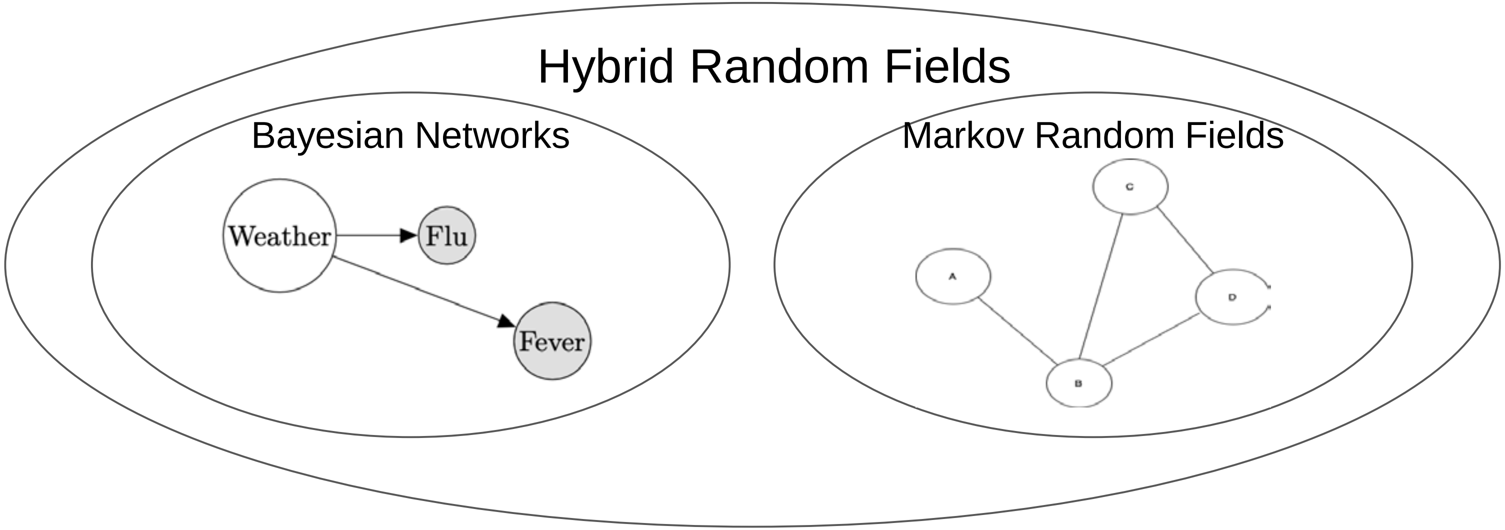
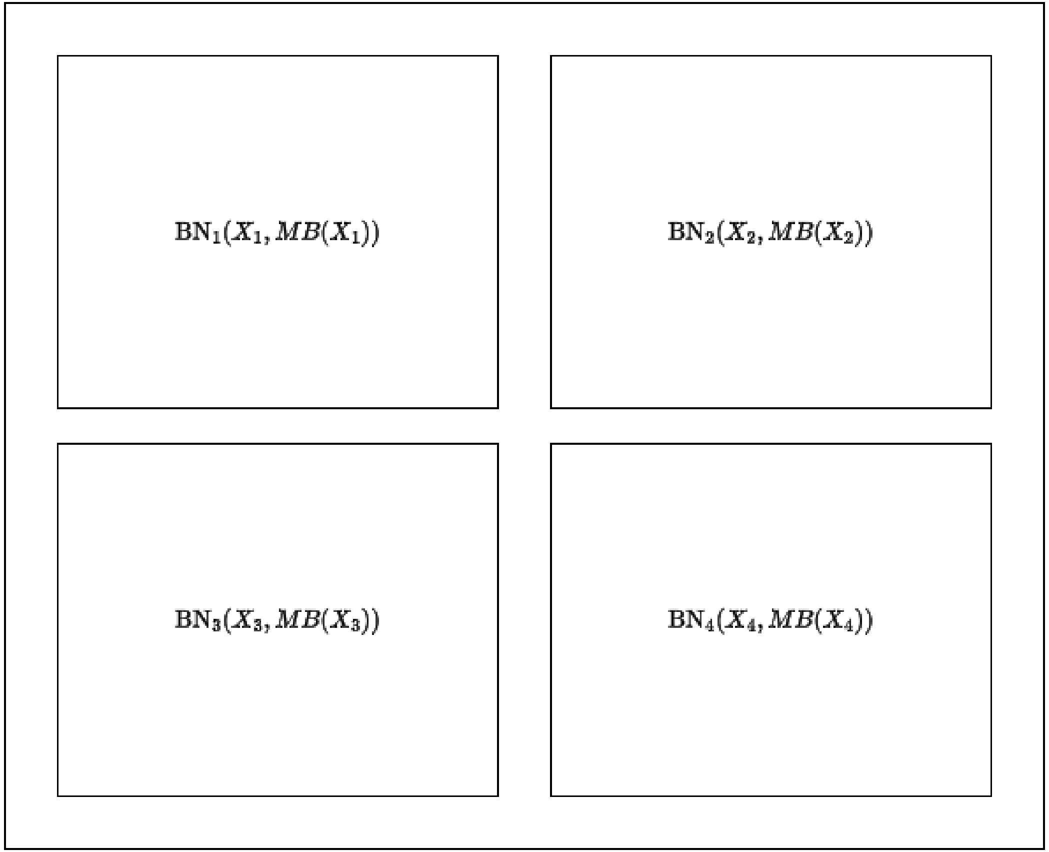
Hybrid Random Fields

Підготував - Жидок Ф. А.
Науковий керівник - Чорней Р. К.

Introduction



Introduction



Problem statement

- Adjust Hybrid Random Fields for supervised learning tasks by adding estimation of $\mathbb{E}[Y|X = x]$, taking into account different parameters estimation for continuous data and discrete data cases.
- Compare the quality of predictions on different datasets with classical machine learning approaches.

Discrete case

$$\hat{y}(x) = argmax_{c_i \in \mathcal{D}_Y} p(Y = c_i | MB(Y) = mb(Y))$$

where $\mathcal{D}_Y$ - domain of Y and $mb(Y) \subseteq x$

Continuous case - Gaussian case

Input:

$$(Y, MB(Y)) \sim N(\mu, \Sigma)$$

Solution - linear model:

$$X = MB(Y)$$

$$\beta = (X^T X)^{-1} X^T Y$$

$$\mathbb{E}[Y|X = x] = \hat{y}(x) = (x - \bar{X})\beta + \bar{Y}$$

Continuous case - mixture of Gaussians case

Input:

$$p(y, x \mid \Xi) = \sum_{i=1}^{C} \Pi_i p_i \left(y, x \mid \xi_i\right), \, p_i - \text{ normal density}$$

$$\text{where } \Xi = \left(\xi_1, \ldots, \xi_C\right), \forall i \in 1 \ldots C : \xi_i = \left(\mu_i, \Sigma_i\right), \sum_{i=1}^{C} \Pi_i = 1$$

Continuous case - mixture of Gaussians case

Solution:

$p(x, y) = \sum_{i=1}^{C} \Pi_i p_i(x, y)$ and $p(y \mid x) \propto p(x, y) \rightarrow$

$p(y \mid X = x) = \left(\sum_{i=1}^{C} \Pi_i p_i(X = x, y)\right) / K_0$, where K_0 - some constant

$$\forall i \in 1...C \exists K_i \in [0; 1] : p_i(X = x, y) = K_i p_i(y|X = x)$$

Using linearity of integral:

$$\mathbb{E}[Y|X = x] = \int_{-\infty}^{\infty} yp(y|X = x)\, dy$$

$$= (\int_{-\infty}^{\infty} y(\sum_{i=1}^{C} \Pi_i p_i(X = x, y))\, dy)/K_0$$

$$= (\sum_{i=1}^{C} \Pi_i K_i \mathbb{E}_i[Y|X = x])/K_0$$

where $K_i = p_i(X = x)$, $K_0 = p(X = x)$, which can be obtained and $\mathbb{E}_i[Y|X = x]$ is just an estimator from previous case

After obtaining $\hat{Y}$ posteriors $P(i \mid \hat{Y}, X)$ can be obtained and new estimate of Y can be calculated with posteriors instead of Π_i

* p, p_i - probability density functions

Continuous case - semiparametric case

Input:

$p_X(x) = p_M(f(x))|\prod_{i \in 1 \dots d} \frac{\partial f_i}{\partial x_i}|$, where $M \sim Normal$

$p_{(X,Y)}((x,y)) = p_N(f((x,y)))|\frac{\partial f_Y}{\partial y} \prod_{i \in 1 \dots d} \frac{\partial f_i}{\partial x_i}|$, where $N \sim Normal$

where $\hat{f}_i(x) = \hat{\mu}_i + \hat{\sigma}_i \hat{h}_i(x)$

$\hat{h}_i(x) = \Phi^{-1}(\hat{F}_i(x))$

$\hat{\mu}_i$ - sample mean of X_i

$\hat{\sigma}_i$ - sample standard deviation of X_i

Φ - cumulative density function of standard normal distribution

$\hat{F}_i(x)$ - sigmoid, fitted to sample cumulative density function of X_i

Continuous case - semiparametric case

Solution - exploit normality of **f(X)**:

1. $\forall X_i \in X$ obtain values $f_i(x_i)$
2. Obtain $\hat{f}_Y(x) = \mathbb{E}[f(Y)|f(X_1) = f(x_1), ..., f(X_n) = f(x_n)]$
3. Calculate $f^{-1}(\hat{f}_Y(x))$

Continuous case - nonparametric case

Input:

$$\hat{p}(y|x) = \frac{\sum_{i \in 1 \ldots n} K_{h1}(y - y_i) K_{h2}(x - x_i)}{\sum_{i \in 1 \ldots n} K_{h2}(x - x_i)}$$

$$K_h(u) = \frac{3}{4h^d}(1 - ||u||^2) \mathbf{1}_{||u|| \le 1}$$

Solution - Nadaraya-Watson regression:

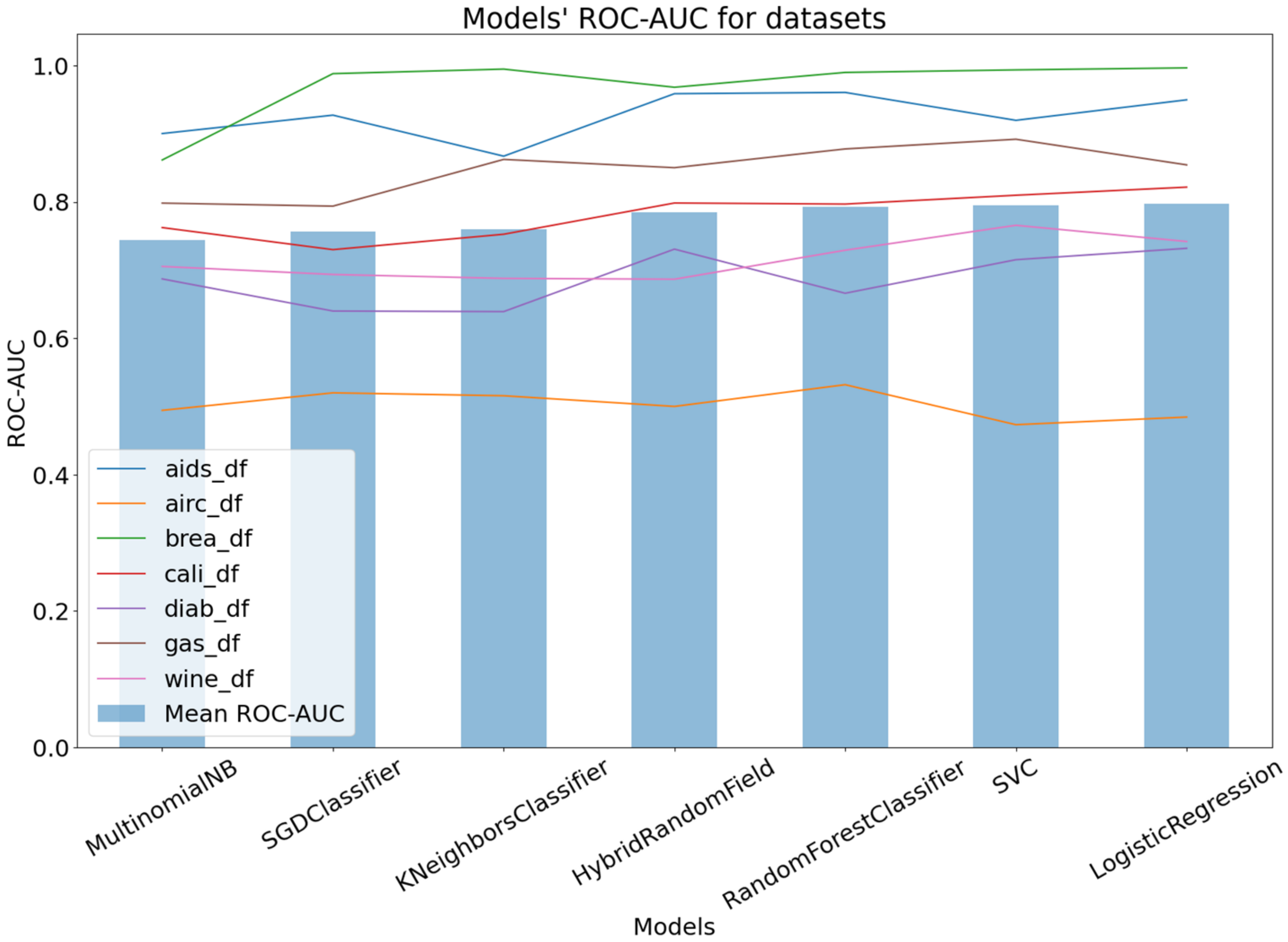
$$\hat{y}(x) = \frac{\sum_{i \in 1 \ldots n} K_{h2}(x - x_i) y_i}{\sum_{i \in 1 \ldots n} K_{h2}(x - x_i)}$$

$$h = 1.06 \min \left\{ s, \frac{IQR}{1.34} \right\} n^{-\frac{1}{5}}$$

Datasets

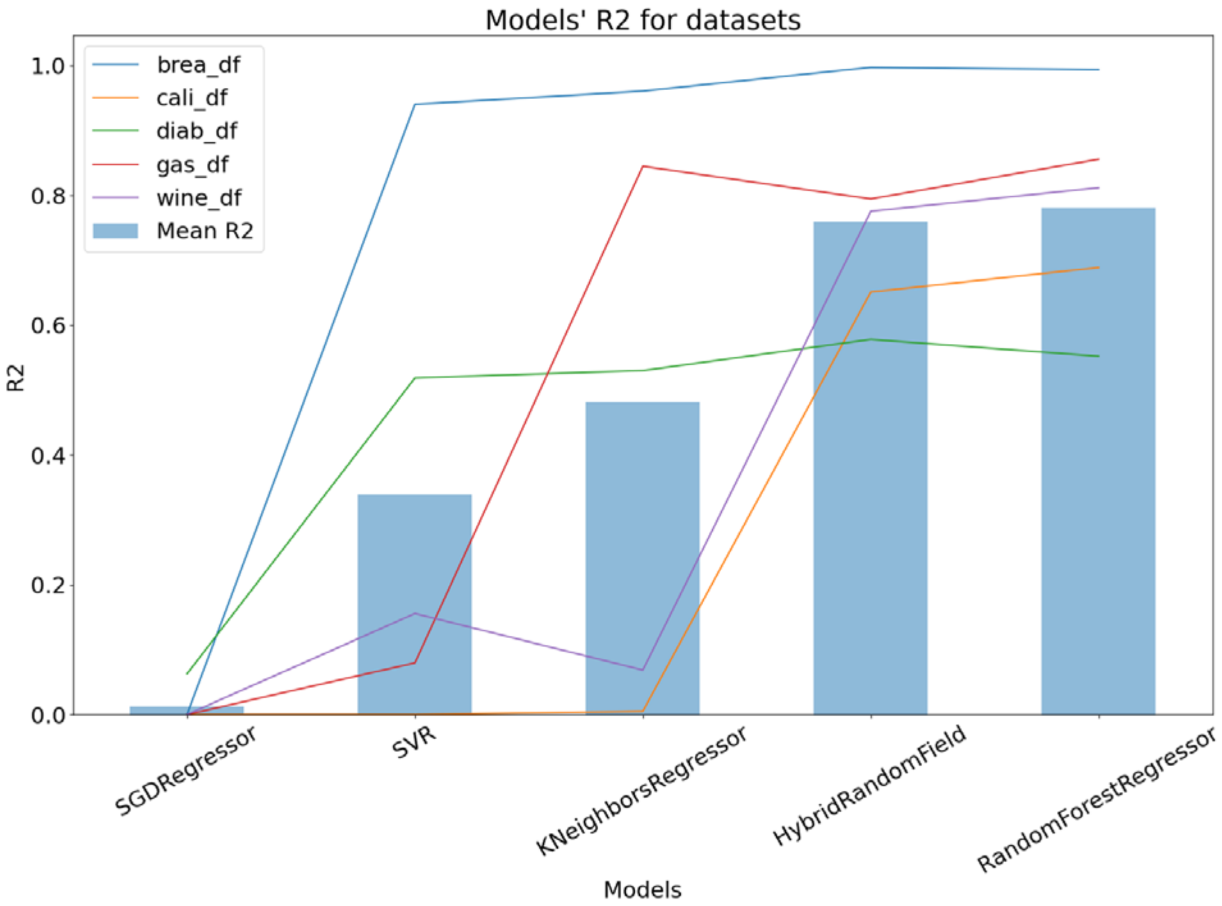
Dataset name	Target variable for classification	Target variable for regression	Dataset shape
California housing dataset	Median house value **(binarized)**	Median house value	(20640, 9)
Diabetes severeness dataset	Diabetes severeness measure **(binarized)**	Diabetes severeness measure	(442, 11)
Breast cancer wisconsin (diagnostic) dataset	Breast cancer disease flag	Area of the worst tissue	(569, 31)
Gas turbine CO and NOx emission dataset	NO + NO^2 emmission amount **(binarized)**	NO + NO^2 emmission amount	(7411, 11)
Aircraft weight and balance dataset	Number of passengers		(25000, 14)
Wine quality dataset	Wine quality	Alcohol level	(6497, 13)
AIDS clinical trials dataset	Indicator of death in determined time		(2139, 24)

Classification comparison



Regression comparison

DSName	Estimation Type	R2
brea_df	nonparametric	0.99
cali_df	semiparametric	0.65
diab_df	normal	0.57
gas_df	mixture model	0.79
wine_df	semiparametric	0.77



References

[1] Antonino Freno and Edmondo Trentin. *Hybrid Random Fields: A Scalable Approach to Structure and Parameter Learning in Probabilistic Graphical Models*, volume 15. 01 2011.

[2] Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10, 04 2009.

[3] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London, 1986.