 28973.953544984728

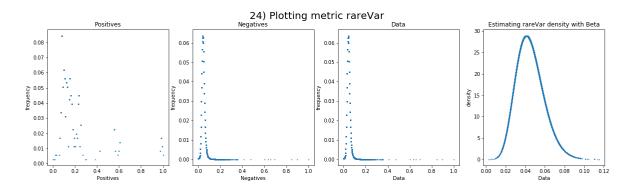


Figura 2.47: Sampling distribution of metric rareVar

2.25.2 Metric values

24) Plotting metric rareVar

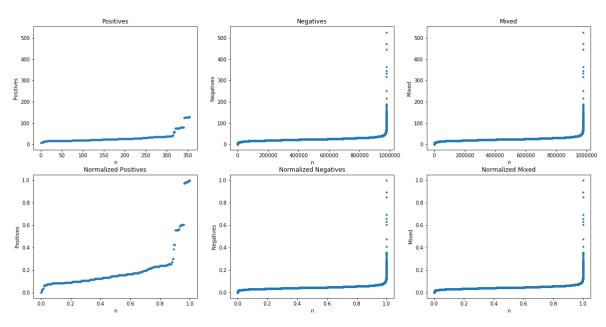


Figura 2.48: Values of metric rareVar

2.26. VERPHASTCONS46WAY CAPITOLO 2. METRICS

2.26 verPhastCons46way

2.26.1 Metric sample distribution

The data points seem to follow a **Gamma** distribution with the following parameters:

 $\alpha = 0.4378982063415524 \qquad \text{loc} = -2.5307968883256733 \\ e - 31 \qquad \text{scale} = 0.43138079305533483$

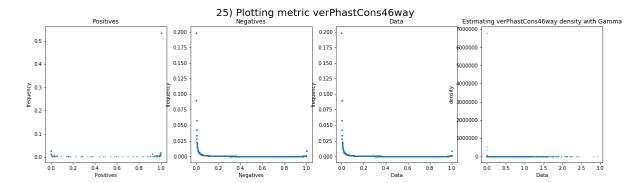


Figura 2.49: Sampling distribution of metric verPhastCons46way

2.26.2 Metric values

25) Plotting metric verPhastCons46way

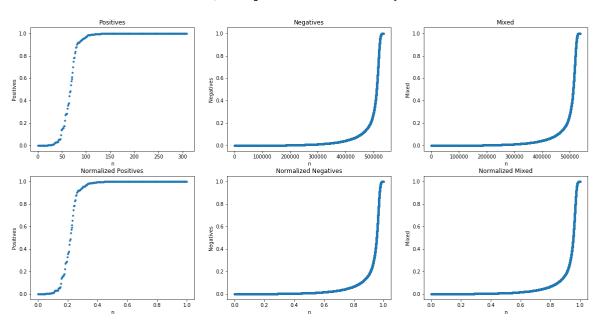


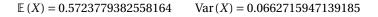
Figura 2.50: Values of metric verPhastCons46way

2.27. VERPHYLOP46WAY CAPITOLO 2. METRICS

2.27 verPhyloP46way

2.27.1 Metric sample distribution

The data points seem to follow a **Gaussian** distribution with the following parameters:



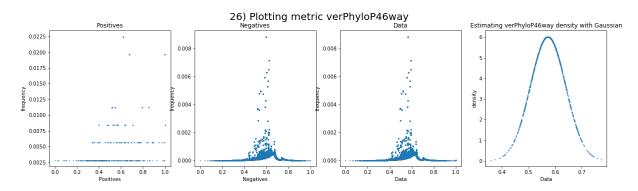


Figura 2.51: Sampling distribution of metric verPhyloP46way

2.27.2 Metric values

26) Plotting metric verPhyloP46way

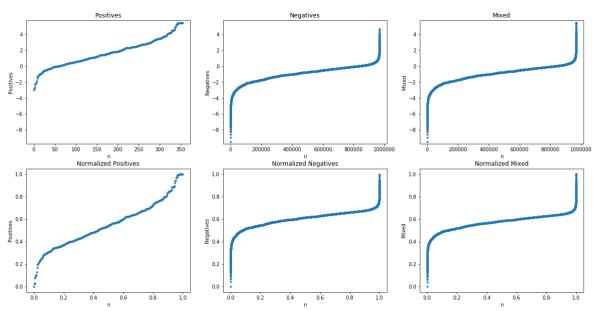


Figura 2.52: Values of metric verPhyloP46way

Metric distribution summary

The metrics seem to follow these sample distributions:

Metric	Distribution
CpGobsExp	Beta
CpGperCpG	Beta
CpGperGC	Gaussian
DGVCount	Gamma
DnaseClusteredHyp	Gamma
EncH3K27Ac	Gamma
GCContent	Gaussian
EncH3K4Me3	Gamma
ISCApath	Gamma
DnaseClusteredScore	Beta
EncH3K4Me1	Gamma
GerpRS	Gamma
GerpRSpv	Gamma
commonVar	Exponential Weibull
dbVARCount	Gamma
fantom5Perm	Gamma
fantom5Robust	Gamma
mamPhastCons46way	Gamma
priPhastCons46way	Gamma
rareVar	Beta
verPhastCons46way	Gamma
numTFBSConserved	Exponential
fracRareCommon	Beta
priPhyloP46way	Beta
verPhyloP46way	Gaussian
mamPhyloP46way	Gaussian

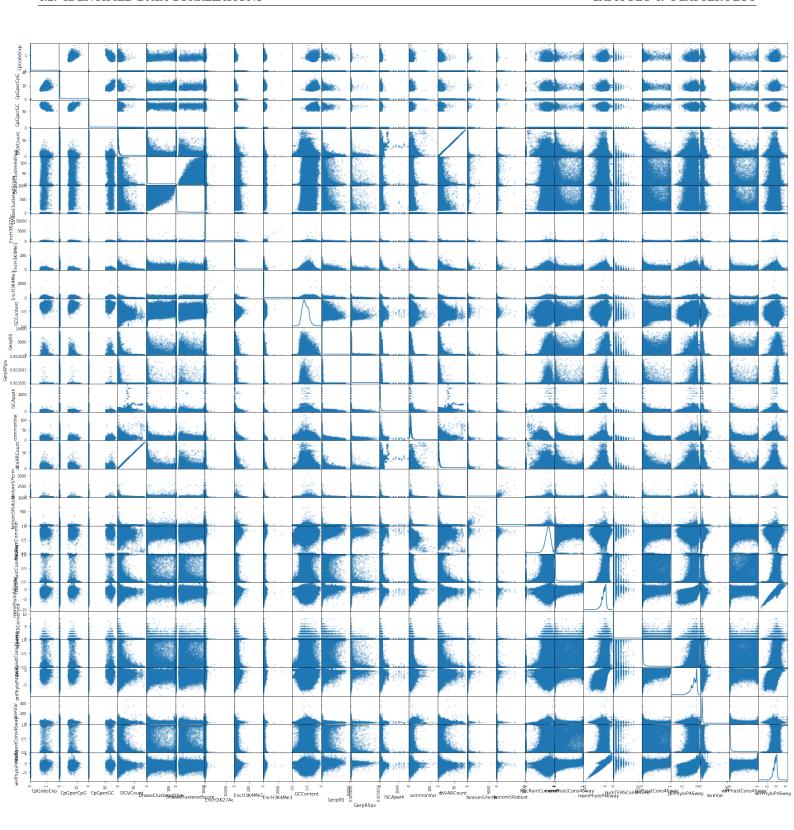
Tabella 3.1: Metrics and their distribution

Scatter plot

We now proceed to draw a scatter plot trying to identify eventual data correlations.

4.1 Scatter plot

A scatter plot with higher resolution is available in the project repository.



4.2 Identified data correlations

Data correlations seem to exist between:

4.2.1 dbVARCount and DGVCount

There is a strong correlation between this two metrics: let's look again at the data plots to see if they follow a similar trend:

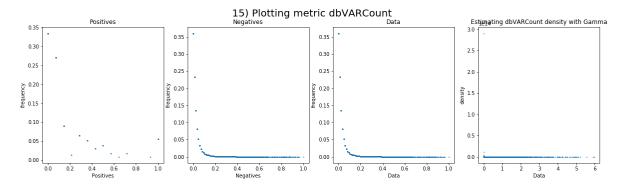


Figura 4.1: Sampling distribution of metric dbVARCount

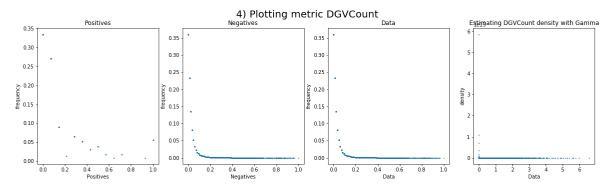


Figura 4.2: Sampling distribution of metric DGVCount

15) Plotting metric dbVARCount

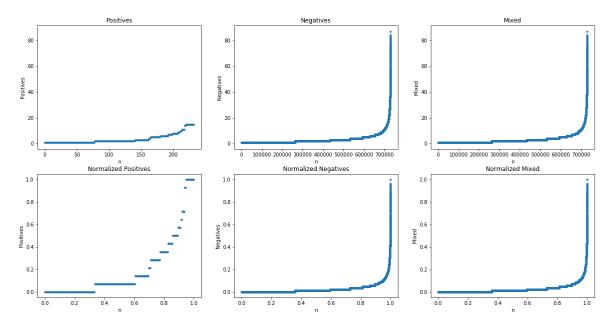


Figura 4.3: Values of metric dbVARCount

4) Plotting metric DGVCount

Figura 4.4: Values of metric DGVCount

The two metrics seem **highly** correlated, if not the **same metric**. This means that one of the two could be removed from the dataset, as it does not add any useful information.

4.2.2 mamPhyloP46way and verPhyloP46way

There is a some correlation between this two metrics: let's look again at the data plots to see if they follow a similar trend:

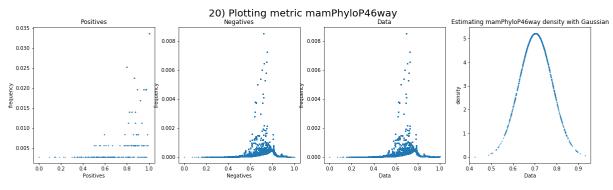


Figura 4.5: Sampling distribution of metric mamPhyloP46way

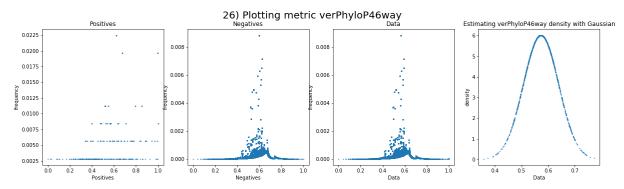


Figura 4.6: Sampling distribution of metric verPhyloP46way

Negatives -10 -10 -10 150 200 1.0 0.8 0.8 0.8 0.6 Nega 0.4 0.2 0.2 0.2 0.0 1.0

20) Plotting metric mamPhyloP46way

Figura 4.7: Values of metric mamPhyloP46way

26) Plotting metric verPhyloP46way

Figura 4.8: Values of metric verPhyloP46way

The two metrics seem **slightly** correlated, but not enough to consider removing one of the two.

Parte II

Theory

5.1 Input values

The values used for each metric are the 3 following:

5.1.1 Normalized metric

Clearly one of the important metrics is the metric itself, that will be normalized to allow for input in [0,1] range:

$$metric' = \frac{metric - min\{metric\ values\}}{max\{metric\ values\} - min\{metric\ values\}}$$

Figura 5.1: Input normalization to [0,1] range

5.1.2 Rarity

Another value we will be using in the input layer of the network is the rarity of the metric value, modelled as the surprise value of the estimated sampling distribution of the metric:

If M is the estimated metric distribution cumulative distribution function (CDF), m is the value assumed by the metric in the given data point and ϵ is a range, we can model **rarity** as follows:

$$\mathbb{P}\left(m-\epsilon \leq X \leq m+\epsilon\right) = M(m+\epsilon) - M(m-\epsilon) \qquad \text{rarity}(m) = 1 - \mathbb{P}\left(m-\epsilon \leq X \leq m+\epsilon\right)$$

Figura 5.2: Rarity

5.1.3 Entropy

The third and final value used will be entropy, obtained using the estimated metric probability:

$$H(x) = -\mathbb{P}(m - \epsilon \le X \le m + \epsilon) \log \mathbb{P}(m - \epsilon \le X \le m + \epsilon)$$

Figura 5.3: Entropy

5.2 Feet

The input layer is comprised of 25 (number of metrics, excluded the one recognized to be in strong correlation to another) *feet*, meaning tiny networks that are used to limit the initial linear combination of the metric input values to themselves.

Each feet is modelled as a locally connected dense layer, with a window of 3 neurons.

5.3 Oversampling of positives

Since the positive values are just the 0.036% of the dataset we'll oversample these to weight more these values. Since the variance of positive data points is too high to extrapolate a distribution to generate significant new fuzzy data points, simple duplication will be used.

5.4 Undersampling of negatives

Since the negative values are more than the 99.96% of the dataset we'll undersample these to weight less these values.

5.5 Oversampling and undersampling targets

Oversampling and undersampling target will be to have a training dataset with 1% of positives and 99% of negatives.

5.6 Absence of information

Absence of information about a given metric will be modelled as **zeros**, meaning all values relative to the given absent metric for that data point will be treated as zero.

6

Output modelling

The output layer of the neural network is modelled by **two** neurons, one representing the positive class and one the negative class, with a **sigmoid** as activation function.

Weight initialization

7.1 Weight distribution based on input distribution

Since input values are not from any particular distribution or hold properties such as $\mathbb{E}(X) = 0$ or Var(X) = 1 (in some metrics mean and variance are far from these values) they do not suggest to use any specific distribution.

7.2 Weight distribution based on activation functions and regularization layers

The codomain values from the activation functions, being SELU for most of the network, tend to hold the properties of $\mathbb{E}(X) = 0$ and Var(X) = 1 (http://arxiv.org/abs/1706.02515). These values are then regularized to penalize extreme weights that may appear when variance starts with a value significantly away from 1.

For these properties weight will be initialized by extracting them from a Gaussian with $\mathbb{E}(X) = 0$ and Var(X) = 1.

Locally connected dense layers

The first two layers will be locally connected dense layers, to exploit the positional information of the input values.

Other than the group of triples, input will be sorted by distribution kind so that the initial interpolations happen mostly with data from the same distribution family.

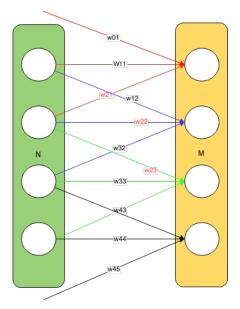


Figura 8.1: Locally connected layer

8.1 Activation function

We'll be using a **leaky relu** for the locally connected dense layers with $\alpha = 0.3$.

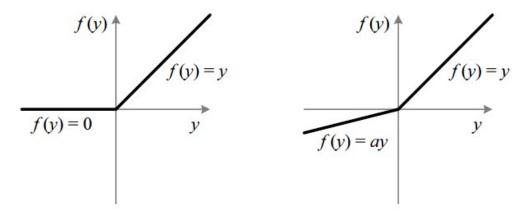


Figura 8.2: RELU and Leaky RELU

For the following hidden layers we will be using dense connected layers, with a piramidal structure (reducing the number of the neurons from 26 to 2).

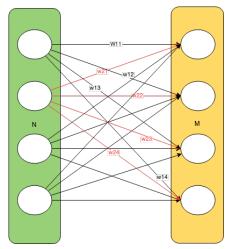


Figura 9.1: Dense connected layer

9.1 Activation function

We'll be using again leaky relu and experimenting with SELU for the dense layers:

$$selu(x) = \lambda \begin{cases} x & x > 0 \\ \alpha e^x - \alpha & x \le 0 \end{cases}$$

Figura 9.2: SELU

9.2 Regularization

Regularization layers will be alternated to the dense layers to penalize weight extreme growth.

9.3 Drop out

In addition to regularization, also **drop out** of 10% of neurons per hidden layer will be applied.

10 References

LatexTools does not compile references at this time.