An Efficient Shapley Value Computation for the Naive Bayes Classifier

Vincent Lemaire, Fabrice Clérot, Marc Boullé

European Conference on Machine Learning – Workshop AIMLAI

Agenda

- Introduction and reminder about Naïve Bayes (NB)
- Indicators of the importance of variables in the literature (NB)
- "Shapley"? What's this? (simply)
- Proposed calculation of a Shapley-type indicator
- Comparison with KernelShap
- Conclusion

A brief reminder about the naïve Bayes

$$P(C_z, X) = P(C_z)P(X|C_z) = P(X)P(C_z|X)$$

$$P(C_z|X) = \frac{P(C_z)P(X|C_z)}{P(X)}$$

A brief reminder about the naïve Bayes

$$P(C_z, X) = P(C_z)P(X|C_z) = P(X)P(C_z|X)$$

$$P(C_z|X) = \frac{P(C_z)P(X|C_z)}{P(X)}$$

$$P(C_z|X) = \frac{P(C_z)\,\prod_{i=1}^{d} P(X_i\,|C_z)}{\sum_{k=1}^{C} P(C_k)\,\prod_{i=1}^{d} P(X_i\,|C_k)}$$

A brief reminder about the naïve Bayes

$$P(C_z, X) = P(C_z)P(X|C_z) = P(X)P(C_z|X)$$

$$P(C_z|X) = \frac{P(C_z)P(X|C_z)}{P(X)}$$

$$P(C_z|X) = \frac{P(C_z)\ \prod_{i=1}^{d} P(X_i\,|C_z)}{\sum_{k=1}^{C} P(C_k)\ \prod_{i=1}^{d} P(X_i\,|C_k)}$$

[HMR99] JA Hoeting, David Madigan, and AE Raftery. Bayesian model averaging : a tutorial. *Statistical science*, 14(4) :382–417, 1999.

[LS94] Pat Langley and S Sage. Induction of Selective Bayesian Classifiers. In R Lopez De Mantras Poole and D, editors, *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pages 399–406. Morgan Kaufmann, 1994.

[Bou06b] Marc Boullé. Regularization and Averaging of the Selective Naive Bayes classifier. *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, pages 1680–1688, 2006.

A brief reminder about the naïve Bayes

$$P(C_z, X) = P(C_z)P(X|C_z) = P(X)P(C_z|X)$$

$$P(C_z|X) = \frac{P(C_z)P(X|C_z)}{P(X)}$$

$$P(C_z|X) = \frac{P(C_z)\ \prod_{i=1}^{d} P(X_i\,|C_z)}{\sum_{k=1}^{C} P(C_k)\ \prod_{i=1}^{d} P(X_i\,|C_k)}$$

$$P(C_z|X) = \frac{P(C_z)\ \prod_{i=1}^{d} P(X_i\,|C_z)^{W_i}}{\sum_{k=1}^{C} P(C_k)\ \prod_{i=1}^{d} P(X_i\,|C_k)^{W_i}}$$

[HMR99] JA Hoeting, David Madigan, and AE Raftery. Bayesian model averaging : a tutorial. *Statistical science*, 14(4) :382–417, 1999.

[LS94] Pat Langley and S Sage. Induction of Selective Bayesian Classifiers. In R Lopez De Mantras Poole and D, editors, *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pages 399–406. Morgan Kaufmann, 1994.

[Bou06b] Marc Boullé. Regularization and Averaging of the Selective Naive Bayes classifier. *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, pages 1680–1688, 2006.

A brief reminder about the naïve Bayes (2/2)

$$P(C_z|x_k) = \frac{P(C_z)\prod_{j=1}^{J} P(V_j = x_{jk}|C_z)^{W_j}}{\sum_{t=1}^{C}\left[P(C_t)\prod_{j=1}^{J} P(V_j = x_{jk}|C_t)^{W_j}\right]}$$

- Each instance is a vector of values (numerical or categorical).
- After discretization / grouping respectively for numerical / categorical variables, each explanatory variable is coded on H values.
- Each instance is then coded as a vector of discrete values.
- Conditional class probabilities ($P(V_j = x_{jk}|C_z)$) are estimated using a discretization method and a modality clustering method.

Agenda

- Introduction and reminder about Naïve Bayes (NB)
- Indicators of the importance of variables in the literature (NB)
- "Shapley"? What's this? (simply)
- Proposed calculation of a Shapley-type indicator
- Comparison with KernelShap
- Conclusion

Indicators of the importance of variables in the literature

Robnik-Sikonja, M. et I. Kononenko (2008). Explaining classifications for individual instances. *Knowledge and Data Engineering, IEEE Transactions on Knowledge and Data Engineering 20*, 589 – 600.

Lemaire, V., Boullé, M., Clérot, F., Gouzien, P.: A method to build a representation using a classifier and its use in a k nearest neighbors-based deployment. In: Proceedings of International Joint Conference on Neural Networks (2010)

- "Information Difference (IDI)"

$$\mathrm{IDI}_j^z = log\left(P(C_z|x_k)\right) - log\left(P(C_z|x_k\backslash V_j)\right)$$

- "Weight of Evidence (WOE)"

$$\mathrm{WoE}_j^z = log\left(odds(C_z|x_k)\right) - log\left(odds(C_z|x_k\backslash V_j)\right)$$

- "Modality probability (MOP)"-

$$\mathrm{MOP}_j^z = P(V_j = x_{jk}|C_z)$$

- "Log Modality probability (LMOP)"-

$$\mathrm{LMOP}_j^z = log\left(P(V_j = x_{jk}|C_z)\right)$$

- "Difference of probabilities (DOP)"

$$\mathrm{DOP}_j^z = P(C_z|x_k) - P(C_z|x_k\backslash V_j)$$

- "Kullback-Leibler divergence (KLD)"

$$\mathrm{KLD}_j^z = P(C_z|x_k)log\left(\frac{P(C_z|x_k)}{P(C_z|x_k\backslash V_j)}\right)$$

- "Minimum of variable probabilities difference (VPD)"

$$\mathrm{VPD}_j^z = P(V_j = x_{jk}|C_z) - \max_{q\neq z} P(V_j = x_{jk}|C_q)$$

Indicators of the importance of variables in the literature

Robnik-Sikonja, M. et I. Kononenko (2008). Explaining classifications for individual instances. *Knowledge and Data Engineering, IEEE Transactions on Knowledge and Data Engineering 20*, 589 – 600.

Lemaire, V., Boullé, M., Clérot, F., Gouzien, P.: A method to build a representation using a classifier and its use in a k nearest neighbors-based deployment. In: Proceedings of International Joint Conference on Neural Networks (2010)

- "Information Difference (IDI)"

$$\mathrm{IDI}_j^z = log\left(P(C_z|x_k)\right) - log\left(P(C_z|x_k \backslash V_j)\right)$$

- "Weight of Evidence (WOE)"

$$\mathrm{WoE}_j^z = log\left(odds(C_z|x_k)\right) - log\left(odds(C_z|x_k \backslash V_j)\right)$$

- "Modality probability (MOP)"-

$$\mathrm{MOP}_j^z = P(V_j = x_{jk}|C_z)$$

- "Log Modality probability (LMOP)"-

$$\mathrm{LMOP}_j^z = log\left(P(V_j = x_{jk}|C_z)\right)$$

- "Difference of probabilities (DOP)"

$$\mathrm{DOP}_j^z = P(C_z|x_k) - P(C_z|x_k \backslash V_j)$$

- "Kullback-Leibler divergence (KLD)"

$$\mathrm{KLD}_j^z = P(C_z|x_k) log\left(\frac{P(C_z|x_k)}{P(C_z|x_k \backslash V_j)}\right)$$

- "Minimum of variable probabilities difference (VPD)"

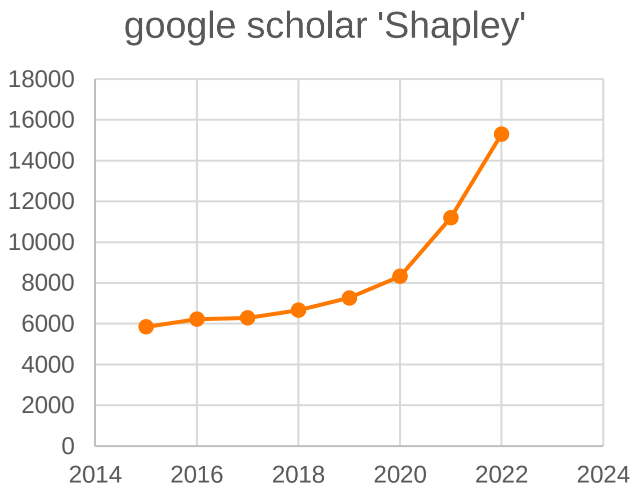
$$\mathrm{VPD}_j^z = P(V_j = x_{jk}|C_z) - \max_{q \neq z} P(V_j = x_{jk}|C_q)$$

Indicators of the importance of variables in the literature

Answers to review ☺

Do we need a new indicator ?

- Not sure, but Shapley is popular at the moment
  - Industrial consequences…



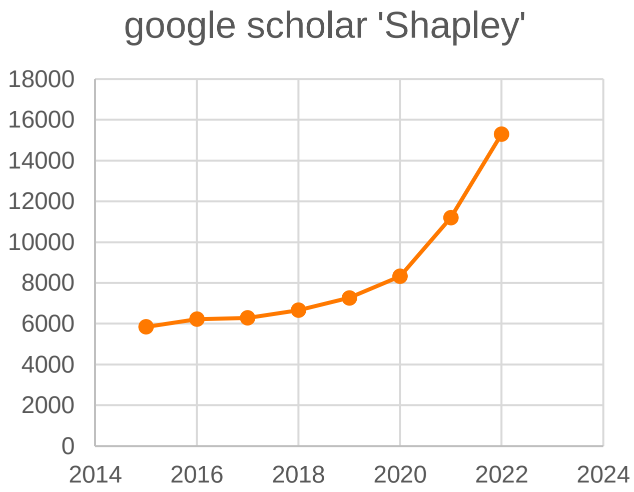
Indicators of the importance of variables in the literature

Answers to review ☺

Do we need a new indicator ?

- Not sure, but Shapley is popular at the moment
  - Industrial consequences…

https://github.com/KhiopsML/khiops



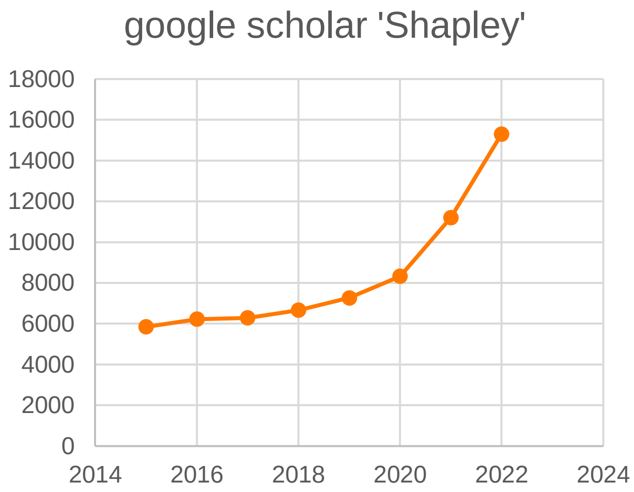
Indicators of the importance of variables in the literature

Answers to review ☺

Do we need a new indicator ?

- Not sure, but Shapley is popular at the moment
  - Industrial consequences…

https://github.com/KhiopsML/khiops



Are the weights sufficiently informative?

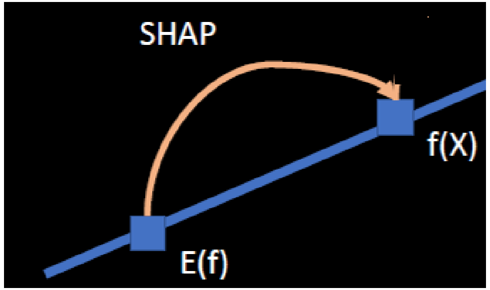
- No, they provide 'global information', whereas we want 'local information'.

$$P(C_z|x_k) = \frac{P(C_z)\prod_{j=1}^{J} P(V_j = x_{jk}|C_z)^{W_j}}{\sum_{t=1}^{C}\left[P(C_t)\prod_{j=1}^{J} P(V_j = x_{jk}|C_t)^{W_j}\right]}$$

Agenda

- Introduction and reminder about Naïve Bayes (NB)
- Indicators of the importance of variables in the literature (NB)
- "Shapley"? What's this? (simply)
- Proposed calculation of a Shapley-type indicator
- Comparison with KernelShap
- Conclusion

"Shapley"? What's this?

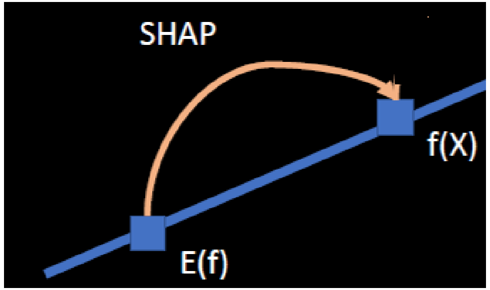


Put simply:

A game-theoretic method of calculating importance

Explains how variable values contribute to shifting predictions f(x) from the mean E[f(x)] of the prediction (f(x) is often taken as a 'value function')

"Shapley"? What's this?



Put simply:

A game-theoretic method of calculating importance

Explains how variable values contribute to shifting predictions f(x) from the mean E[f(x)] of the prediction (f(x) is often taken as a 'value function')

To do this:

- Step 1: When calculating Shapley values for a given individual, simulate different combinations of values for the input variables.
- Step 2: For each combination, calculate the difference between the predicted value and the mean of the predictions.
- The Shapley value of a variable then corresponds to the average contribution of its value according to the different combinations.

Agenda

- Introduction and reminder about Naïve Bayes (NB)
- Indicators of the importance of variables in the literature (NB)
- "Shapley"? What's this? (simply)
- Proposed calculation of a Shapley-type indicator
- Comparison with KernelShap
- Conclusion

Proposed calculation of a Shapley-type indicator (1/5)

In the case of NB, we propose to use the log ratio of probabilities as the Value Function (for a two-class classification problem):

$$
\begin{aligned}
LR &= log\left(\frac{P(C_1|X)}{P(C_0|X)}\right) \\
&= log\left(\frac{P(C_1)\prod_{i=1}^{d} P(X_i|C_1)^{W_i}}{\sum_{j=1}^{K}(P(C_j)\prod_{i=1}^{d} P(X_i|C_j)^{W_i})} \frac{\sum_{j=1}^{K}(P(C_j)\prod_{i=1}^{d} P(X_i|C_j)^{W_i})}{P(C_0)\prod_{i=1}^{d} P(X_i|C_1)^{W_i}}\right) \\
&= log\left(\frac{P(C_1)\prod_{i=1}^{d} P(X_i|C_1)^{W_i}}{P(C_0)\prod_{i=1}^{d} P(X_i|C_1)^{W_i}}\right) \\
&= log\left(\frac{P(C_1)}{P(C_0)}\right) + \sum_{i=1}^{d} W_i log\left(\frac{P(X_i|C_1)}{P(X_i|C_0)}\right)
\end{aligned}
$$

There are three reasons for choosing the log odd ratio as the value function

(i) the log odd ratio is in bijection with the score produced by the classifier

(ii) the log odd ratio has a linear form which simplifies calculations

(iii) this is also what is considered in the WoE.

Proposed calculation of a Shapley-type indicator (1/5)

In the case of NB, we propose to use the log ratio of probabilities as the Value Function (for a two-class classification problem):

$$
\begin{aligned}
LR &= log\left(\frac{P(C_1|X)}{P(C_0|X)}\right)\\
&= log\left(\frac{P(C_1)\prod_{i=1}^{d}P(X_i|C_1)^{W_i}}{\sum_{j=1}^{K}(P(C_j)\prod_{i=1}^{d}P(X_i|C_j)^{W_i})}\frac{\sum_{j=1}^{K}(P(C_j)\prod_{i=1}^{d}P(X_i|C_j)^{W_i})}{P(C_0)\prod_{i=1}^{d}P(X_i|C_1)^{W_i}}\right)\\
&= log\left(\frac{P(C_1)\prod_{i=1}^{d}P(X_i|C_1)^{W_i}}{P(C_0)\prod_{i=1}^{d}P(X_i|C_1)^{W_i}}\right)\\
&= log\left(\frac{P(C_1)}{P(C_0)}\right)+\sum_{i=1}^{d}W_i log\left(\frac{P(X_i|C_1)}{P(X_i|C_0)}\right)
\end{aligned}
$$

We stress here that the derivation above is only valid in the case of independent variables conditionally to the class variable, which is the standard assumption for the naive Bayes classifier.

In practice, we expect a variable selection method to result in a classifier relying on variables which are uncorrelated or only weakly correlated conditionally to the class.

Proposed calculation of a Shapley-type indicator (2/5)

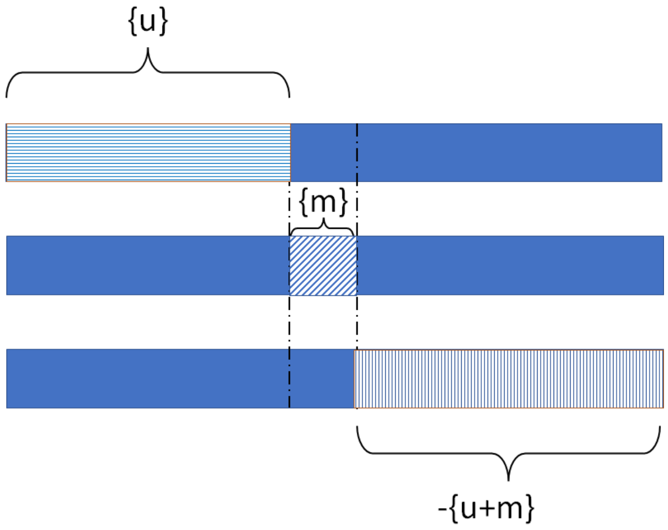
We need to calculate quantities such as

$$v(u) = \mathbb{E}_{X_{-u}|X_u=x_u}\left[LR(X_u = x_u^*, X_{-u})\right]$$

which we will write in "simplified" form below
$v(u) = E[(LR(X)|X_u = x^*_u)]$

Following the example of [Lundberg and Lee (2017)] and the Corollary1 with a linear model whose covariates are the log odd ratio as 'value function' we can decompose the subsets of variables into 3 groups:

- {u}
- {m}
- −{u + m}



Proposed calculation of a Shapley-type indicator (3/5)

On {u}, we condition on $X_u = x_u$ while on {m}, {u+m}, we perform averaging

$$
\begin{aligned}
v(u) &= \mathbb{E}\left[LR(X)|X_u = x_u^*)\right] \\
&= log(P(Y_1)/P(Y_0)) \\
&+ \sum_{k(k \in u)} w_k log\left(\frac{P(X_k = x_k{}^*|Y_1)}{P(X_k = x_k{}^*|Y_0)}\right) \\
&+ w_m \mathbb{E}_{X_m}\left[P(X_m = x_m) log\left(\frac{P(X_m = x_m|Y_1)}{P(X_m = x_m|Y_0)}\right)\right] \\
&+ \sum_{k(k \in -\{u+m\})} w_k \mathbb{E}_{X_k}\left[P(X_k = x_k) log\left(\frac{P(X_k = x_k|Y_1)}{P(X_k = x_k|Y_0)})\right)\right]
\end{aligned}
$$

Proposed calculation of a Shapley-type indicator (4/5)

Calculation of v(u + m) : The only difference is that we also condition on X_m

$$
\begin{aligned}
v(u+m) &= \mathbb{E}\left[LR(X)|X_{u+m} = x^*_{u+m})\right] \\
&= log(P(Y_1)/P(Y_0)) \\
&+ \sum_{k(k\in u)} w_k log\left(\frac{P(X_k = x_k{}^*|Y_1)}{P(X_k = x_k{}^*|Y_0)}\right) \\
&+ w_m\left[log\left(\frac{P(X_m = x^*_m|Y_1)}{P(X_m = x^*_m|Y_0)}\right)\right] \\
&+ \sum_{k(k\in -\{u+m\})} w_k \mathbb{E}_{X_k}\left[P(X_k = x_k) log\left(\frac{P(X_k = x_k{}^*|Y_1)}{P(X_k = x_k{}^*|Y_0)})\right)\right]
\end{aligned}
$$

Proposed calculation of a Shapley-type indicator (5/5)

So v(u + m) – v(u) :

$$v(u+m) - v(u) = w_m \left(log \left(\frac{P(X_m = x_m^* | Y_1)}{P(X_m = x_m^* | Y_0)} \right) - \mathbb{E}_{X_m} \left[P(X_m = x_m) log \left(\frac{P(X_m = x_m | Y_1)}{P(X_m = x_m | Y_0)} \right) \right] \right)$$

Interpretation and Discussion (1/3)

$$v(u+m)-v(u)=w_m\left(log\left(\frac{P(X_m=x_m^*|Y_1)}{P(X_m=x_m^*|Y_0)}\right)-\mathbb{E}_{X_m}\left[P(X_m=x_m)log\left(\frac{P(X_m=x_m|Y_1)}{P(X_m=x_m|Y_0)}\right)\right]\right)$$

The equation is the difference between the information content of X_m conditionally on $X_m = x_m^*$ and the expectation of this information.

In other words, it is the information contribution of the variable X_m for the value \$ $X_m = x_m^*$ \$ of the considered instance, contrasted by the average contribution on the entire database.

Interpretation and Discussion (2/3)

$$v(u+m)-v(u) = w_m \left(log \left(\frac{P(X_m = x_m^* | Y_1)}{P(X_m = x_m^* | Y_0)} \right) - \mathbb{E}_{X_m} \left[P(X_m = x_m) log \left(\frac{P(X_m = x_m | Y_1)}{P(X_m = x_m | Y_0)} \right) \right] \right)$$

$$- \left[log \left(\frac{1}{P(X_m = x_m^* | Y_1)} \right) - \sum_{X_m} \left(P(X_m = x_m) log \left(\frac{1}{P(X_m = x_m | Y_1)} \right) \right) \right]$$

$$+ \left[log \left(\frac{1}{P(X_m = x_m^* | Y_0)} \right) - \sum_{X_m} \left(P(X_m = x_m) log \left(\frac{1}{P(X_m = x_m | Y_0)} \right) \right) \right]$$

The terms in brackets $[\ldots]$ in equation are the difference between the information content related to the conditioning $X_m = x_m^*$ and the entropy of the variable X_m for each class (Y_0 and Y_1). This term measures how much conditioning on $X_m = x_m^*$ brings information about the target classes.

Interpretation and Discussion (3/3)

$$v(u+m)-v(u) = w_m \left(log \left(\frac{P(X_m = x_m^*|Y_1)}{P(X_m = x_m^*|Y_0)} \right) - \mathbb{E}_{X_m} \left[P(X_m = x_m) log \left(\frac{P(X_m = x_m|Y_1)}{P(X_m = x_m|Y_0)} \right) \right] \right)$$

When the numerical (resp. categorical) variables have been previously discretized into intervals (resp. groups of values), the complexity of the equation is linear in the number of discretized parts.

For an input vector made up of d variables, this complexity is

$O(\sum_{i=1}^{d} P_i)$

where P_i is the number of discretized parts of variable *i*.

Other points in the paper…

What's new or different about WoE [Good 1950] ?

$$\phi_m = w_m \left(log \left(\frac{P(X_m = x_m^* | Y_1)}{P(X_m = x_m^* | Y_0)} \right) - \mathbb{E}_{X_m} \left[P(X_m = x_m) log \left(\frac{P(X_m = x_m | Y_1)}{P(X_m = x_m | Y_0)} \right) \right] \right)$$

$$(WoE)_m = w_m \left(log \left(\frac{P(X_m = x_m^* | Y_1)}{P(X_m = x_m^* | Y_0)} \right) + log \left(\frac{1}{1} \right) \right)$$

See the « proof » in the paper

it's the reference that changes ...

both results have high agreements and the WoE doesn’t suffer from computation exhaustion

Answers to review ☺ (thanks for them) : So use the one you prefer versus the reference ... ☺

Agenda

- Introduction and reminder about Naïve Bayes (NB)
- Indicators of the importance of variables in the literature (NB)
- "Shapley"? What's this? (simply)
- Proposed calculation of a Shapley-type indicator
- Comparison with KernelShap
- Conclusion

Comparison with KernelShap
Datasets

Name	#Cont	#Cat	#Inst (N)	Maj. class.	Accuracy	AUC	#Var
Twonorm	20	0	7400	0.5004	0.9766	0.9969	20
Crx	6	9	690	0.5550	0.8112	0.9149	7
Ionosphere	34	0	351	0.6410	0.9619	0.9621	9
Spam	57	0	4307	0.6473	0.9328	0.9791	29
Tictactoe	0	9	958	0.6534	0.6713	0.7383	5
German	24	0	1000	0.7	0.7090	0.7112	9
Telco	3	18	7043	0.7346	0.8047	0.8476	10
Adult	7	8	48842	0.7607	0.8657	0.9216	13
KRFCC	28	7	858	0.9358	0.9471	0.8702	3
Breast	10	0	699	0.9421	0.975	0.9915	8

Table 1. Description of the datasets used in the experiments (KRFCC = KagRisk-FactorsCervicalCancer dataset)

Comparison with KernelShap
Results

Knowledge base

Name	N_k	Pearson	Kendall
Twonorm	200 (7400)	0.9027	0.7052
Crx	690 (690)	0.9953	0.9047
Ionosphere	351 (351)	0.9974	0.8888
Spam	200 (4307)	0.8829	0.7684
Tictactoe	958 (958)	1.0000	1.00
German	1000 (1000)	0.9974	0.9047
Telco	1000 (7043)	0.9633	0.7333
Adult	1000 (48842)	0.8373	0.7692
KRFCC	858 (858)	0.9993	1.00
Breast	699 (699)	0.9908	0.8571

Table 3. Correlation between our analytic Shapley and Kernelshap

good correlations for both coefficients

KernelShap 2 hours

Our Shapley < 10 secondes

Answers to review ☺ :

* since KernelShap is very slow the calculated importance values may not be reliable.

Comparison with KernelShap
Results

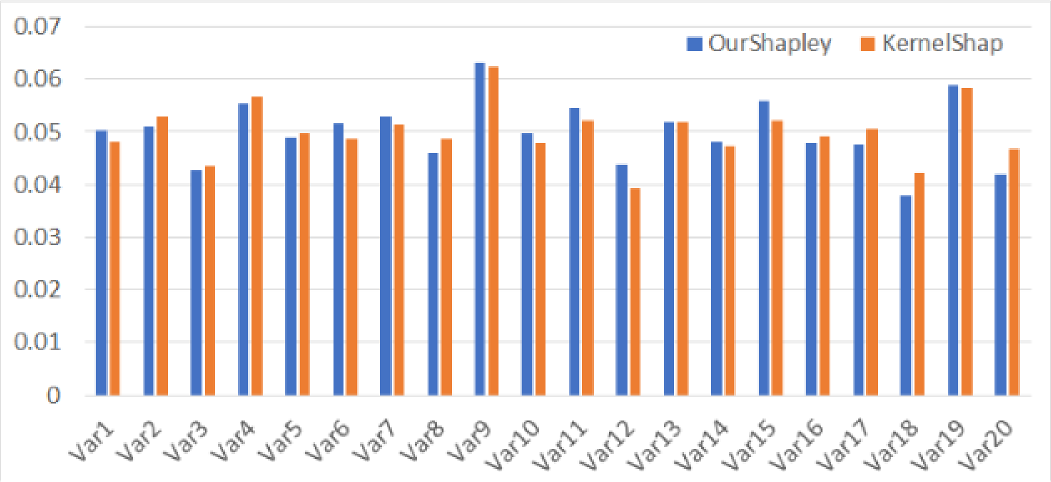


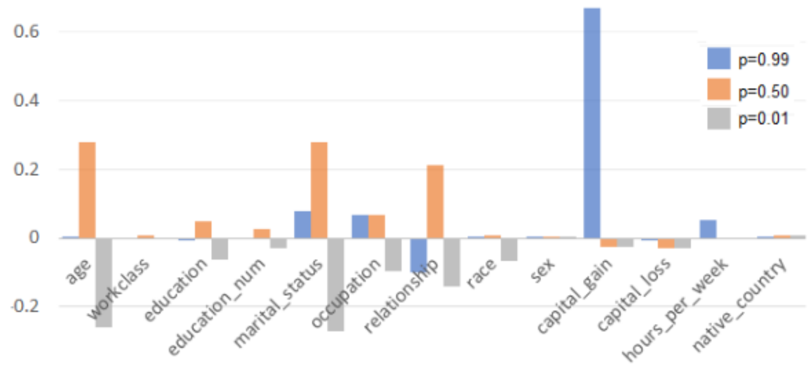
Fig. 3. Two Norm dataset: Comparison of our Shapley proposal and KernelShap.

the lower Kendall coefficient value is due to the fact that many variables have close Shapley values, resulting in differences in their value ranks.

Agenda

- Introduction and reminder about Naïve Bayes (NB)
- Indicators of the importance of variables in the literature (NB)
- "Shapley"? What's this? (simply)
- Proposed calculation of a Shapley-type indicator
- Comparison with KernelShap
- Conclusion

Conclusion …



In this paper,

we propose

a method for analytically calculating Shapley values in the case of the naive Bayes classifier.

This method exploits the hypothesis of independence of the variables conditional on the target to obtain the exact value of the Shapley values, with algorithmic complexity linear with the number of variables.

Unlike alternative evaluation/approximation methods, we use assumptions that are perfectly consistent with the underlying classifier

and we avoid approximation methods that are particularly time-consuming to compute.

Many more details in the article

GitHub The code and data used in this section are available in the GitHub repository at `https://tinyurl.com/ycxzkffk`.