Estimates of the learning set size for k-NN and IINC methods in HEP

František Hakl

SPMS 2018

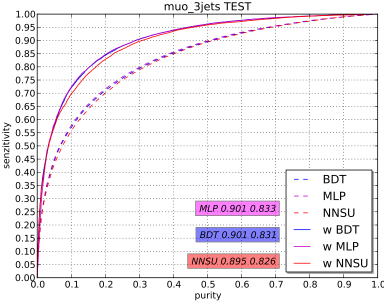
hakl@cs.cas.cz

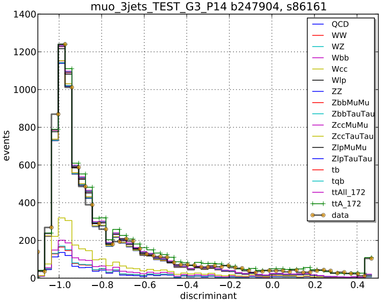
Institute of computer science, Prague

Jun 2018

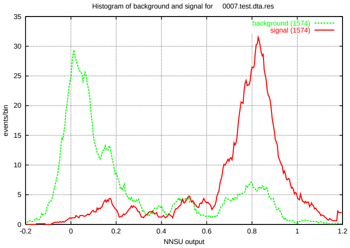
Data driven verification use data to test

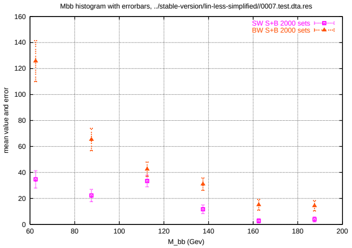
- over-learning of training method, e.g. resemble behavior on learn, train and test data
- results robustness on different portions of data (cross-validation)
- resemblance of discriminant distribution over different data sets or over simulated and measured data (so called control plots)

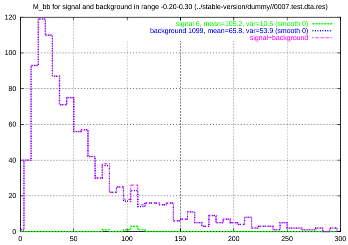




- validity of apriori known statistical characteristics of data (enhanced by cross-validation)







These approaches do not provide an estimation of convenient size of data sets and information about expected accuracy of separation.

Probably approximately correct learning

Definition ((ϵ, δ)-learning algorithm in PAC model)

1. $\mathrm{e}_{\widetilde{P}}\left(\bar{h}, \bar{c}\right) \stackrel{\text{def}}{=} \widetilde{P}\left(\bar{c} \triangle \bar{h}\right) = \widetilde{P}\left(\left(\bar{c} \dot{-} \bar{h}\right) \cup \left(\bar{h} \dot{-} \bar{c}\right)\right)$
2. $\bar{h}$ is consistent if and only if $\{x_i, \ldots, x_m\} \cap \left(\bar{c} \triangle \bar{h}\right) = \emptyset$
3. $\bar{S}_{\mathrm{C}}$ denote the set of all samples $\left(\breve{x}, \vec{\boldsymbol{z}}\right)$ of fixed $\bar{c} \in \mathrm{C}$, where $\vec{\boldsymbol{z}} \in \{-1, +1\}^m$, $\breve{x} \in \bar{X}^m$, $m \in Z$.
4. (ϵ, δ)-LEARNING ALGORITHM is each mapping $\widetilde{\mathrm{A}^*} : \bar{S}_{\mathrm{C}} \to \mathrm{C}$ such that for all $\bar{c} \in \mathrm{C}$, $\epsilon, \delta \in (0, 1)$ and $\widetilde{P}$ on $\bar{X}$, the probability of the set

$$\left\{ \breve{x} \,\middle|\, \left(\breve{x}, \vec{\boldsymbol{z}}\right) \text{ is } m\text{-sample of } \bar{c} \text{ and } \mathrm{e}_{\widetilde{P}}\left(\bar{c}, \widetilde{\mathrm{A}^*}\left(\left(\breve{x}, \vec{\boldsymbol{z}}\right)\right)\right) \geq \epsilon \right\}$$

is smaller than the number δ.
5. VC-dimension: Let $\bar{X}$ be arbitrary set, $\mathrm{C} \subset 2^{\bar{X}}$ and

$$\Pi_{\mathrm{C}}(m) \stackrel{\text{def}}{=} \max_{\bar{A} \subset \bar{X}, |\bar{A}| = m} \left|\left\{ \bar{b} \,\middle|\, (\exists \bar{c} \in \mathrm{C})\left(\bar{b} = \bar{A} \cap \bar{c}\right) \right\}\right|$$

Then

$$\mathrm{VC}_{dim}(\mathrm{C}) \stackrel{\text{def}}{=} \sup\left\{ m \,\middle|\, \Pi_{\mathrm{C}}(m) = 2^m \right\}.$$

Theorem (main result of PAC theory)

Let C *satisfy* $(\exists \bar{c}_1, \bar{c}_2 \in \mathrm{C})\,(\bar{c}_1 \neq \bar{c}_2$ *and* $(\bar{c}_1 \cap \bar{c}_2 \neq \emptyset$ *or* $\bar{c}_1 \cup \bar{c}_2 \neq \bar{X}))$ *and* C *be well-behaved . Then:*

1. *If* $\mathrm{VC}_{dim}(\mathrm{C}) < +\infty$*. Then*
    1. *for any* $0 < \epsilon < \frac{1}{2}$ *there is no* (ϵ, δ)*-learning algorithm with number of queries less than*

    $$\max\left(\frac{1-\epsilon}{\epsilon}\ln\left(\frac{1}{\delta}\right), \mathrm{VC}_{dim}(\mathrm{C}) \cdot (1 - 2(\epsilon(1-\delta)+\delta))\right) . \quad (1)$$

    2. *for arbitrary* $0 < \epsilon < 1$*, any learning algorithm using at least*

    $$\max\left(\frac{4}{\epsilon}\log_2\left(\frac{2}{\delta}\right), \frac{8\mathrm{VC}_{dim}(\mathrm{C})}{\epsilon}\log_2\left(\frac{13}{\epsilon}\right)\right) \quad (2)$$

    queries and returning a consistent hypothesis is an (ϵ, δ)*-learning algorithm.*
2. (ϵ, δ)*-learning algorithm for* C *exists* $\Leftrightarrow \mathrm{VC}_{dim}(\mathrm{C}) < +\infty$*.*

Sketch of the proof:

1. 1.
   - $\frac{1-\epsilon}{\epsilon}\ln\left(\frac{1}{\delta}\right)$: (c&c) Any nontrivial concept class can be reduced to one of the cases discussed above. For uniform probability we get a contradiction.
   - $d\left(1-2\left(\epsilon\left(1-\delta\right)+\delta\right)\right)$: (c&c) Reduce $\bar{X}$ to d-element subset with uniform probability. Then use the "matrix" $\boldsymbol{Z}_{\bar{c},\bar{h}} \stackrel{\text{def}}{=} \mathrm{e}_{\widetilde{P}}\left(\bar{c},\bar{h}\right)$ to show, that $m > d\left(1-2\left(\epsilon\left(1-\delta\right)+\delta\right)\right)$ imply that $\left(\exists \bar{h}^*\right)$ contradicts (ϵ,δ)-property ... "broadly speaking".

   2. In more steps we show that from (2) follows that

$$Prob_{\widetilde{P}}\left(\left\{x_i,\ldots,x_m\right\}\,\middle|\,\left(\forall \bar{T}\in\left\{\bar{h}\triangle\bar{c}\,\middle|\,\bar{h}\in \mathrm{H}\right\}\,\middle|\,Prob_{\widetilde{P}}\left(\bar{T}\right)>\epsilon\right)\right.$$

$$\left.\left(\left\{x_i,\ldots,x_m\right\}\cap\bar{T}=\emptyset\right)\right)\le\delta\;.$$

2.
   - $\Leftarrow$ (construction) Use Zermelo's well-ordering theorem to well-order $\bar{H}$. Let algorithm get m-sample of $\bar{c}$ and return the first hypothesis consistent with $\bar{c}$. The statement follows from 1)-2).
   - $\Rightarrow$ (by contradiction) For any $d\in N$ we carry out steps 1)-1)-(second term). Choose (ϵ,δ) such that $\left(1-2\left(\epsilon\left(1-\delta\right)+\delta\right)\right)>0$. Hence m can't be upper-bounded.

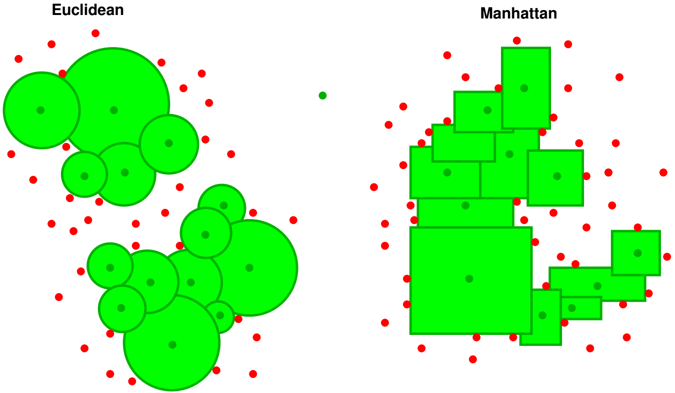
Nearest neighbor (NN) is consistent and has a known $\mathrm{VC}_{dim}(NN)$ u.b.

Lemma (Union, Intersection)

Let $\mathrm{U}_{k,\mathrm{C}} \stackrel{def}{=} \left\{ \bigcup_{i=1}^{k} \bar{c}_i \,\middle|\, \left(\forall i \in \hat{k}\right) (\bar{c}_i \in \mathrm{C}) \right\}$, $\mathrm{I}_{k,\mathrm{C}} \stackrel{def}{=} \left\{ \bigcap_{i=1}^{k} \bar{c}_i \,\middle|\, \left(\forall i \in \hat{k}\right) (\bar{c}_i \in \mathrm{C}) \right\}$ *and* $\mathrm{VC}_{dim}(\mathrm{C}) = d \geq 1$ *be finite. Then*

$$\mathrm{VC}_{dim}(\mathrm{U}_{k,\mathrm{C}}) \leq 2dk\log_2(3k) \quad \textit{and} \quad \mathrm{VC}_{dim}(\mathrm{I}_{k,\mathrm{C}}) \leq 2dk\log_2(3k)\ .$$

$\bar{X} = \Re^n$, k=number of Balls (or Rect.), $\mathrm{VC}_{dim}(Ball_n) = n+1$, $\mathrm{VC}_{dim}(Rect_n) = 2n$



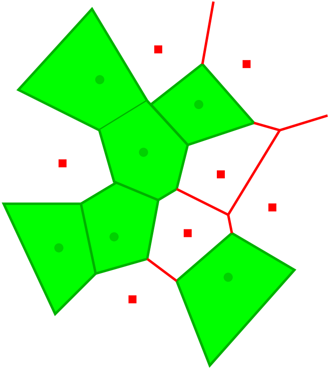
$\mathrm{VC}_{dim}(NN_{Ball_n}) \leq 2(n+1)k\log_2(3k)$, consistent, $\mathrm{VC}_{dim}(NN_{Rect_n}) \leq 4nk\log_2(3k)$

IINC algorithm is consistent and has a known $\mathrm{VC}_{dim}(IINC)$ upper bound

Lemma

Let $\bar{X}$ *be an arbitrary set,* $\mathrm{C} \subset 2^X$*. Then*

1. *if any two sets in* C *are disjoint then* $\mathrm{VC}_{dim}(\mathrm{C}) = 1$,
2. $\mathrm{VC}_{dim}(\mathrm{C}) = 1 \Rightarrow \mathrm{VC}_{dim}(\mathrm{U}_{k,\mathrm{C}}) \leq k$.



IINC outline (basic)

- for a new (unspecified) point compute distances to all k known points
- sort points by inverted distances
- put new point to the set of the first point in sorted sequence
- ... so the "green" set is an union of pairwise disjoint sets

$\mathrm{VC}_{dim}(IINC) \leq k$ and consistent hypothesis

Corollary

Let $|\{x_i, \dots, x_m\} \cap \bar{c}| = \rho m$, e.g. ρ is the ratio of positive examples. It follows (recall second lower bound $m > \frac{8\mathrm{VC}_{dim}(C)}{\epsilon}\log_2\left(\frac{13}{\epsilon}\right)$, n is dimension of examples):

NN Euclidean:	$1 > \rho \cdot 16(n+1)$	$\times$	$\log_2(3\rho m) \times \left[\frac{1}{\epsilon}\log_2\left(\frac{13}{\epsilon}\right)\right]$
NN Manhattan:	$1 > \rho \cdot 32n$	$\times$	$\log_2(3\rho m) \times \left[\frac{1}{\epsilon}\log_2\left(\frac{13}{\epsilon}\right)\right]$
IINC:	$1 > \rho \cdot 8$	$\times$	$\left[\frac{1}{\epsilon}\log_2\left(\frac{13}{\epsilon}\right)\right]$

(note that the first constrain implies $\log_2(3\rho m) > \log_2\left(\frac{12\rho}{\epsilon}\log_2\left(\frac{2}{\delta}\right)\right)$)

Discussion

- unusable for "large" values of ρ (e.g. $\rho \simeq \epsilon$)
- dimension of examples can be considered constant; corresponds to the number of relevant and reasonable features
- for NN ρ should be proportionate to $\dfrac{\text{desired accuracy of separation } (\epsilon)}{\text{logarithm of positive examples}}$
- applicable in the case of very rare positive examples

Example of HEP data set size
(source Measurement of Electroweak Top Quark Production at D∅, Yun-Tse Tsai, Rochester, 2013)

Pre-tagged event yields

	Run IIa, 1 fb^{-1}				Run IIb, 8.7 fb^{-1}			
	Electron Channel		Muon Channel		Electron Channel		Muon Channel	
	2 jets	3 jets	2 jets	3 jets	2 jets	3 jets	2 jets	3 jets
Signals								
tb	20	8.1	20	9.4	158	39	133	34
Background Sum	14962	3586	18610	5125	78502	11526	72382	11192
Background + Signal	15021	3611	18672	5156	78941	11642	72764	11294
Data	15021	3611	18672	5156	78936	11641	72762	11293

Table 5.13 Pre-tagged event yields after selection.

Estimated range of ρ for selected processes:

Top Quark Production at D∅	$\rho \in \langle 0.001, 0.003 \rangle$
Higgs boson search at ATLAS, LHC	$\rho \simeq 10^{-4} - 10^{-6}$
NOvA: muon antineutrinos $\to$ electron antineutrinos (press release, June 4, 2018)	18 events over three years

Conclusion

- method of learn data size estimation is suggested for very rare processes
- upper bound of the the Vapnik-Chervonenkis dimension for consistent nearest neighbor and IINC like methods is derived
- set size estimation is applicable in applications in which the ratio ρ of positive examples is extremely small

$$\text{NN:} \qquad \rho \lessapprox \frac{const. \cdot dim(\bar{X}) \cdot \epsilon}{\log(\text{\# of pos. examples})}$$

$$\text{IINC:} \qquad \rho \lessapprox const. \cdot \epsilon$$