Analysis of information-based selection criteria: supplemental material

Małgorzata Łazęcka[1,2][0000−0003−0975−4274]
and Jan Mielniczuk[1,2][0000−0003−2621−2303]

[1] Polish Academy of Sciences, Institute of Computer Science, Warsaw, Poland
[2] Warsaw University of Technology, Faculty of Mathematics and Information Sciences, Warsaw, Poland
{miel,malgorzata.lazecka}@ipipan.waw.pl

1 Möbius representation and formula for $I(X, Y|X_S)$

The p-way interaction information [2, 4] is

$$II(X_1, \ldots, X_p) = -\sum_{T\subseteq\{1,\ldots,p\}} (-1)^{p-|T|} H(X_T). \quad (1)$$

For $p = 2$, (1) reduces to mutual information, whereas for $p = 3$ it reduces to $II(X_1, X_2, X_3)$ introduced in (6) in the main text.

We state now Möbius representation of mutual information which plays an important role in the following development. For $S \subseteq \{1, 2, \ldots, p\}$ let X_S be a random vector coordinates of which have indices in S. Möbius representation [2, 3, 5] states that $I(X_S, Y)$ can be recovered from interaction informations

$$I(X_S, Y) = \sum_{k=1}^{|S|} \sum_{\{t_1,\ldots,t_k\}\subseteq S} II(X_{t_1}, \ldots, X_{t_k}, Y), \quad (2)$$

where $|S|$ denotes number of elements of set S. The natural way to approximate the conditional mutual information (CMI) is to use Möbius representation (2) which gives

$$\begin{aligned} I(X_{S\cup\{j\}}, Y) - I(X_S, Y) &= I(X_j, Y|X_S) \\ &= \sum_{k=0}^{|S|} \sum_{\{t_1,\ldots,t_k\}\subseteq S} II(X_{t_1}, \ldots, X_{t_k}, X_j, Y). \quad (3) \end{aligned}$$

The above formula allows however to obtain various natural approximations of CMI. Considering only the first term of the sum in (3) leads to first-order approximation $I(X_j, Y)$, which is a simple univariate filter. In the paper we focus on second order approximation, which is a balance between relatively accurate approximation of CMI and low computational cost. The first order approximation does not take interactions into account and that is why the second order approximation obtained by taking first two terms in (3) is usually considered. The corresponding score for candidate feature is

$$\begin{aligned} CIFE(X_j, Y|X_S) &= I(X_j, Y) + \sum_{i \in S} II(X_i, X_j, Y) \\ &= I(X_j, Y) + \sum_{i \in S} [I(X_i, X_j|Y) - I(X_i, X_j)]. \quad (4) \end{aligned}$$

2 Derivation of JMI

We show now reasoning leading to Joint Mutual Information Criterion JMI (cf. [6] and [5] on which the derivation below is based). Namely, if we define $S = \{j_1, \ldots, j_{|S|}\}$ we have for $i \in S$

$$I(X_j, X_S) = I(X_j, X_i) + I(X_j, X_{S \setminus \{i\}}|X_i)$$

Summing these equalities over all $i \in S$ and dividing by $|S|$ we obtain

$$I(X_j, X_S) = \frac{1}{|S|} \sum_{i \in S} I(X_j, X_i) + \frac{1}{|S|} \sum_{i \in S} I(X_j, X_{S \setminus \{i\}}|X_i)$$

and analogously

$$I(X_j, X_S|Y) = \frac{1}{|S|} \sum_{i \in S} I(X_j, X_i|Y) + \frac{1}{|S|} \sum_{i \in S} I(X_j, X_{S \setminus \{i\}}|X_i, Y).$$

Using $II(X_1, X_2, Y) = I(Y, X_1|X_2) - I(Y, X_1)$ and subtracting two last equations we obtain

$$I(X_j, Y|X_S) = I(X_j, Y) + \frac{1}{|S|} \sum_{i \in S} II(X_j, X_i, Y) + \frac{1}{|S|} \sum_{i \in S} II(X_j, X_{S \setminus \{i\}}, Y|X_i).$$

Moreover it follows from $II(X_1, X_2, Y) = I(Y, X_1|X_2) - I(Y, X_1)$ that when X_k is independent from $X_{S \setminus \{i\}}$ given X_i and these quantities are independent given X_i and Y the last sum is 0 and we obtain equality

$$\begin{aligned} JMI(X_j, Y|X_S) &= I(X_j, Y) + \frac{1}{|S|} \sum_{i \in S} II(X_j, X_i, Y) \\ &= I(X_j, Y) + \frac{1}{|S|} \sum_{i \in S} [I(X_j, X_i|Y) - I(X_j, X_i)]. \quad (5) \end{aligned}$$

3 Proof of Theorem 1

Theorem 1. *Differential entropy of X in (11) in the main text equals*

$$H(X) = h(\|\mu\|) + \frac{d-1}{2} \log(2\pi e),$$

where h is the differential entropy of one-dimensional gaussian mixture equal to $2^{-1}\{N(0,1)+N(a,1)\}$:

$$h(a) = -\int\limits_R \frac{1}{2\sqrt{2\pi}}\left(e^{-\frac{x^2}{2}} + e^{-\frac{(x-a)^2}{2}}\right) \cdot \log\left(\frac{1}{2\sqrt{2\pi}}\left(e^{-\frac{x^2}{2}} + e^{-\frac{(x-a)^2}{2}}\right)\right) dx. \quad (6)$$

Proof. In order to avoid burdensome notation we prove the theorem for $d = 2$ only. By the definition of differential entropy we have

$$H(X) = -\int\limits_{R^2} \frac{1}{2}\left(f_0(x_1,x_2) + f_\mu(x_1,x_2)\right) \cdot \log\left(\frac{1}{2}(f_0(x_1,x_2) + f_\mu(x_1,x_2))\right) dx_1 dx_2,$$

where X is defined in (11) in the main text for $d = 2$.

We calculate the integral above changing the variables according to the following rotation

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} \frac{\mu_1}{\|\mu\|} & -\frac{\mu_2}{\|\mu\|} \\ \frac{\mu_2}{\|\mu\|} & \frac{\mu_1}{\|\mu\|} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

Transformed densities f_0 and f_μ are equal $f_0(y_1,y_2) = \exp\left(-\left(y_1^2 + y_2^2\right)/2\right)/2\pi$ and $f_\mu(y_1,y_2) = \exp\left(-\left((y_1 - \|\mu\|)^2 + y_2^2\right)/2\right)/2\pi$. Applying above transformation, we can decompose $H(X)$ into sum of two integrals as follows

$$\begin{aligned} H(X) = &\int\limits_R \frac{1}{2\sqrt{2\pi}}\left(e^{-\frac{1}{2}y_1^2} + e^{-\frac{1}{2}(y_1-\|\mu\|)^2}\right) \\ &\cdot \log\left(\frac{1}{2\sqrt{2\pi}}\left(e^{-\frac{1}{2}y_1^2} + e^{-\frac{1}{2}(y_1-\|\mu\|)^2}\right)\right) dy_1 \\ &+ \int\limits_R \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y_2^2} \log\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y_2^2}\right) dy_2 \\ = &\, h(\|\mu\|) + \frac{1}{2}\log(2\pi e), \end{aligned}$$

where in the last equality the value $H(Z) = \log(2\pi e)/2$ for $N(0,1)$ variable Z is used. This ends the proof.

4 Proof of monotonicity of h

Lemma 1. *Differential entropy $h(a)$ of gaussian mixture defined in Theorem 1 is strictly increasing function of a.*

Proof. It is easy to see that h is differentiable and for calculation of its derivative integration in (6) and taking derivative can be interchanged. We show that derivative of h is positive. We obtain that by standard manipulations, using the fact that $x\exp(-x^2/2)$ is an odd function for the second equality below and change of variables for the third and the fifth equality.

$$
\begin{aligned}
-\frac{1}{2\sqrt{2\pi}}h'(a) &= \int\limits_R \Bigg((x-a)e^{-\frac{(x-a)^2}{2}} \\
&\qquad \cdot \log\left(\frac{1}{2\sqrt{2\pi}}\left(e^{-\frac{x^2}{2}}+e^{-\frac{(x-a)^2}{2}}\right)\right) + (x-a)e^{-\frac{(x-a)^2}{2}}\Bigg)\,dx \\
&= \int\limits_R (x-a)e^{-\frac{(x-a)^2}{2}} \log\left(\frac{1}{2\sqrt{2\pi}}\left(e^{-\frac{x^2}{2}}+e^{-\frac{(x-a)^2}{2}}\right)\right) dx \\
&= \int\limits_R xe^{-\frac{x^2}{2}} \log\left(\frac{1}{2\sqrt{2\pi}}\left(e^{-\frac{x^2}{2}}+e^{-\frac{(x+a)^2}{2}}\right)\right) dx \\
&= \int\limits_0^\infty xe^{-\frac{x^2}{2}} \log\left(\frac{1}{2\sqrt{2\pi}}\left(e^{-\frac{x^2}{2}}+e^{-\frac{(x+a)^2}{2}}\right)\right) dx \\
&\qquad + \int\limits_{-\infty}^0 xe^{-\frac{x^2}{2}} \log\left(\frac{1}{2\sqrt{2\pi}}\left(e^{-\frac{x^2}{2}}+e^{-\frac{(x+a)^2}{2}}\right)\right) dx \\
&= \int\limits_0^\infty xe^{-\frac{x^2}{2}} \Bigg(\log\left(\frac{1}{2\sqrt{2\pi}}\left(e^{-\frac{x^2}{2}}+e^{-\frac{(x+a)^2}{2}}\right)\right) \\
&\qquad - \log\left(\frac{1}{2\sqrt{2\pi}}\left(e^{-\frac{x^2}{2}}+e^{-\frac{(x-a)^2}{2}}\right)\right)\Bigg) dx.
\end{aligned}
$$

It follows from the last expression that $h'(a) > 0$ as $(x-a)^2 < (x+a)^2$ for $x > 0$ and $a > 0$ and therefore h is increasing.

5 Proof of Theorem 2

Theorem 2. *Differential entropy of*

$$X \sim \frac{1}{2}\mathcal{N}(0,\Sigma) + \frac{1}{2}\mathcal{N}(\mu,\Sigma)$$

equals

$$H(X) = h\left(\left\|\Sigma^{-1/2}\mu\right\|\right) + \frac{d-1}{2}\log(2\pi e) + \frac{1}{2}\log(\det\Sigma).$$

Proof. We apply Theorem 1 to multivariate random variable $Y = \Sigma^{-\frac{1}{2}}X$. We obtain

$$H(Y) = h\left(\left\|\Sigma^{-1/2}\mu\right\|\right) + \frac{d-1}{2}\log(2\pi e).$$

Using scaling property of differential entropy [1] we have

$$H(X) = H(Y) + \frac{1}{2}\log(\det \Sigma)$$

which completes the proof.

6 Proof of Theorem 3

Theorem 3. *Mutual information of X and Y where $Y \sim Bern(1/2)$ and $X|Y \sim \mathcal{N}(Y\mu, \Sigma)$ equals*

$$I(X,Y) = h\left(\left\|\Sigma^{-1/2}\mu\right\|\right) - \frac{1}{2}\log(2\pi e).$$

Proof. We will use here the fact that the entropy of multidimensional normal distribution $Z \sim \mathcal{N}(\mu_Z, \Sigma)$ equals (cf. [1], Theorem 8.4.1)

$$H(Z) = \frac{d}{2}\log(2\pi e) + \frac{1}{2}\log(\det \Sigma).$$

Therefore we have

$$I(X,Y) = H(X) - H(X|Y) = h\left(\left\|\Sigma^{-1/2}\mu\right\|\right) - \frac{1}{2}\log(2\pi e), \tag{7}$$

as

$$H(X|Y) = \frac{1}{2}H(X|Y=0) + \frac{1}{2}H(X|Y=1), \tag{8}$$

where $H(X|Y=i)$ stands for the entropy of X on the stratum $Y=i$. We notice that $H(X|Y=i) = H(Z)$, as the distribution of X on stratum $Y=i$ is normal with covariance matrix Σ and its entropy does not depend on the mean.

Remark 1. Note that Theorems 2 and 3 in conjunction with Lemma 1 show that entropy of mixture of two gaussians with the same covariance matrix and its mutual information with mixing distribution is strictly increasing function of the norm $\left\|\Sigma^{-1}\mu\right\|$. In particular, for $\Sigma = I$ entropy increases as the distance between centers of two gaussians increases.

References

1. Cover, T.M., Thomas, J.A.: Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing). Wiley-Interscience (2006)
2. Han, T.S.: Multiple mutual informations and multiple interactions in frequency data. Information and Control **46**(1), 26 – 45 (1980)
3. Meyer, P., Schretter, C., Bontempi, G.: Information-theoretic feature selection in microarray data using variable complementarity. IEEE Journal of Selected Topics in Signal Processing **2**(3), 261–274 (2008)
4. Ting, H.K.: On the amount of information. Theory Probab. Appl. **7**(4), 439–447 (1960)
5. Vergara, J.R., Estévez, P.A.: A review of feature selection methods based on mutual information. Neural Computing and Applications **24**(1), 175–186 (2014)
6. Yang, H.H., Moody, J.: Data visualization and feature selection: new algorithms for nongaussian data. Advances in Neural Information Processing Systems **12**, 687–693 (1999)