1 Statistical study of MCC

1.1 General framework

Let $\mathcal{T} = (T_n)_{0\leq n\leq N-1}$ be a corpus of N texts we want to classify into K classes. We have at our disposal a classifier that associates to each text its predicted class k_n, l_n denotes the true class of the text T_n.
Let $\widehat{C} = (\widehat{c}_{i,j})_{0\leq i,j\leq N-1}$ be the empirical confusion square matrix of size K defined by

$$\widehat{c}_{i,j} = \sum_{n=0}^{N-1} X_n(i,j) \tag{1}$$

with

$$X_n(i,j) = \mathbb{1}_{k_n=i}\mathbb{1}_{l_n=j}, \tag{2}$$

and let $\widehat{P} = (\widehat{p}_{i,j})$ be the proportion matrix defined by $\widehat{P} = \frac{1}{N}\widehat{C}$.

Assumption 1

For all $(i,j) \in \{1,...,K\}^2$, the elements of the sequence $(X_n(i,j))_{0\leq n\leq N-1}$ are i.i.d random variables of the same law as $X(i,j)$ with:

$$\mathbb{P}(X_n(i,j)=1) = 1 - \mathbb{P}(X_n(i,j)=0) = p_{i,j}. \tag{3}$$

Remark 1. *$p_{i,j}$ represents the probability that a text categorized in the i-class belongs to the j-class. By the strong law of large number we have*

$$\widehat{p}_{i,j} \xrightarrow[N\to\infty]{a.s.} p_{i,j} \tag{4}$$

1.2 Quantifying the efficiency of the classifying procedure

Definition 1. *The empirical Matthews Correlation Coefficient (MCC) metric is defined by:*

$$\widehat{MCC} = \frac{N\times\sum_{k=0}^{K-1}\widehat{c}_{k,k} - \sum_{k=0}^{K-1}(\sum_{i=0}^{K-1}\widehat{c}_{i,k}\times\sum_{j=0}^{K-1}\widehat{c}_{k,j})}{\sqrt{N^2-\sum_{k=0}^{K-1}(\sum_{i=0}^{K-1}\widehat{c}_{i,k})^2}\times\sqrt{N^2-\sum_{k=0}^{K-1}(\sum_{j=0}^{K-1}\widehat{c}_{k,j})^2}}. \tag{5}$$

and the true MCC is defined by

$$MCC_{true} = \frac{\sum_{k=0}^{K-1}p_{k,k} - \sum_{k=0}^{K-1}(\sum_{i=0}^{K-1}p_{i,k}\times\sum_{j=0}^{K-1}p_{k,j})}{\sqrt{1-\sum_{k=0}^{K-1}(\sum_{i=0}^{K-1}p_{i,k})^2}\times\sqrt{1-\sum_{k=0}^{K-1}(\sum_{j=0}^{K-1}p_{k,j})^2}}. \tag{6}$$

Remark 2. *Note that by multiplying the numerator and denominator of* (5) *by $\frac{1}{N^2}$, we have*

$$\widehat{MCC} = \frac{\sum_{k=0}^{K-1}\widehat{p}_{k,k} - \sum_{k=0}^{K-1}(\sum_{i=0}^{K-1}\widehat{p}_{i,k}\times\sum_{j=0}^{K-1}\widehat{p}_{k,j})}{\sqrt{1-\sum_{k=0}^{K-1}(\sum_{i=0}^{K-1}\widehat{p}_{i,k})^2}\times\sqrt{1-\sum_{k=0}^{K-1}(\sum_{j=0}^{K-1}\widehat{p}_{k,j})^2}}. \tag{7}$$

Hence, by Equation (4)

$$\widehat{MCC} \xrightarrow[N\to\infty]{a.s.} MCC_{true}. \tag{8}$$

In practice, we only have access to the empirical value $\widehat{MCC}$. It is desirable to quantify how far $\widehat{MCC}$ is from MCC_{true}. One classical way to answer this question is to provide a confidence interval for MCC_{true}. Moreover, if one wants to compare two different classification procedures, one can compute for each procedure the associated $\widehat{MCC}_1$ and $\widehat{MCC}_2$, and investigate whether there is statistical evidence that:

$$\widehat{MCC}_1 \leq \widehat{MCC}_2 (resp. \geq) \implies MCC_{true1} \leq MCC_{true2} (resp. \geq).$$

This can be done thanks to a confidence interval for $\widehat{MCC}_1 - \widehat{MCC}_2$ which is classically derived from a joint Central Limit Theorem for $(\widehat{MCC}_1, \widehat{MCC}_2)$.

1.2.1 Vectorial central limit theorem for $\widehat{MCC}$ and application to confidence interval

Let P_n be the square matrix of size K, where the variable at the column j and line i is $X_n(i,j)$. We obtain that

$$\widehat{P} = \frac{1}{N} \sum_{n=0}^{N-1} P_n. \tag{9}$$

Let $\mathcal{M}_K$ be the space of all the square real matrices of size K, and g be the application defined by:

$$\begin{array}{rccc} g & : & \mathcal{M}_K & \to & \mathbb{R} \\ & & (x_{i,j}) & \mapsto & g((x_{i,j})), \end{array}$$

where

$$g((x_{i,j})) = \frac{\sum_{k=0}^{K-1} x_{k,k} - \sum_{k=0}^{K-1} (\sum_{i=0}^{K-1} x_{i,k} \times \sum_{j=0}^{K-1} x_{k,j})}{\sqrt{1 - \sum_{k=0}^{K-1} (\sum_{i=0}^{K-1} x_{i,k})^2} \times \sqrt{1 - \sum_{k=0}^{K-1} (\sum_{j=0}^{K-1} x_{k,j})^2}}. \tag{10}$$

Theorem 1.1. *If Assumption 1 holds, then*

1. $$\sqrt{N}((\widehat{MCC} - MCC_{true}) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_1(0\,, \sigma_1), \tag{11}$$

   where $\sigma_1 = \sqrt{Dg^t \Sigma Dg}$, *and* Σ *is the covariance matrix of size* K^2

   $$\Sigma = \begin{bmatrix} Cov(X(0,0), X(0,0)) & ... & Cov(X(0,0), X(K-1,K-1)) \\ ... & ... & ... \\ Cov(X(K-1,K-1), X(0,0)) & ... & Cov(X(K-1,K-1), X(K-1,K-1)) \end{bmatrix},$$

   and $Dg = (\frac{\partial g}{\partial x_{i,j}})_{(i,j)\in\{0,K-1\}^2}$ *is the gradient at the coordinates* $\widehat{P}$ *of the application g.*

2. Let $0 \leq \alpha \leq 1$ and $\widehat{\sigma_1}$ a consistent estimator of σ_1, then

$$\lim_{N\to\infty} \mathbb{P}\left(\widehat{MCC} - \frac{q_\alpha \times \widehat{\sigma}_1}{\sqrt{N}} \leq MCC_{true} \leq \widehat{MCC} + \frac{q_\alpha \times \widehat{\sigma}_1}{\sqrt{N}}\right) = 1 - \alpha \tag{12}$$

where q_α is such that $\mathbb{P}\left(-q_\alpha \leq \mathcal{N}(0,1) \leq q_\alpha\right) = 1 - \alpha$.

Remark 3. *In particular, the confidence interval with a 95% confidence level is given by*

$$]\widehat{MCC} - \frac{1,96 \times \widehat{\sigma}_1}{\sqrt{N}}, \widehat{MCC} + \frac{1,96 \times \widehat{\sigma}_1}{\sqrt{N}}[. \tag{13}$$

Remark 4. *Closed formula for Dg and $\widehat{\Sigma}$ provided in the Appendix.*

Proof 1. *1. According to the vectorial central limit theorem, we have*

$$\sqrt{N}(\frac{1}{N}\sum_{n=0}^{N-1}(P_n - E[P_n]) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_{K^2}(\mathbf{0}, \Sigma)$$

Now setting $\sigma_1 = \sqrt{Dg^t \Sigma Dg}$, the Delta method ensures

$$\sqrt{N}((g(\widehat{P}) - g(\frac{1}{N}\sum_{n=0}^{N-1} E[P_n])) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_1(0, \sigma_1^2)$$

2. According to the Slutsky Lemma

$$\frac{\sqrt{N}((g(\widehat{P}) - g(\frac{1}{N}\sum_{n=0}^{N-1} E[P_n]))}{\widehat{\sigma_1}} \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_1(0, 1)$$

where,

$$\widehat{\sigma_1} = \sqrt{Dg^t \widehat{\Sigma} Dg} \tag{14}$$

and $\widehat{\Sigma}$ any consistant estimator of Σ.

1.2.2 Application to statistical test

In this Section, we aim to provide statistical evidence that two models' performance differ significantly. To do so, we test $H_0 : H_0 : MCC_{true1} = MCC_{true2}$ against $H_1 : H_1 : MCC_{true1} \neq MCC_{true2}$, where MCC_{true1} is the true MCC value for model 1, and MCC_{true2} is the MCC value for model 2.

Let $X_1(i,j), ..., X_n(i,j)$ be i.i.d variables and of the same law as $X(i,j)$. Let P_n be the square matrix of size K, where the variable at the column j and line i is $X_n(i,j)$.

Let $Y_1(i,j), ..., Y_n(i,j)$ be i.i.d variables and of the same law as $Y(i,j)$. Let Q_n be the square matrix of size K, where the variable at the column j and line i is $Y_n(i,j)$.

Let U_n be the vector of such that U_n is the juxtaposition of P_n and Q_n. Let $\overline{U}_N$ be defined by:

$$\overline{U}_N = \frac{1}{N}\sum_{n=0}^{N-1} U_n. \tag{15}$$

We define the functions J and f by

$$\begin{array}{rccc} J & : & M_K \times M_K & \longrightarrow & \mathbb{R}^2 \\ & & (x,y) & \longmapsto & (g(x), g(y)). \\ f & : & \mathbb{R}^2 & \longrightarrow & \mathbb{R} \\ & & (x,y) & \longmapsto & x - y. \end{array} \tag{16}$$

Theorem 1.2. *Under H_0,*

$$\sqrt{N}(\widehat{MCC}_1 - \widehat{MCC}_2) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}(0\,, \sigma_2^2). \tag{17}$$

where

$$\sigma_2 = \sqrt{Df^t \Sigma_b Df}, \tag{18}$$

Df is the gradient of the application f, and $\Sigma_b = DJ^t \Sigma_a DJ$ with DJ the gradient of the application J, and $\Sigma_a = Cov(U_n)$ is a covariance matrix of size $2K^2$.

Remark 5. *Hence, we will reject H_0 as soon as $|\widehat{MCC}_1 - \widehat{MCC}_2| > q_\alpha$. The threshold q_α is qualified thanks to Theorem 1.2 and the slutsky Lemma. For example the threshold for a test of level $\alpha = 5\%$ is given by $\frac{1.96\widehat{\sigma_2}}{\sqrt{N}}$.*

Remark 6. *Note that the framework developed in Section 1 can be applied on any of the classical metrics such as precision, recall, or f1 score.*

Remark 7. *Closed formula for DJ and Σ_a is in the Appendix.*

Proof 2. *According to the vectorial central limit theorem*

$$\sqrt{N}(\frac{1}{N}\sum_{n=0}^{N-1}(U_n - E[U_n]) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_{2K^2}(\mathbf{0}\,, \Sigma_a),$$

where $\Sigma_a = Cov(U_n)$ is a covariance matrix of size $2K^2$. According to the delta method,

$$\sqrt{N}((J(\overline{U}_N) - J(\frac{1}{N}\sum_{n=0}^{N-1} E[U_n])) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}_2(\mathbf{0}\,, \Sigma_b).$$

where $\Sigma_b = DJ^t \Sigma_a DJ$. Now since $J(\overline{U}_N) = (\widehat{MCC}_1, \widehat{MCC}_2)$ we have

$$\sqrt{N}(\begin{bmatrix}\widehat{MCC}_1\\ \widehat{MCC}_2\end{bmatrix} - \begin{bmatrix}MCC_{1true}\\ MCC_{2true}\end{bmatrix}) \xrightarrow[N\to\infty]{L} \mathcal{N}_2(\mathbf{0}\,, \Sigma_b).$$

Applying again the delta method, we get

$$\sqrt{N}((\widehat{MCC}_1 - \widehat{MCC}_2) - (MCC_{1true} - MCC_{2true}) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}(0\,, \sigma_2^2).$$

where $\sigma_2 = \sqrt{Df^t \Sigma_b Df}$, *with* Df *the gradient of the application* f
Under H_0, *we have*

$$\sqrt{N}(\widehat{MCC}_1 - \widehat{MCC}_2) \xrightarrow[N\to\infty]{\mathcal{L}} \mathcal{N}(0\,, \sigma_2^2).$$

APPENDIX:

Calculating $\widehat{\Sigma}$

For elements in the diagonal of Σ, we have:

$$\begin{aligned} Cov(X_{l_1,m_1}, X_{l_1,m_1}) &= E(X_{l_1,m_1}^2) - E(X_{l_1,m_1})^2 \\ &= p_{l_1,m_1} \times (1 - p_{l_1,m_1}). \end{aligned} \tag{1}$$

We obtain that a consistent estimation for $\widehat{\Sigma}$ is: $\widehat{p}_{l_1,m_1} \times (1 - \widehat{p}_{l_1,m_1})$.

For elements outside the diagonal, we obtain that:

$$\begin{aligned} Cov(X_{l_1,m_1,l_2,m_2}) &= E(X_{l_1,m_1} \times X_{l_2,m_2}) - E(X_{l_1,m_1}) \times E(X_{l_2,m_2}) \\ &= -p_{l_1,m_1} * p_{l_2,m_2}. \end{aligned} \tag{2}$$

We obtain that a consistent estimation for $\widehat{\Sigma}$ is: $-\widehat{p}_{l_1,m_1} * \widehat{p}_{l_2,m_2}$.

Calculating Dg

We introduce the following notation for the purpose of clarity in our calculations: Let $\widehat{p}_{i,.}$ designates the sum of the elements of row i of $\widehat{P}$.
Let $\widehat{p}_{.,j}$ designates the sum of the elements of column j of $\widehat{P}$.
Let $\widehat{p}_{i,-[j_0,\dots j_A]}$ with $A \in [0, K-1]$ designates $\widehat{p}_{i,.} - \sum_{a=0}^{A} \widehat{p}_{i,j_a}$
Let $\widehat{p}_{-[i_0,\dots i_A],j}$ with $A \in [0, K-1]$ designates $\widehat{p}_{.,j} - \sum_{a=0}^{A} \widehat{p}_{i_a,j}$
Let $Tr(\widehat{p})$ designates the sum of the elements in the diagonal of $\widehat{P}$.
We must obtain $\frac{\partial g}{\partial \widehat{p}_{i,j}}$. We distinguish between the case where $i = j$ and $i \neq j$.

When $i = j$, we have:

$$g(x_{l,l}) = \frac{x_{l,l} + a - (b + x_{l,l}^2 + x_{l,l} \times c)}{\sqrt{1 - d - (x_{l,l} + e)^2} \times \sqrt{1 - f - (x_{l,l} + g)^2}}. \tag{3}$$

Then we have:

$$\frac{\partial g}{\partial x_{l,l}} = \frac{(1 - 2 \times x_{l,l} - c) \times A \times B + ((x_{l,l} + e) \times \frac{B}{A} + (x_{l,l} + g) \times \frac{A}{B}) \times (x_{l,l} + a - (b + x_{l,l}^2 + x_{l,l} \times c))}{A^2 \times B^2}$$

with:

$$A = \sqrt{1 - d - (x_{l,l} + e)^2} \quad \text{and} \quad B = \sqrt{1 - f - (x_{l,l} + g)^2}.$$

When the input for g is our matrix $\widehat{P}$, we have:

$$a = Tr(\widehat{p}) - \widehat{p}_{l,l}, \quad b = \sum_{k \neq l}^{K-1} (\widehat{p}_{.,j} * \widehat{p}_{i,.}) + \widehat{p}_{-[l],l} \times \widehat{p}_{l,-[l]}, \quad c = \widehat{p}_{-[l],l} + \widehat{p}_{l,-[l]}$$

$$d = \sum_{k \neq l}^{K-1} \widehat{p}_{.,k}^2, \quad e = \widehat{p}_{-[l],l}, \quad f = \sum_{k \neq l}^{K-1} \widehat{p}_{k,.}^2$$

$$g = \widehat{p}_{l,-[l]} \quad \text{and} \quad x_{l,l} = \widehat{p}_{l,l}.$$

When $i \neq j$, we have:

$$g(x_{l,m}) = \frac{a_1 - (b_1 + (x_{l,m} + \widehat{c}_1) \times d_1 + (x_{l,m} + e_1) \times f_1)}{\sqrt{1 - g_1 - (x_{l,m} + h_1)^2} \times \sqrt{1 - i_1 - (x_{l,m} + j_1)^2}}, \tag{4}$$

then we have:

$$\frac{\partial g}{\partial x_{l,m}} =$$

$$\frac{-(d_1 + f_1)A_1B_1 + ((x_{l,m} + h_1)B_1/A_1 + (x_{l,m} + j_1)A_1/B_1)(a_1 - (b_1 + (x_{l,m} + \widehat{c}_1)d_1 + (x_{l,m} + e_1)f_1)}{A_1^2 B_1^2}$$

with:

$$A_1 = \sqrt{1 - g_1 - (x_{l,m} + h_1)^2} \quad \text{and} \quad B_1 = \sqrt{1 - i_1 - (x_{l,m} + j_1)^2}.$$

When the input for g is our matrix $\widehat{P}$, we have:

$$a_1 = Tr(\widehat{p}), \quad b_1 = \sum_{k \neq (l,m)}^{K-1} (\widehat{p}_{.,k} * \widehat{p}_{k,.}), \quad c_1 = \widehat{p}_{l,-[m]}$$

$$d_1 = \widehat{p}_{.,l}, \quad e_1 = \widehat{p}_{-[l],m}, \quad f_1 = \widehat{p}_{m,.}$$

$$g_1 = \sum_{k \neq m}^{K-1} \widehat{p}_{.,k}^2, \quad h_1 = \widehat{p}_{-[l],m}, \quad i_1 = \sum_{k \neq l}^{K-1} \widehat{p}_{k,.}^2$$

$$j_1 = \widehat{p}_{l,-[m]} \quad \text{and} \quad x_{l,m} = \widehat{p}_{l,m}.$$

calculating $\widehat{\Sigma_a}$

We have 6 cases:

- $Cov(X_{l_1,m_1}, X_{l_2,m_2})$
- $Cov(X_{l_1,m_1}, X_{l_1,m_1})$
- $Cov(Y_{l_1,m_1}, Y_{l_2,m_2})$
- $Cov(Y_{l_1,m_1}, Y_{l_1,m_1})$
- $Cov(X_{l_1,m_1}, Y_{l_1,m_1})$
- $Cov(X_{l_1,m_1}, Y_{l_2,m_2})$

with $(l_1, m_1) \neq (l_2, m_2)$.

The four first cases are already covered by our previous work on confidence intervals.
For the two next cases, we obtain that:

$$Cov(X_{l_1,m_1}, Y_{l_1,m_1}) = E(X_{l_1,m_1} \times Y_{l_1,m_1}) - E(X_{l_1,m_1}) \times E(Y_{l_1,m_1}) \tag{5}$$

We obtain that a consistent estimation for $\widehat{\Sigma_a}$ is:

$$\frac{\sum_{n=0}^{N-1} X_{n(l_1,m_1)} \times Y_{n(l_1,m_1)}}{N} - \widehat{p}_{l_1,m_1} \times \widehat{q}_{l_1,m_1}$$

$$Cov(X_{l_1,m_1}, Y_{l_2,m_2}) = E(X_{l_1,m_1} \times Y_{l_2,m_2}) - E(X_{l_1,m_1}) \times E(Y_{l_2,m_2}) \tag{6}$$

We obtain that a consistent estimation for $\widehat{\Sigma_a}$ is:

$$\frac{\sum_{n=0}^{N-1} X_{n(l_1,m_1)} \times Y_{n(l_2,m_2)}}{N} - \widehat{p}_{l_1,m_1} \times \widehat{q}_{l_2,m_2}$$

Notice that when $X_n(i1, j1) = 1$, it means that $j1$ is the true label for the n^{th} text. As such, $Y_n(i2, j2) = 0$ when $j1 \neq j2$. This means that for cases where $j1 \neq j2$, a consistent estimation for $\widehat{\Sigma_a}$ is:

$$-\widehat{p}_{l_1,m_1} \times \widehat{q}_{l_2,m_2}$$

Calculating DJ

We obtain:

$$\frac{\partial DJ_1(x,y)}{\partial y} = \frac{\partial g(x)}{\partial y} = 0, \quad \frac{\partial DJ_2(x,y)}{\partial x} = \frac{\partial g(y)}{\partial x} = 0$$