Sample Size Determination Methods for Classification

A. Motrenko, V. Strijov

Moscow Institute of Physics and Technology

International Digital Processing Conference
Barcelona, 2016.

Sample size determination (SSD) problem

The goal is to design a method of sample size determination, that would accurately estimate the number of observations, required to obtain classification results given the model.

From (Sadia and Hossain, 2014)
"A good statistical study is one that is well designed and leads to a valid conclusion. "

- Sadia and Hossain, 2014. Contrast of Bayesian and Classical Sample Size Determination, Journal of Modern Applied Statistical Methods.
- Wang and Gelfand, 2002. A simulation-based approach to bayesian sample size determination for performance under a given model and for separating models. Statistical Science.
- Motrenko, Strijov, and Weber, 2014. Bayesian sample size estimation for logistic regression. Journal of Computational and Applied Mathematics

Various cases of SSD problem

1. No data is not available, $m = 0$. Use the *data generation hypothesis* to derive sample size estimate m^*.

2. A large amount of data is available, $m \to \infty$. Verify the criteria of interest directly on the observed data for various sample sizes m.

3. Some data has been observed $0 < m < m^*$. Adjust the data generation hypothesis according to the observed data before making predictions of m^*.

Motivation: Classification of patients with Cardio-Vascular Disease

Consider two groups of patients: $y \in \{\text{A1}, \text{A3}\}$; each patient is described by a set of markers $\mathbf{x}$.

Classes ⟶	Groups of patients	The patients have classification labels “A1” and“A3”.
Objects ⟶	Patients	We have measured data for 14 patients in the group “A1” and 17 patients in the group “A3”.
Features ⟶	Markers	We have 20 markers: K, L, K/M, L/M, K/N, K/O, L/O, K/P, L/P, K/Q, K/R, L/R, L/R/SA, L/T/SA, L/T/SO, U/V, U/W, U/X, U/Y, U/Z

Object–Feature (Patient–Marker) table, an extract

Class	Patient name	K	L	K/M	L/M	
A1	C001	58.3	16.7	0.52	0.00	
A1	C004	40.2	6.0	NaN	NaN	
A1	C005	54.3	13.1	NaN	NaN	
A1	C008	48.7	9.8	0.05	0.02	etc.
A3	023	46.6	21.2	0.40	0.08	
A3	026	50.7	26.2	0.12	0.00	
A3	027	45.3	24.5	0.05	0.02	
A3	D037	46.3	13.1	1.23	0.13	
				etc.		

How much more data do we need?

Classification problem

Let $D_m = (\mathbf{y}, \mathbf{X}) = \{(y_i, \mathbf{x}_i)\}_{i=1}^{m}$ denote a sample of m i.i.d random variables generated by unknown distribution $\mathbb{P}(y, \mathbf{x})$, $D_m \sim \mathbb{P}^m$,

$$\hat{y} = \arg\max_{y \in [0,1]} \mathbb{P}(y, \mathbf{x}_{\text{new}}).$$

Fix a parametric family

$$\mathcal{F} = \{f(y, \mathbf{x}, \theta) |\ \theta \in \Theta\} \text{ s. t. } \int\limits_{\{0,1\} \times \mathbb{R}^n} f(y, \mathbf{x}, \theta) dy d\mathbf{x} \equiv 1,\ \theta \in \Theta.$$

The optimal $\hat{\theta}$ maximizes approximate likelihood

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \prod_{i=1}^{m} f(y_i, \mathbf{x}_i, \theta).$$

Further inference is made with respect to $f(y, \mathbf{x}, \hat{\theta})$.

How many m^* observations $(y_i, \mathbf{x}_i)$ do we need to obtain reasonable approximation of $\mathbb{P}(y, \mathbf{x})$?

Sample size determination, frequentist approach

Let $\theta_m = \theta(D_m)$ be estimate of the parameter θ.
To estimate sample size m^*, formulate null and alternative hypothesis:

$$H_0 : \theta \in A(\theta_0), \quad H_1 : \theta \in A_1(\theta_0).$$

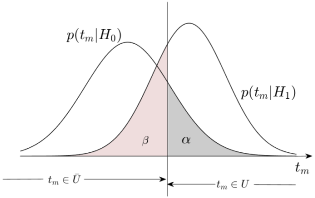
Let U be the critical area for statistics t_m for H_0 vs H_1.

Definition. The sample size m^* defines as follows: m^* s.t. for $m \geq m^*$

$$\mathsf{P}\{t_m \in U | H_0\} \geq 1 - \alpha$$

$$\text{and } \mathsf{P}\{t_m \in \bar{U} | H_1\} \leq \beta,$$

where α and β are type I and type II errors.



Example

Let $\mathbb{P}(y) = y^{\theta}(1-y)^{(1-\theta)}$, which corresponds to $y \sim B(\theta)$,

$$\theta_m = \frac{1}{m}\sum_{i=1}^{m} y_i.$$

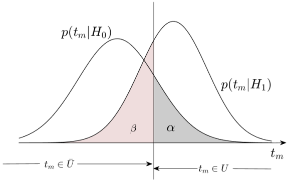
Under $H_0:\ \theta = \theta_0$ against $H_1:\ \theta \neq \theta_0$ as $m \to \infty$

$$t_m = \frac{\theta_m - \theta_0}{\sqrt{\theta_0(1-\theta_0)}}\sqrt{m} \to \mathcal{N}(0,1) \Rightarrow m^* = \frac{z^2_{\alpha/2}\theta_m(1-\theta_m)}{(\theta_m-\theta_0)^2},$$

where $z_{\alpha/2} = F_{\mathcal{N}}^{-1}(1-\alpha)$. Alternatively, with $H_1:\ \theta = \theta_1$

$$t_m|H_1 \to \mathcal{N}\left(\theta_1 - \theta_0, \frac{\theta_1(1-\theta_1)}{\theta_0(1-\theta_0)}\right),\ \text{and}$$

$$m^* = \frac{\left(z_{1-\beta}\sqrt{\theta_1(1-\theta_1)} + z_{\alpha}\sqrt{\theta_0(1-\theta_0)}\right)^2}{(\theta_1-\theta_0)^2}.$$



Statistical SSD methods

Method	Expression for m^*
1) $f(y,\mathbf{x};\theta) = y^{\theta}(1-y)^{(1-\theta)}$, 2) $H_0 : \theta = \theta_0$, $H_1 : \theta = \theta_1$, 3) Statistics: $t_m = \frac{\bar{y}-\theta_0}{\sqrt{\theta_0(1-\theta_0)}}\sqrt{m}$	$m^* = \frac{\left(z_{1-\beta}\sqrt{\theta_1(1-\theta_1)}+z_{\alpha}\sqrt{\theta_0(1-\theta_0)}\right)^2}{(\theta_1-\theta_0)^2}$
1) $f(y,\mathbf{x};p) = y^{\theta}(1-y)^{(1-\theta)}$, 2) $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$, 3) Statistics: $t_m = \frac{\bar{y}-\theta_0}{\sqrt{\theta_m(1-\theta_m)}}\sqrt{m}$	$m^* = \frac{z^2_{\alpha/2}\theta_m(1-\theta_m)}{(\theta_m-\theta_0)^2}$
1) $f(y,\mathbf{x};p) = y^{\theta}(1-y)^{(1-\theta)}$, 2) $H_0 : \ \theta-\theta_0 \leq \delta$, $H_1 : \theta-\theta_0 > \delta$, 3) Statistics: $t_m = \frac{\bar{y}-\theta_0}{\sqrt{\theta_0(1-\theta_0)}}\sqrt{m}$	$m^* = \frac{(z_{1-\beta}+z_{\alpha/2})^2\theta_m(1-\theta_m)}{(\lvert\theta_m-\theta_0\rvert-\delta)^2}$
1) $f(y,\mathbf{x},\theta) = y^{\sigma(\mathbf{x}^{\mathsf{T}}\theta)}(1-y)^{(1-\sigma(\mathbf{x}^{\mathsf{T}}\theta))}$, 2) $H_0 : \ \theta_j = 0$, $H_1 : \theta_j \neq 0$, 3) Statistics: $t_m = 2\ln\frac{f(D_m,\theta)}{f(D_m,\theta_0)}$	$m^* = \frac{\gamma_m}{\Delta^*}$, where $\gamma_m : \ \chi^2_{n,\beta}(\gamma_m) = \chi^2_{p,\alpha}$, $\Delta^* = \mathrm{E}_{\mathbf{X}}\left[-\mathbf{X}(\theta-\theta_0)\sigma(\mathbf{X}\theta) - \ln\left(\frac{\sigma(\mathbf{X}\theta_0)}{\sigma(\mathbf{X}\theta)}\right)\right]$
1) $f(y,\mathbf{x},\theta) = y^{\sigma(\mathbf{x}^{\mathsf{T}}\theta)}(1-y)^{(1-\sigma(\mathbf{x}^{\mathsf{T}}\theta))}$, 2) $H_0 : \ \theta_j = 0$, $H_1 : \theta_j \neq 0$, 3)Statistics: $t_m = \frac{w-w_0}{\sqrt{\mathrm{var}[\theta]}}\sqrt{m}$	$m^* = \frac{\left(\sqrt{V_1}z_{1-\beta}-\sqrt{V_0}z_{\alpha/2}\right)^2}{(\theta-\theta_0)^2}$

Bayesian sample size determination

Instead of parameter estimates, focus on parameter distributions $p(\theta|D, f) \equiv p(\theta|D)$.

Let $p(\theta)$ be the prior for θ, then $p(\theta|D_m) \propto f(y, \mathbf{x}, \theta)p(\theta)$.

The triplet $\langle \ell, \mathbb{P}, \xi \rangle$ defines the criterion $T(m)$ as

$$T(m) = \mathbb{I}\left[\int L(\mathbf{y}, \mathbf{X}) \prod_{i=1}^{m} \mathbb{P}(y_i, \mathbf{x}_i) dy_i d\mathbf{x}_i \le \xi\right],$$

where $\mathbb{I}[\cdot]$ is the indicator function, $L(\mathbf{y}, \mathbf{X})$ is the expectation of $\ell(\mathbf{y}, \mathbf{X}, \theta)$ with respect to $p(\theta|\mathbf{y}, \mathbf{X})$.

Definition. The sample size m^* is called *sufficient* according to posterior criterion T, if $T(m)$ holds for all $m \ge m^*$.

This definition allows $m^* = \infty$.

Computation of $T(m)$

According to Bayes' rule, model approximation of $\mathbb{P}(y, \mathbf{x})$ is given by

$$p(D_m) = \prod_{i=1}^{m} \int f(y_i, \mathbf{x}_i, \theta) p(\theta) d\theta.$$

Compute posterior criteria $T(m)$ using $p(D_m)$

$$T(m) = \mathbb{I}\left[\int L(\mathbf{y}, \mathbf{X}) p(D_m) \prod_{i=1}^{m} dy_i d\mathbf{x}_i \leq \xi\right].$$

or use the sample mean instead of integration:

$$T(m) = \mathbb{I}\left[\frac{1}{K} \sum_{k=1}^{K} L(\mathbf{y}_k, \mathbf{X}_k) \leq \xi\right],$$

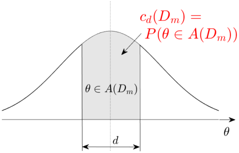
where $D_m^{(k)} = (\mathbf{y}_k, \mathbf{X}_k) \sim p^m(D_m)$.

Average posterior criteria: $T(m) \leftarrow \langle \ell, p(D_m), \xi \rangle$

- Average coverage criterion (ACC):

  Ensure that coverage probability

$$c_d(D_m) = \mathrm{P}(\theta | \theta \in A(D_m)) = \int_{\theta \in A(D_m)} p(\theta | D_m) d\theta$$

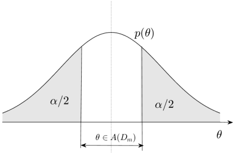


  exceeds the threshold: $T(m) = \mathbb{I}[\mathrm{E}_{D_m}(1 - c_d(m)) \leq \xi]$.

- Average length criterion (ALC):

$$T(m) = \mathbb{I}[\mathrm{E}_{D_m} |A(D_m)| \leq \xi],$$

  where $c_d(D_m) = \alpha$.



- Average posterior variance criterion (APVC):
  $T(m) = \mathbb{I}[\mathrm{E}_{D_m} V(D_m) \leq \xi]$.

Relation between ACC, ALC and APVC

From Chebyshev inequality:

$$\mathrm{P}\big(|\theta - \mathrm{E}(\theta|D_m)| < d\big) \geq \frac{\mathrm{var}[\theta|D_m]}{4d^2},$$

which is equivalent to

$$c_d(D_m) \geq 1 - \frac{V(D_m)}{4d(D_m)^2}.$$

Relation	Fixed	Controled	Follows
APVC ⇒ ACC	$d(D_m) = d$	$V(D_m) \to 0$	$c_d(D_m) \to 1$
APVC ⇒ ALC	$c_d(D_m) = 1 - \xi < 1$	$V(D_m) \to 0$	$d(D_m) \to 0$
ACC ⇒ APVC	$d(D_m) = d$	$c_d(D_m) \to 1$	$V(D_m) \to 0$
ALC ⇒ APVC	$c_d(D_m) = 1 - \xi < 1$	$d(D_m) \to 0$	$V(D_m) \to 0$

Interaction between ACC, ALC and APVC

From Chebyshev inequlity:

$$\mathrm{P}\left(|\theta - \mathrm{E}(\theta|D_m)| < d\right) \geq \frac{\mathrm{var}[\theta|D_m]}{4d^2},$$

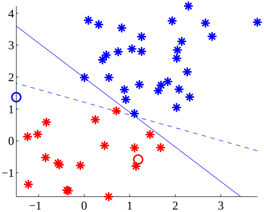
which is equivalent to

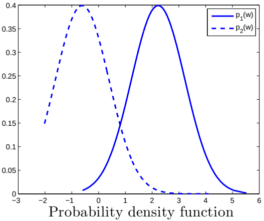
$$c_d(D_m) \geq 1 - \frac{V(D_m)}{4d(D_m)^2}.$$

Relation	Fixed	Controled	Follows
APVC $\Rightarrow$ ACC	$d(D_m) = d$	$V(D_m) \to 0$	$c_d(D_m) \to 1$
APVC $\Rightarrow$ ALC	$c_d(D_m) = 1 - \xi < 1$	$V(D_m) \to 0$	$d(D_m) \to 0$
ACC $\Rightarrow$ APVC	$d(D_m) = d$	$c_d(D_m) \to 1$	$V(D_m) \to 0$
ALC $\Rightarrow$ APVC	$c_d(D_m) = 1 - \xi < 1$	$d(D_m) \to 0$	$V(D_m) \to 0$

Average KL-divergence criterion

Small variation of data sample D_m leads to significant change of model parameters θ and posterior probability estimates $p(\theta|D_m)$.





Let $\mathsf{KL}(m)$ denote the expected KL-divergence

$$D_{\mathsf{KL}}\big(p(\theta|D_m), p(\theta|D_{m-1})\big) = \int p(\theta|D_m)\ln\frac{p(\theta|D_m)}{p(\theta|D_{m-1})}d\theta$$

between the posterior distributions:

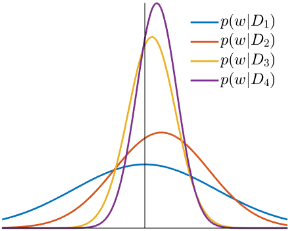
$$\mathsf{KL}(m) = \mathsf{E}_{D_m} D_{\mathsf{KL}}\big(p(\theta|D_m), p(\theta|D_{m-1})\big).$$

Motivation for average KL-divergence criterion

In Bayesian statistics, KL-divergence between posterior and prior is used as measure of information gain when prior $p(\theta)$ is updated to $p(\theta|D)$.

Consider a sequence of prior updates:

$$p(\theta) \to p(\theta|D_1) \to \dots$$
$$p(\theta|D_{m-1}) \to p(\theta|D_m)$$



The AKLC observes average information gain from D_{m-1} to D_m and such sample size m^* that for all $m > m^*$ less then ξ information gain is expected.

Asymptotics of KL-divergence criterion

Consider empirical estimate $q_m(\theta) = \sum_{k=1}^{K} \delta(\theta - \theta_m^k)$ of posterior distribution $p(\theta|D_m)$.

If θ_m is an MLE at D_m, then, given consistency conditions hold,

$$\theta_m - \theta_0 \to^P \mathcal{N}\big(\mathbf{0}, I(\theta_0)\big) \text{ and } q_m(\theta) \to F_{\mathcal{N}}\big(\theta|\mathbf{0}, I(\theta_0)\big),$$

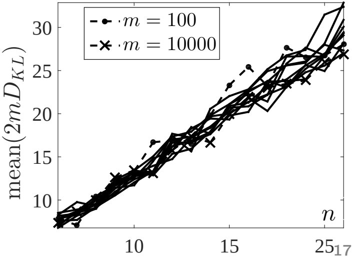
thus $\mathrm{KL}(p(\theta|D_m)||p(\theta|D_{m-1})) \approx$

$$\approx \frac{1}{2}\left[\frac{1}{m}(\theta_m - \theta_{m-1})^{\mathsf{T}}\mathbf{H}^{-1}(\theta_{m-1})(\theta_m - \theta_{m-1}) - \right.$$

$$\left. - n + \mathrm{Tr}[\mathbf{H}^{-1}(\theta_{m-1})\mathbf{H}(\theta_m)] + \ln\frac{\det\big(\mathbf{H}(\theta_{m-1})\big)}{\det\big(\mathbf{H}(\theta_m)\big)}\right]$$

As m tends to infinity,
$2m\mathrm{KL}\big(p(\theta|D_m)||p(\theta|D_{m-1})\big) \to C\chi_n^2$,

hence $2mD_{\mathrm{KL}}(m) \to Cn$

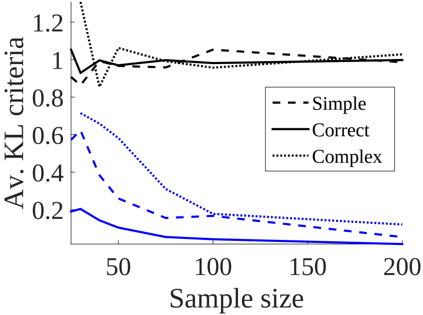


Computation of average KL criterion

0. Fix $\mathbb{P}$, $\mathcal{F}$ and the number K of samples D_m, used to perform numerical integration.
1. For each $m = 1, \dots, M$ generate K samples $D_m^{(k)} \sim \mathbb{P}^m$, $k = 1, \dots, K$.
2. For each $D_m^{(k)}$ generate a sample of posterior parameters $p(\theta|D_m^{(k)})$ using $p(\theta)$ and $f(y, \mathbf{x}, \theta)$. Compute $T(D_m^{(k)})$.
3. Average the values of $T(D_m^{(k)})$ over $k = \dots 1, \dots, K$.

$\mathbb{P}$: data $(\mathbf{y}, \mathbf{X})$, $\mathbf{x} \in \mathbb{R}^{20}$ is linearly separable with $\mathbf{x}_{1:5} = [x_{i1}, \dots, x_{i5}]^\mathsf{T}$.

Simple model:
$f(y, \mathbf{x}, \theta) \equiv f(y, x_1, \theta)$.
Correct model:
$f(y, \mathbf{x}, \theta) \equiv f(y, \mathbf{x}_{1:5}, \theta)$.
Complex model: $f(y, \mathbf{x}, \theta)$.



Conclusion

- A new criterion for bayesian sample size determination was formulated.
- The proposed criterion is based on the minimizing divergence between posterior distributions of model parameters.
- The proposed criterion attempts to generalize the existing criteria.
- Convergence and applicability of the proposed criteria were demonstrated.