GCRF for Classification, Fast Approximation, and Applications

Andrija Petrovic

University of Belgrade

April 17, 2019

GCRF

Gaussian Conditional Random Fields for regression

- Discriminative model

Advantages of GCRF:

- Combination of models and spatio-temporal correlation
- Additional information provided by structure
- Learning coefficients $\boldsymbol{\alpha}$,$\boldsymbol{\beta}$
  not correlation matrix $\boldsymbol{\Sigma}$ and expectation $\boldsymbol{\mu}$

GCRF

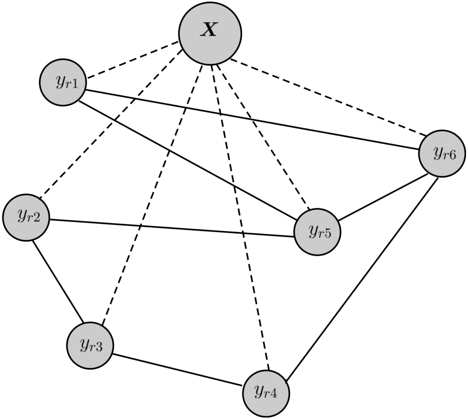

Figure: Graphical representation of dependencies expressed by GCRF

GCRF

- The generalized form of the GCRF is:

$$P(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})=\frac{1}{Z(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})}\exp\Big(\sum_{i=1}^{N} A(\boldsymbol{\alpha},y_i,\boldsymbol{x_i})+\sum_{i\neq j} I(\boldsymbol{\beta},y_i,y_j)\Big) \quad (1)$$

- Two different feature functions are used: *association potential* $A(\alpha, y_i, \boldsymbol{x})$ to model relations between outputs y_i and corresponding input vector $\boldsymbol{x_i}$ and *interaction potential* $I(\beta, y_i, y_j)$ to model pairwise relations between nodes.

GCRF

- The association potential is defined as:

$$A(\boldsymbol{\alpha}, y_i, \mathbf{x_i}) = -\sum_{k=1}^{K} \alpha_k \big(y_i - R_k(\mathbf{x_i})\big)^2 \tag{2}$$

- The interaction potential functions are defined as:

$$I(\boldsymbol{\beta}, y_i, y_j) = -\sum_{l=1}^{L} \sum_{k=1}^{K} \beta_l S_{ij}^{l} (y_i - y_j)^2 \tag{3}$$

The canonical form of GCRF is:

$$P(\mathbf{y}|\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{(2\pi)^{\frac{N}{2}} |\Sigma|^{\frac{1}{2}}} \exp\Big(-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{y} - \boldsymbol{\mu})\Big) \tag{4}$$

GCRFCB

One way of adapting GCRF to classification problem is by approximating discrete outputs by suitably defining continuous outputs. Namely, GCRF can provide dependence structure over continuous variables which can be passed through sigmoid function.

- The model is applicable to classification problems with undirected graphs, intractable for standard classification CRFs.
- Defining correlations directly between discrete outputs may introduce unnecessary noise to the model.
- In case that unstructured predictors are unreliable, which is signaled by their large variance (diagonal elements in the covariance matrix), it is simple to marginalize over latent variable space and obtain better results.

Representation

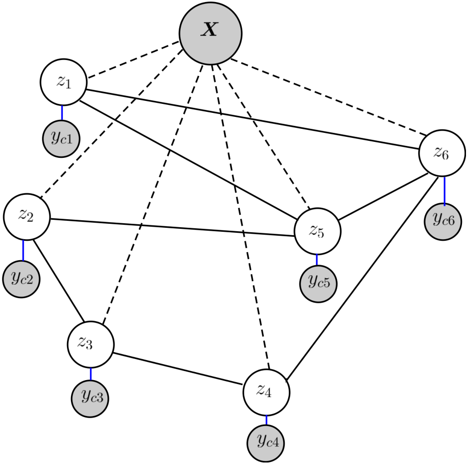

Figure: Graphical representation of GCRFCB

Representation

- The conditional probability distribution $P(y_i|z_i)$ is defined as Bernoulli distribution:

$$P(y_i|z_i) = Ber(y_i|\sigma(z_i)) = \sigma(z_i)^{y_i}(1-\sigma(z_i))^{1-y_i} \quad (5)$$

- The joint distribution of outputs y_i can be expressed as:

$$P(y_1, y_2, \ldots, y_N|\mathbf{z}) = \prod_{i=1}^{N} \sigma(z_i)^{y_i}(1-\sigma(z_i))^{1-y_i} \quad (6)$$

- The conditional distribution $P(\mathbf{z}|\mathbf{x})$ is the same as in the classical GCRF model and has canonical form defined by multivariate Gaussian distribution.

Representation

- The joint distribution of continuous latent variables $\boldsymbol{z}$ and outputs $\boldsymbol{y}$ is:

$$P(\boldsymbol{y},\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})=\prod_{i=1}^{N}\sigma(z_i)^{y_i}(1-\sigma(z_i))^{1-y_i}\cdot\frac{1}{(2\pi)^{N/2}|\Sigma(\boldsymbol{x},\boldsymbol{\theta})|^{1/2}} \cdot\exp\Big(-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{\mu}(\boldsymbol{x},\boldsymbol{\theta}))^T\Sigma(\boldsymbol{x},\boldsymbol{\theta})^{-1}(\boldsymbol{z}-\boldsymbol{\mu}(\boldsymbol{x},\boldsymbol{\theta}))\Big) \tag{7}$$

where $\boldsymbol{\theta}=(\alpha_1,\ldots,\alpha_K,\beta_1,\ldots,\beta_L)$.

Two ways of inference and learning were considered in GCRFBC model:

1. GCRFBCb - with conditional probability distribution $P(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta})$, in which variables $\boldsymbol{z}$ are marginalized over, and
2. GCRFBCnb - with conditional probability distribution $P\left(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{\theta},\mu_{\boldsymbol{z}}\right)$, in which variables $\boldsymbol{z}$ are substituted by their expectations.

- Due to conditional independence between nodes, it is possible to obtain $P(y_i = 1|\boldsymbol{x}, \boldsymbol{\theta})$.

$$P(y_i|\boldsymbol{x},\boldsymbol{\theta})=\int_{\boldsymbol{z}} P(y_i|\boldsymbol{z})P(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{z} \tag{8}$$

$$P(y_i=1|\boldsymbol{x},\boldsymbol{\theta})=\int_{\boldsymbol{z}} \sigma(z_i)P(\boldsymbol{z}|\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{z} \tag{9}$$

- As a result of independence properties of the distribution, it holds $P(y_i = 1|\boldsymbol{z}) = P(y_i = 1|z_i)$, and it is possible to marginalize $P(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta})$ with respect to latent variables $\boldsymbol{z}' = (z_1, \ldots, z_{i-1}, z_{i+1}, \ldots, z_N)$:

$$P(y_i=1|\boldsymbol{x},\boldsymbol{\theta})=\int_{z_i} \sigma(z_i)\Big(\int_{\boldsymbol{z}'} P(\boldsymbol{z}',z_i|\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{z}'\Big)dz_i \tag{10}$$

- It holds:

$$P(y_i=1|\boldsymbol{x},\boldsymbol{\theta})=\int_{-\infty}^{+\infty} \sigma(z_i)\mathcal{N}(z_i|\mu_i,\sigma_i^2)dz_i \tag{11}$$

- To predict $\boldsymbol{y}$, it is necessary to evaluate posterior maximum of latent variable $\boldsymbol{z}_{\max} = \underset{\boldsymbol{z}}{\text{argmax}} P(\boldsymbol{z}|\boldsymbol{x}, \boldsymbol{\theta})$, which is straightforward due to normal form of GCRF. Therefore, it holds $\boldsymbol{z}_{\max} = \boldsymbol{\mu_{z,i}}$. The conditional distribution $P(y_i = 1|\boldsymbol{x}, \boldsymbol{\mu_{z,i}}, \boldsymbol{\theta})$ can be expressed as:

$$P(y_i=1|\boldsymbol{x},\boldsymbol{\mu_z},\boldsymbol{\theta})=\sigma(\mu_{z,i})=\frac{1}{1+\exp(-\mu_{z,i})} \tag{12}$$

where $\mu_{z,i}$ is expectation of latent variable z_i.

- Evaluation of the conditional log likelihood is intractable, since latent variables cannot be analytically marginalized. The conditional log likelihood is expressed as:

$$\mathcal{L}(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\theta})=\log\left(\int_{\boldsymbol{Z}} P(\boldsymbol{Y},\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta})d\boldsymbol{Z}\right)=\sum_{j=1}^{M}\log\left(\int_{\boldsymbol{z_j}} P(\boldsymbol{y_j},\boldsymbol{z_j}|\boldsymbol{x_j},\boldsymbol{\theta})d\boldsymbol{z_j}\right)$$
$$=\sum_{j=1}^{M}\mathcal{L}_j(\boldsymbol{y_j}|\boldsymbol{x_j},\boldsymbol{\theta}) \tag{13}$$

$$\mathcal{L}_j(\boldsymbol{y_j}|\boldsymbol{x_j},\boldsymbol{\theta})=\log\int_{\boldsymbol{z_j}}\prod_{i=1}^{N}\sigma(z_{ji})^{y_{ji}}(1-\sigma(z_{ji}))^{1-y_{ji}}\frac{\exp(-\frac{1}{2}(\boldsymbol{z_j}-\boldsymbol{\mu_j})^T\Sigma_j^{-1}(\boldsymbol{z_j}-\boldsymbol{\mu_j}))}{(2\pi)^{N/2}\left|\Sigma_j\right|^{1/2}}d\boldsymbol{z_j} \tag{14}$$

- One way to approximate integral in conditional log likelihood is by local variational approximation. Lower bound for sigmoid function, can be expressed as:

$$\sigma(x) \geqslant \sigma(\xi) \exp\{(x-\xi)/2 - \lambda(\xi)(x^2-\xi^2)\} \tag{15}$$

where $\lambda(\xi) = -\frac{1}{2\xi} \cdot [\sigma(\xi) - \frac{1}{2}]$ and ξ is a variational parameter.

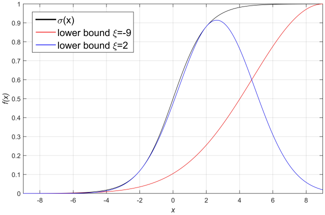

Figure: The sigmoid function with its lower bound

- This approximation can be applied to the model, such that:

$$P(\mathbf{y_j},\mathbf{z_j}|\mathbf{x_j},\boldsymbol{\theta})=P(\mathbf{y_j}|\mathbf{z_j})P(\mathbf{z_j}|\mathbf{x_j},\boldsymbol{\theta})\geq\underline{P}(\mathbf{y_j},\mathbf{z_j}|\mathbf{x_j},\boldsymbol{\theta},\boldsymbol{\xi_j}) \tag{16}$$

$$\underline{P}(\mathbf{y_j},\mathbf{z_j}|\mathbf{x_j},\boldsymbol{\theta},\boldsymbol{\xi_j})=\prod_{i=1}^{N}\sigma(\xi_{ji})\exp\left(z_{ji}y_{ji}-\frac{z_{ji}+\xi_{ji}}{2}-\lambda(\xi_{ji})(z_{ji}^2-\xi_{ji}^2)\right)\cdot \frac{1}{(2\pi)^{N/2}\left|\Sigma_j\right|^{1/2}}\exp\left(-\frac{1}{2}(\mathbf{z_j}-\boldsymbol{\mu_j})^T\Sigma_j^{-1}(\mathbf{z_j}-\boldsymbol{\mu_j})\right) \tag{17}$$

- The lower bound of conditional log likelihood $\underline{\mathcal{L}}(\mathbf{y_j}|\mathbf{x_j},\boldsymbol{\theta},\boldsymbol{\xi_j})$ is defined as:

$$\underline{\mathcal{L}_j}(\mathbf{y_j}|\mathbf{x_j},\boldsymbol{\theta},\boldsymbol{\xi_j})=\log\underline{P}(\mathbf{y_j}|\mathbf{x_j},\boldsymbol{\theta},\boldsymbol{\xi_j})=\sum_{i=1}^{N}\left(\log\sigma(\xi_{ji})-\frac{\xi_{ji}}{2}+\lambda(\xi_{ji})\xi_{ji}^2\right)-\frac{1}{2}\boldsymbol{\mu_j}^T\Sigma_j^{-1}\boldsymbol{\mu_j}+\frac{1}{2}\mathbf{m_j}^TS_j^{-1}\mathbf{m_j}+\frac{1}{2}\log|S_j| \tag{18}$$

- Learning in GCRFBCnb model is simpler compared to the GCRFBCb algorithm, because instead of marginalization, the mode of posterior distribution of continuous latent variable $\boldsymbol{z}$ is evaluated directly, so there is no need for approximation technique. The conditional log likelihood can be expressed as:

$$\begin{aligned}\mathcal{L}\big(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{\mu}\big) &= \log P(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{\theta},\boldsymbol{\mu}) \\ &= \textstyle\sum_{j=1}^{M}\sum_{i=1}^{N}\log P(y_{ji}|\mathbf{x_j},\boldsymbol{\theta},\mu_{ji}) \\ &= \textstyle\sum_{j=1}^{M}\sum_{i=1}^{N}\mathcal{L}_{ji}(y_{ji}|\mathbf{x_j},\boldsymbol{\theta},\mu_{ji})\end{aligned} \tag{19}$$

$$\mathcal{L}_{ji}(y_{ji}|\mathbf{x_j},\boldsymbol{\theta},\mu_{ji}) = y_{ji}\log\sigma(\mu_{ji}) + (1-y_{ji})\log\big(1-\sigma(\mu_{ji})\big) \tag{20}$$

GCRFBCb - Fast approximation

- Due to large number of variational parameters it is necessary to decrease memory and computational cost.
- However, in the case of GCRFBCb, memory complexity during training is $O(M)$ due to dependency of variational parameters on the number of instances. Computational complexity is also higher – $O(TMN^3)$, which can also be reduced to $O(TMN^2)$ in case of sparse precision matrix.
- Decreasing costs by decreasing number of variational parameters

- Partial derivatives of conditional log likelihood with respect to ξ_{ji} are:

$$\frac{\partial \mathcal{L}_j(\boldsymbol{y_j}|\boldsymbol{x_j},\boldsymbol{\theta},\boldsymbol{\xi_j})}{\partial \xi_{ji}} = -\frac{1}{2}\mathrm{Tr}\left(2S_j \frac{\partial \Lambda_j}{\partial \xi_{ji}}\right) - \left[2\left(\boldsymbol{y_j} - \frac{1}{2}\boldsymbol{I}\right)S_j \frac{\partial \Lambda_j}{\partial \xi_{ji}} S_j\right] S_j^{-1}\boldsymbol{m_j} + \boldsymbol{m_j}^T \frac{\partial \Lambda_j}{\partial \xi_{ji}} m_j + \sum_{i=1}^{N}\left(\left(\frac{1}{\sigma(\xi_{ji})} + \frac{1}{2}\xi_{ji}\right)\frac{\partial \sigma(\xi_{ji})}{\partial \xi_{ji}} + \frac{1}{2}\left(\sigma(\xi_{ji}) - \frac{3}{4}\right)\right) \quad (21)$$

- One way to solve this it to cluster ξ_{ji} and use group representative as approximation to all variational parameters in group

- In each iteration of optimization for each instance and node expectation of $P(\boldsymbol{z_j}|\boldsymbol{x_j},\boldsymbol{\theta})$ is evaluated for current value of $\boldsymbol{\theta}$
- The obtained μ_{ji} or $\sigma(\mu_{ji})$ are clustered
- Group representative is used for gradient evaluation

Results

YEAST					
No.	Model	Cluster	Cluster No.	AUC	Calculation time in sec
1	GCRFCNB	-	-	0.793	6.574
2	GCRFCB	-	-	0.797	146.527
3	GCRFCB_fast	K-means	5	0.793	81.332
4	GCRFCB_fast	K-means	50	0.793	76.412
5	GCRFCB_fast	K-means	150	0.794	126.419
6	GCRFCB_fast	K-means	250	0.794	182.050
7	GCRFCB_fast	MiniBatch K-means	5	0.793	86.667
8	GCRFCB_fast	MiniBatch K-means	50	0.794	87.942
9	GCRFCB_fast	MiniBatch K-means	150	0.794	130.509
10	GCRFCB_fast	MiniBatch K-means	250	0.793	139.425
11	GCRFCB_fast	Gaussian Mix	5	0.794	86.542
12	GCRFCB_fast	Gaussian Mix	50	0.793	90.905
13	GCRFCB_fast	Gaussian Mix	150	0.794	122.325
14	GCRFCB_fast	Gaussian Mix	250	0.794	125.200
15	GCRFCB_fast	Gaussian Mix prob	5	0.793	87.269
16	GCRFCB_fast	Gaussian Mix prob	50	0.793	157.059
17	GCRFCB_fast	Gaussian Mix prob	150	0.793	358.438
18	GCRFCB_fast	Gaussian Mix prob	250	0.794	891.755
19	Log_regre_L2	-	-	0.581	-
20	Log_regre_L1	-	-	0.581	-
21	NN	-	-	0.579	-
22	RF	-	-	0.608	-

Figure: Gene functional classification

Results

SCENE					
No.	Model	Cluster	Cluster No.	AUC	Calculation time in sec
1	GCRFCNB	-	-	0.943	5.283
2	GCRFCB	-	-	0.928	82.995
3	GCRFCB_fast	K-means	5	0.892	60.832
4	GCRFCB_fast	K-means	50	0.893	83.825
5	GCRFCB_fast	K-means	500	0.872	145.776
6	GCRFCB_fast	K-means	1200	0.767	273.152
7	GCRFCB_fast	MiniBatch K-means	5	0.894	65.224
8	GCRFCB_fast	MiniBatch K-means	50	0.891	75.232
9	GCRFCB_fast	MiniBatch K-means	500	0.801	135.654
10	GCRFCB_fast	MiniBatch K-means	1200	0.761	259.348
11	GCRFCB_fast	Gaussian Mix	5	0.890	69.725
12	GCRFCB_fast	Gaussian Mix	50	0.892	93.138
13	GCRFCB_fast	Gaussian Mix	500	0.910	369.583
14	GCRFCB_fast	Gaussian Mix	1200	0.907	783.624
15	GCRFCB_fast	Gaussian Mix prob	5	0.890	63.357
16	GCRFCB_fast	Gaussian Mix prob	50	0.911	104.404
17	GCRFCB_fast	Gaussian Mix prob	500	0.910	1145.447
18	GCRFCB_fast	Gaussian Mix prob	1200	0.892	2552.967
19	Log_regre_L2	-	-	0.898	-
20	Log_regre_L1	-	-	0.917	-
21	NN	-	-	0.930	-
22	RF	-	-	0.934	-

Figure: Semantic scene classification

Results

Pediatric					
	Model	Cluster	Cluster No.	AUC	Calculation time in sec
1	GCRFCNB	-	-	0.820	36.767
2	GCRFCB	-	-	0.834	851.645
3	GCRFCB_fast	K-means	5	0.832	392.249
4	GCRFCB_fast	K-means	50	0.831	675.800
5	GCRFCB_fast	K-means	500	0.802	1395.052
7	GCRFCB_fast	MiniBatch K-means	5	0.831	546.011
8	GCRFCB_fast	MiniBatch K-means	50	0.831	541.520
9	GCRFCB_fast	MiniBatch K-means	500	0.827	734.008
11	GCRFCB_fast	Gaussian Mix	5	0.831	481.539
12	GCRFCB_fast	Gaussian Mix	50	0.831	537.168
13	GCRFCB_fast	Gaussian Mix	500	0.831	1283.313
15	GCRFCB_fast	Gaussian Mix prob	5	0.830	553.819
16	GCRFCB_fast	Gaussian Mix prob	50	0.831	783.791
17	GCRFCB_fast	Gaussian Mix prob	500	0.832	3124.255
19	Log_regre_L2	-	-	0.782	-
20	Log_regre_L1	-	-	0.797	-
21	NN	-	-	0.783	-
22	RF	-	-	0.811	-

Figure: Pediatric readmission

GCRF-GCRFBC

- If a dataset is sparse, GCRF cannot be directly applied since it cannot handle missing values and handling them in a naive way (e.g., substituting them by 0 in case of average vehicle speed) leads to poor performance.
- The learning procedure is straightforward. Firstly, all instances and corresponding nodes are used for learning parameters of binary classification algorithm (GCRFBCb or GCRFBCnb). Subsequently, for each instance only nodes with non-null values of outputs are used in maximization of GCRF log likelihood with respect to the parameters α and β.
- It is important to emphasize that structure of GCRF during learning and inference is changing with respect to the nodes with non-null values of variables.

GCRF-GCRFBC

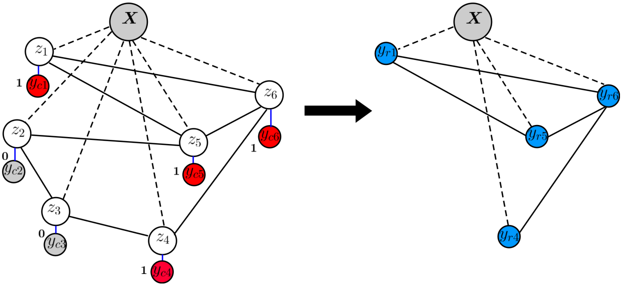

Figure: Graphical representation of classification-regression methodology expressed by GCRFBC and GCRF models

Results
Highway TSE

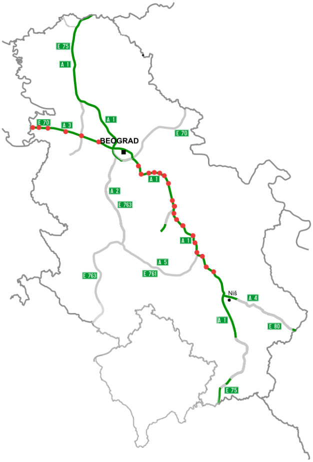

Figure: Map of highways in Serbia - red marked toll stations were used as nodes in GCRFBC model

Table: Prediction performance and total computation time of GCRF and unstructured predictors for average moving velocity when null and non null values are not classified

		Niš - Belgrade		Belgrade - Adaševci	
No.	Model	R^2	Computation time in minutes	R^2	Computation time in minutes
1	GCRF	**0.509**	227.134	**0.947**	19.272
2	Ridge	0.449	5.148	0.928	0.391
3	Lasso	0.452	3.204	0.928	0.328
4	NN	-0.060	25.179	0.891	8.310
5	RF	0.507	184.720	0.946	7.378

Table: Prediction performance and total computation time of GCRFBC and GCRF for classification of null versus non null values and regression of average moving velocity

Classification of null values					
		Niš - Belgrade		Belgrade - Adaševci	
No.	Model	AUC/R^2	Computation time in minutes	AUC/R^2	Computation time in minutes
1	GCRFBCnb	**0.893**	262.799	0.648	49.518
2	GCRFBCb	0.878	465.683	**0.999**	89.949
4	Ridge	0.853	0.216	0.995	0.015
5	Lasso	0.854	1.842	0.995	0.218
6	NN	0.832	249.933	0.997	47.181
7	RF	0.858	4.437	0.998	0.581
Regression of average moving velocity					
1	GCRF	**0.859**	175.458	**0.983**	14.136
2	Ridge	0.826	3.891	0.970	0.295
3	Lasso	0.828	2.421	0.971	0.247
4	NN	0.796	19.031	0.961	6.271
5	RF	0.834	139.622	0.980	5.568

Conclusion

- Structured classification and regression algorithms have better prediction performance compared to the unstructured predictors, meaning that the variable dependence structure can be exploited for better prediction
- Both GCRFBCb and GCRFBCnb models have better prediction performance compared to the unstructured predictors
- Due to high memory and computational complexity of GCRFBCb compared to GCRFBCnb, it is reasonable to use GCRFBCnb or fast GCRFBCb.

References

- Jaakkola, Tommi S., and Michael I. Jordan. "Bayesian parameter estimation via variational methods.Statistics and Computing 10.1 (2000): 25-37.
- Petrović, Andrija, et al. "Gaussian Conditional Random Fields for Classification." arXiv preprint arXiv:1902.00045 (2019).