How to optimize sample in active learning : Dispersion, an optimum criterion for classification ?

B. GANDAR, G. LOOSLI & G. DEFFUANT

We want to generate a sample adapted to active classification learning.
There exist theoretical arguments showing that discrepancy is a criterion of quality of a learning set for active functions learning (regression). We show that these theoretical arguments do not apply to classification and we propose dispersion as a new indicator. We give theoretical and experimental arguments in favour of this indicator.

Motivations

We consider function f, from I^s to $\mathbb{R}$ in regression problems, from I^s to $\{-1,1\}$ in classification problems. We want to choose a learning set of size n $\{(x_1,f(x_1)),(x_2,f(x_2)),\ldots,(x_n,f(x_n))\}$, allowing us to approximate as closely as possible function f with function $\hat{f}$, obtained from a learning algorithm (for instance minimising empirical risk with some regularisation criteria).
For any function $\tilde{f}$ of a hypothesis space, the generalisation error $L(\tilde{f})$ is defined by: $L(\tilde{f}) = \int \left|\tilde{f} - f\right| d\mu(x)$, and we experimentally estimate it by: $\hat{L}(\tilde{f}) = \frac{1}{n}\sum_{i=1}^{n}\left|\tilde{f}(x_i) - f(x_i)\right|$.

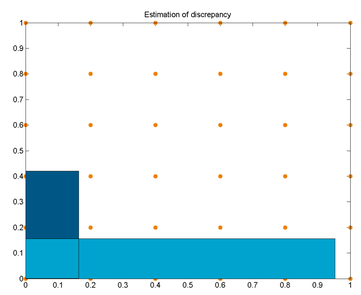
Notations

- I^s: unit cube in dimension s
- I^{s*}: set of all hyper-rectangles of I^s containing the origin
- d: euclidean distance
- λ: Lebesgue measure
- #: operator which, for a sequence $(x(n))$ with n elements, and a set P, gives the number of elements of $x(n)$ in set P

Discrepancy and low discrepancy sequences

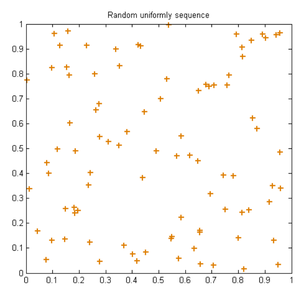
Discrepancy of $x_{(n)}$
$= D_n^*(x) = \sup_{P\in I^{s*}} \left|\frac{\#(P,(x_{(n)}))}{n} - \lambda(P)\right|$

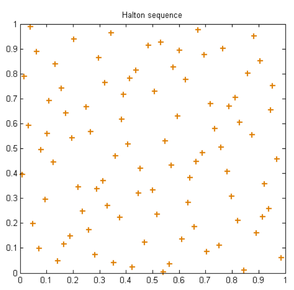
Discrepancy is defined by the maximal deviation on all the hyper-rectangles of I^s containing the origin between the proportion of points of the sequence in a hyper-rectangle and the volume of this hyper-rectangle. It is a measure of good " uniformity " of the sequence.
A *sequence is said to be low discrepancy* if its discrepancy is inferior to $\frac{1}{\sqrt{n}}$.



Estimation of discrepancy.

A regular grid has a discrepancy about $\frac{1}{\sqrt{n}}$, discrepancy of a low discrepancy sequence is about $O\left(\frac{log^s(n)}{n}\right)$.





(a) Uniformly random. (b) Low discrepancy (Halton).
Two sequences of 90 points.

Low discrepancy sequence are optimum for regression problem:

Applying Koksma-Hlawka theorem, we obtain: $\left|L\left(\tilde{f}\right) - \hat{L}\left(\tilde{f}\right)\right| \leq V_{HK}\left(\left|\tilde{f} - f\right|\right) D^*(X)$ where $V_{HK}(g)$ is the variation of the function g in the sense of Hardy-Krause.

- Contrary to classical learning theory, convergence obtained is **deterministic** and its speed is the same as discrepancy: $O\left(\frac{log^s(n)}{n}\right)$.
- We don't need a 0 empirical risk to obtain a convergence of $O\left(\frac{1}{n}\right)$ instead of $O\left(\frac{1}{\sqrt{n}}\right)$.
- The condition to be in finite VC-dimension is substituted by a condition to be in finite variation (...)

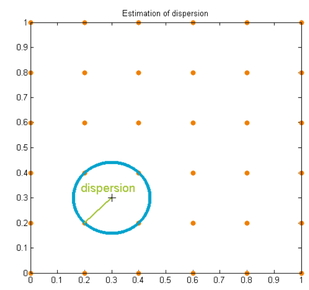
Theoretical arguments used for regression do not apply to classification

In classification problems, function f is so an indicator function, i.e. it takes its values in $\{-1,1\}$. In the majority of cases, its Hardy-Krause variation is infinite, $\Longrightarrow$ last inequality is unexploitable.
Moreover, experimentally, we note that classification using SVMs yields better results with regular grids than low discrepancy sequences.
$\Longrightarrow$ **Hence it seems that discrepancy is not the relevant criterion to choose learning samples in classification problems.**

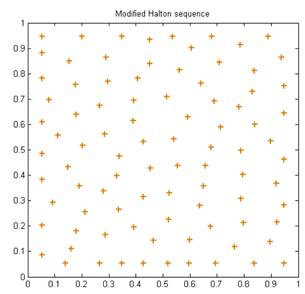
Dispersion of a sequence

Dispersion of $x_{(n)}$
$= \delta(x_{(n)}) = \max_{y\in I^s} \min_{i=1,\ldots,n} d(y,x_i)$

Dispersion is the radius of the biggest empty ball of the space.
Shukarev grid minimizes dispersion when the number of points is appropriate.



Estimation of dispersion.



Low dispersion sequence with 90 points.

For a classification problem, the generalisation error is linked to the dispersion of the sample, not to its discrepancy:

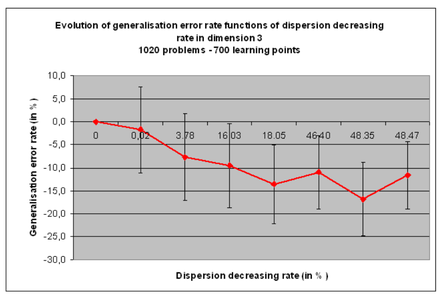
Let $\chi_{f^+} = \{x \in I^s | f(x) = +1\}$ et $\chi_{f^-} = \{x \in I^s | f(x) = -1\}$.
We suppose that f has this propriety of regularity: $\exists R$ tel que $\forall x \in \chi_{f^+}, \exists x_0 \in \chi_{f^+} | x \in B(x_0,R)$ and $B(x_0,R) \subset \chi_{f^+}$ and $\forall x \in \chi_{f^-}, \exists x_0 \in \chi_{f^-} | x \in B(x_0,R)$ and $B(x_0,R) \subset \chi_{f^-}$

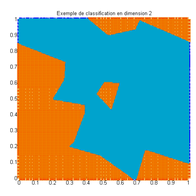
Let the learning algorithm A approximing the function f by $A(E) = \hat{f}$ as: $\hat{f}(x) = \begin{cases} +1 & \text{si } \forall x_i^- \in E \cap \chi_{f^-}, d(x_i^-,x) \geq 2\delta_{x_{(n)}}. \\ -1 & \text{si } \forall x_i^+ \in E \cap \chi_{f^+}, d(x_i^+,x) \geq 2\delta_{x_{(n)}}. \\ \text{random} & \text{otherwise} \end{cases}$

Then procedure A provides an approximation of f with a generalisation error $L(A(E))$ such that: $L(A(E)) < cte_f * \delta_{x_{(n)}}$.

Numerical experiments and conclusion

The numerical experiments involve artificially generated indicator functions, and we used SVMs as learning technique.
In each experiment, we generate the learning set of low dispersion for a given indicator function, run the SVM and then evaluate the generalisation error. We iterate this process while decreasing dispersion of the sample. For each dispersion level, we use several different indicator functions.





Example of classification in dimension 2.

Conclusion:

- generalisation error decreases with dispersion for classification
- a small decrease of dispersion may lead to a significant decrease of generalisation error

Prospects:

- strengthen experimental results, increasing the number of simulations and the dimension
- verify these results with others classification methods
- **derive an active learning method using these results**

Cemagref Science, water & land management
Centre de Clermont-Ferrand
Laboratoire d'Ingénierie pour les Systèmes Complexes
wwwlisc.clermont.cemagref.fr
www.isima.fr/limos
benoit.gandar@cemagref.fr

ENBIS 2009 Conference
Saint-Etienne, Fr
1-3 July 2009

LIMOS
Université Blaise Pascal