# Influence functions of the error rates of classification based on clustering

C. Ruwet and G. Haesbroeck

University of Liège, Belgium

cruwet@ulg.ac.be

## Generalized $k$-means

Assume that $X \sim F$ arises from 2 groups $G_1$ and $G_2$ with probabilities $\pi_i(F) = \mathbb{P}_F[X \in G_i]$ (called the prior probabilities), then $F$ is a mixture of two distributions : $F = \pi_1(F)F_1 + \pi_2(F)F_2$ (with density $f = \pi_1 f_1 + \pi_2 f_2$). In this particular setting, cluster analysis may be performed in order to find the underlying groups. Several procedures to construct clusters are available, a classical one is the 2-means algorithm that is a particular case of the generalized 2-means procedure. For a suitable nondecreasing penalty function $\Omega : \mathbb{R}^+ \to \mathbb{R}^+$, the clusters'centers $T_1(F)$ and $T_2(F)$ are solutions of the minimization problem

$$\min_{\{t_1,t_2\}\subset\mathbb{R}^p} \int \Omega\left(\inf_{1\leq j\leq 2} \|x - t_j\|\right) dF(x) \tag{1}$$

while, for all $x \in \mathbb{R}^p$, the corresponding classification rule is given by

$$R_F(x) = C_j(F) \Leftrightarrow \Omega(\|x - T_j(F)\|) = \min_{1\leq i\leq 2} \Omega(\|x - T_i(F)\|). \tag{2}$$

As García-Escudero and Gordaliza (1999), who studied some robustness properties of these estimators, we restrict ourselves to the univariate case. In this case, the clusters take the simplest forms $C_1(F) = ]-\infty, C(F)[$ and $C_2(F) = ]C(F), +\infty[$ where $C(F) = \dfrac{T_1(F) + T_2(F)}{2}$ is the cut-off point.

## Empirical and theoretical error rates

The performance of a classification rule can be measured by the error rate which is the probability to misclassify data. Two types of error rates can be computed : a theoretical one and a more empirical one. The first one can be written as $\mathrm{ER}(F, F_m)$ where $F$ is the distribution of the training sample used to set up the classification rule and $F_m$ (model distribution) is the distribution under which the quality of the rule is assessed (via a test sample). Often, this test sample is also used to estimate the prior probabilities. The empirical error rate corresponds to $\mathrm{ER}(F, F)$, meaning that the classification rule is tested on the same sample as the one used to set up the rule.

More formally, the theoretical error rate is defined as

$$\mathrm{ER}(F, F_m) = \sum_{j=1}^{2} \pi_j(F_m)\mathbb{P}_{F_m}\left[R_F(X) \neq C_j(F)\middle| G_j\right] \tag{3}$$

while the empirical error rate takes the form

$$\mathrm{ER}(F, F) = \sum_{j=1}^{2} \pi_j(F)\mathbb{P}_{F}\left[R_F(X) \neq C_j(F)\middle| G_j\right]. \tag{4}$$

In ideal circumstances, $F = F_m$ and the two error rates are identical. However, if the distribution $F$ is contaminated, $F_\varepsilon$ say, the two error rates are different.

Equation (3) becomes

$$\mathrm{ER}(F_\varepsilon, F_m) = \sum_{j=1}^{2} \pi_j(F_m)\mathbb{P}_{F_m}\left[R_{F_\varepsilon}(X) \neq C_j(F_\varepsilon)\middle| G_j\right] \tag{5}$$

and equation (4) becomes

$$\mathrm{ER}(F_\varepsilon, F_\varepsilon) = \sum_{j=1}^{2} \pi_j(F_\varepsilon)\mathbb{P}_{F_\varepsilon}\left[R_{F_\varepsilon}(X) \neq C_j(F_\varepsilon)\middle| G_j\right]. \tag{6}$$

One sees that the theoretical error rate is only contaminated through the classification rule while the contamination is everywhere in the empirical error rate. This difference will be stressed in the sequel when focusing on point mass contamination, i.e. $F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_x$, in the same spirit as Croux et al. (2008) and Croux et al. (2008) did in discriminant analysis.

The model distribution of interest is a mixture of two homoscedastic normal distributions : $F = \pi_1\, N(\mu_1, \sigma^2) + (1-\pi_1)\, N(\mu_2, \sigma^2)$ with $\mu_1 < \mu_2$ since Qiu and Tamhane (2007) proved the optimality of the 2-means clustering method under a particular case of normal mixture : the model $F_N = 0.5\, N(\mu_1, \sigma^2) + 0.5\, N(\mu_2, \sigma^2)$. Under the optimal model, the cut-off point corresponds to the midpoint between the means of the two groups $C(F_N) = \dfrac{\mu_1 + \mu_2}{2}$.

## Influence function

Hampel et al. (1986) defined the influence function of a statistical functional $T$ at a distribution $F$ as $\mathrm{IF}(x; T, F) = \dfrac{\partial}{\partial\varepsilon} T[(1-\varepsilon)F + \varepsilon\Delta_x]\Big|_{\varepsilon=0}$ for those $x$ where this derivative exists. It measures the impact on the statistical functional of an infinitesimal contamination at the point $x$. Two interesting properties of this function are :

- $E_F[\mathrm{IF}(X; T, F)] = 0$,
- $T(F_\varepsilon) \approx T(F) + \varepsilon\mathrm{IF}(x; T, F)$ (First order Taylor expansion of $T$ at $F$).

## Theoretical error rate


FIGURE 1: Optimal error rate (solide line) and contaminated theoretical error rate (dashed line) as a function of the amount of contamination (left panel) and as a function of the position of the contaminated mass (right panel) under model $F_N$ with $\mu_1 = -1$, $\mu_2 = 1$ and $\sigma = 1$. The position of the contamination is set to $x = -0.5$ (left panel) and the percentage of contamination is set at 10% (right panel).

Figure 1 shows that the contaminated theoretical error rate $\mathrm{ER}(F_\varepsilon, F_N)$ is always bigger than the optimal error rate $\Phi\left(\frac{\mu_1-\mu_2}{2\sigma}\right) = \mathrm{ER}(F_N, F_N)$. Then, the first order Taylor expansion of this error rate at the distribution $F_N$ implies that $\mathrm{IF}(x; \mathrm{ER}, F_N) \geq 0$ and the first property of the influence function leads to $\mathrm{IF}(x; \mathrm{ER}, F_N) \equiv 0$ for all $x \in \mathbb{R}$. Under this model, one then needs to compute the second order influence function (second derivative of the contaminated theoretical error rate) to study the impact of contamination on the theoretical error rate. The right panel shows that the smallest impact on the theoretical error rate comes from contamination near the centers of the groups.

**Proposition 1.** *The influence function of the theoretical error rate of the generalized 2-means procedure is given by*

$$\mathrm{IF}(x; \mathrm{ER}, F) = \frac{1}{2}\{\mathrm{IF}(x; T_1, F) + \mathrm{IF}(x; T_2, F)\}\{\pi_2(F)f_2(C(F)) - \pi_1(F)f_1(C(F))\} \tag{7}$$

*for all $x \neq C(F)$.*

Expressions of $\mathrm{IF}(x; T_1, F)$ and $\mathrm{IF}(x; T_2, F)$ have been computed by García-Escudero and Gordaliza (1999).

Due to the symmetry under $F_N$, one has $f_{N,1}\left(\frac{\mu_1+\mu_2}{2}\right) = f_{N,2}\left(\frac{\mu_1+\mu_2}{2}\right)$ and then the influence function of the theoretical error rate is null.

## Empirical error rate


FIGURE 2: Optimal error rate (solide line) and contaminated empirical error rate (dashed line) as a function of the amount of contamination (left panel) and as a function of the position of the contaminated mass (right panel) under model **(N)** with $\mu_1 = -1$, $\mu_2 = 1$ and $\sigma = 1$. The position of the contamination is set to $x = -0.5$ (left panel) and the percentage of contamination is set at 10% (right panel).

Figure 2 shows that the behaviour of the empirical error rate under contamination is quite different because contamination may make this error rate decrease even under optimality. Indeed, when the contamination is classified in the cluster corresponding to its group, i.e. when it is well classified, the empirical error rate $\mathrm{ER}(F_\varepsilon, F_\varepsilon)$ is lower than the optimal one $\Phi\left(\frac{\mu_1-\mu_2}{2\sigma}\right) = \mathrm{ER}(F_N, F_N)$. Consequently, the influence function of this error rate under $F_N$ is not identically null any more. As the cut-off point is near zero, this behaviour is highlighted on the right panel. This is the same kind of phenomenon as in regression where good leverage points are outliers that may improve the regression outputs whereas bad leverage points have a negative impact.

**Proposition 2.** *The influence function of the empirical error rate of the generalized 2-means method is given by*

$$\begin{aligned}\mathrm{IF}(x; \mathrm{ER}, F) = {} & \frac{1}{2}(\mathrm{IF}(x; T_1, F) + \mathrm{IF}(x; T_2, F))\{\pi_2(F)f_2(C(F)) - \pi_1(F)f_1(C(F))\} \\ & + I\{x \leq C(F)\}(1 - 2\delta_1(x)) + \delta_1(x) - \pi_1(F)\{1 - F_1(C(F))\} - \pi_2(F)F_2(C(F))\end{aligned} \tag{8}$$

*for all $x \neq C(F)$.*

Under $F_N$, this expression reduces to

$$\mathrm{IF}(x; \mathrm{ER}, F_N) = \mathrm{I}\{x \leq C(F_N)\}(1 - 2\delta_1(x)) + \delta_1(x) - \{1 - F_{N,1}(C(F_N))\}$$

which is not identically null.

**References** :

- Croux C., Filzmoser P., and Joossens K. (2008), Classification efficiences for robust linear discriminant analysis. *Statistica Sinica*, 18, 581–599.
- Croux C., Filzmoser P., and Joossens K. (2008), Logistic discrimination using robust estimators : an influence function approach. *The Canadian Journal of Statistics*, 36, 157–174.
- Garcia-Escudero L.A., and Gordaliza A. (1999), Robustness Properties of $k$ Means and Trimmed $k$ Means. *Journal of the American Statistical Association*, 94, 956–969.
- Hampel F.R., Ronchetti E.M., Rousseeuw P.J., and Stahel W.A.(1986), Robust Statistics : The Approach Based on Influence Functions, John Wiley and Sons, New-York.
- Qiu D., and Tamhane A.C. (2007), A comparative study of the $k$-means algorithm and the normal mixture model for clustering : Univariate case. *Journal of Statistical Planning and Inference*, 137, pp. 3722-3740.