Applied Model Selection and Error Estimation: Some New Results and Open problems

Third Italian Workshop on Machine Learning and Data Mining
Workshop of the XIII AI*IA Symposium on Artificial Intelligence

Luca Oneto, Alessandro Ghio, Davide Anguita

University of Genoa

DIBRIS

DITEN

SmartLab

December 10, 2014

luca.oneto@unige.it - www.lucaoneto.com

Outline

Introduction: Model Selection and Error Estimation

Out–Of–Sample Methods (Hold–Out)

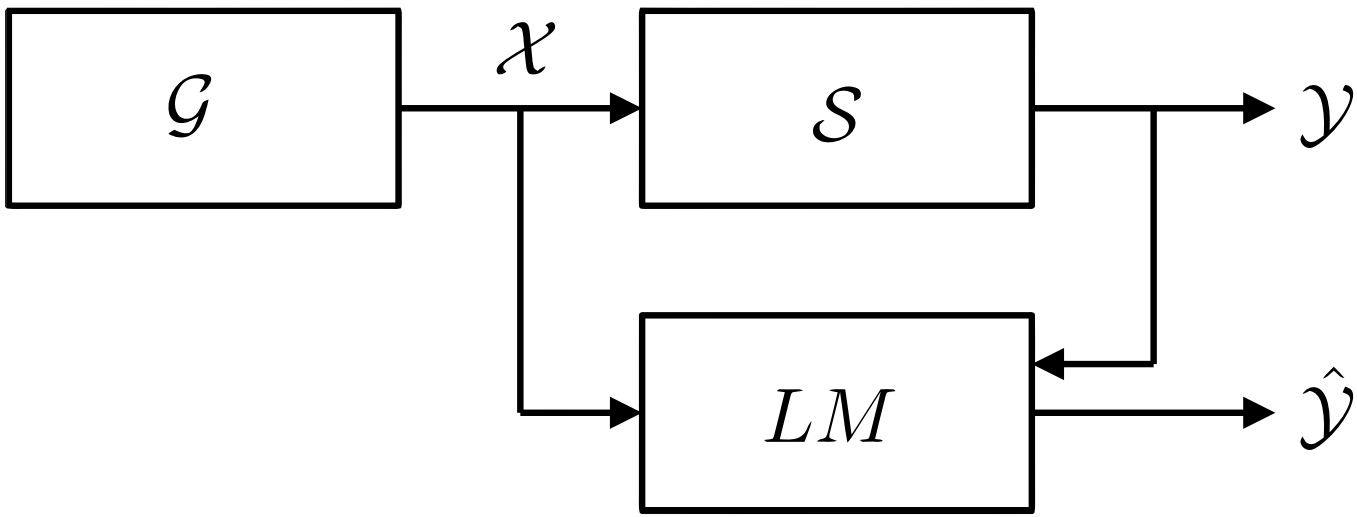
In–Sample Methods
- Hypotheses Space–Based Methods
- Algorithm–Based Methods

Final Remarks

Outline

Supervised Learning Framework[1]

- $\mathcal{G}$ Data Generator
- $\mathcal{D}$ Supervisor
- *LM* Learning Machine
- Generalization (True) Error $L(f) = \mathbb{E}_{(X,Y)}\ell\left(f(X), Y\right)$
- Empirical Error $\hat{L}(f) = \frac{1}{n}\sum_{i=1}^{n}\ell\left(f(X_i), Y_i\right)$



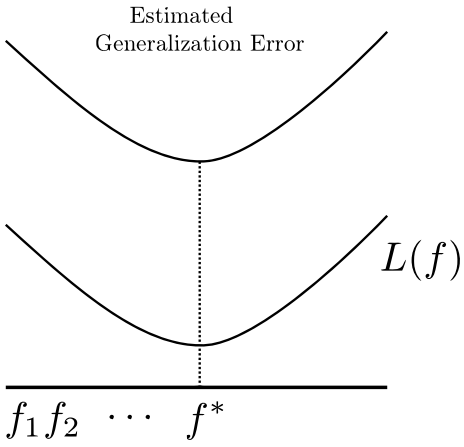
[1] Vapnik V. N. Statistical learning theory.1998

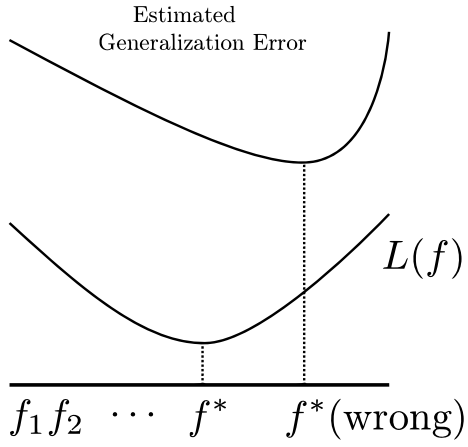
Model Selection and Error Estimation (I)

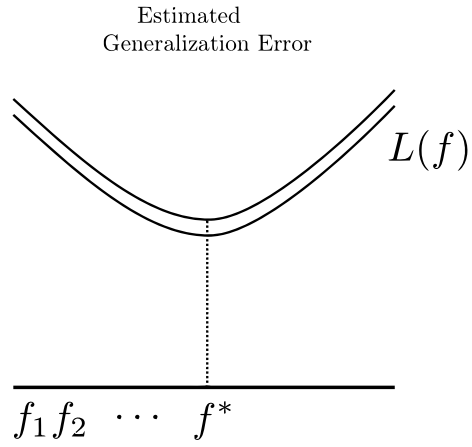
Model Selection is the effort to obtain a function f with the smallest possible $L(f)$ where $L(f)$ is obviously unknown since μ is unknown.
(Selecting the number of hidden neurons in Multi Layer Perceptrons, the regularization parameter in Support Vector Machines or the depth of a Decision Tree)

Error Estimation is the effort to obtain the rigorously tightest possible estimate (with high probability) of $L(f)$

The two problems seem to be two faces of the same coin. The truth is that Model Selection is an easier task respect to Error Estimation.







Model Selection and Error Estimation (II)[2,3]

- **Out–Of–Sample Methods (Hold–Out)**: the idea behind this approach is to divide the data in different parts (Test Set, K-Fold Cross-Validation, Bootstrap, etc.). These methods are quite general and can be used for every possible Learning Algorithm.

- **In–Sample Methods**: all the data are used for learning the model f, for model selection (selecting the best hyperparameters of the algorithms or the right hypothesis space) and for error estimation.

  - **Hypotheses Space–Based Methods**: these methods require to know explicitly the hypotheses space from which we choose f

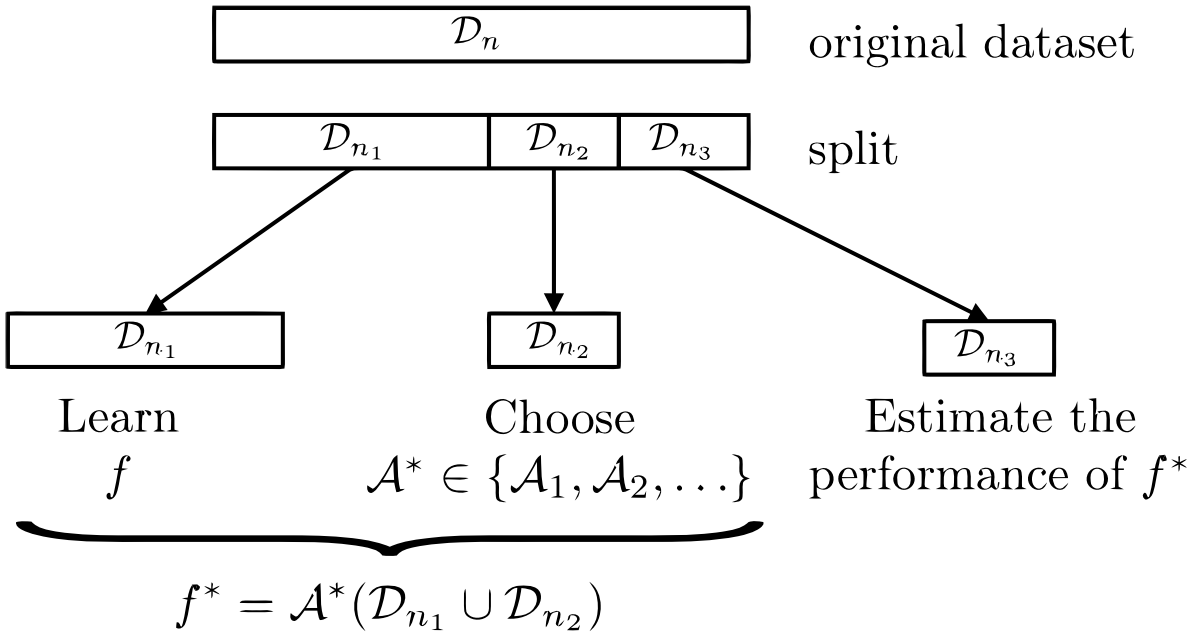
  - **Algorithm–Based Methods**: these methods are quite general and can be used for every possible Learning Algorithm

[2] Anguita D. Ghio A. Oneto L. Ridella S. In–sample and out-of-sample model selection and error estimation for support vector machines. IEEE Transactions on Neural Networks and Learning Systems. 2012

[3] Oneto L. Ghio A. Anguita D. Ridella S. Fully Empirical and Data–Dependent Stability–Based Bounds. IEEE Transactions on Cybernetics. –

Outline

Out–Of–Sample Methods



The key is the I.I.D. hypothesis

Model Selection: if the model is good, the error over previously unseen data ($\mathcal{D}_{n_2}$) should be small

Error Estimation: since the data in $\mathcal{D}_{n_3}$ are I.I.D. with respect to $\mathcal{D}_{n_1} \cup \mathcal{D}_{n_2}$, the errors over the sample in $\mathcal{D}_{n_3}$ are I.I.D. so we can rigorously estimate $L(f)$ (Clopper–Pearson, Hoeffding, etc.)

Out–Of–Sample Methods: Open Problems (And Some Solutions)

- How to deal with small–sample setting ($n \leq 100$)?[4]
- Are the state–of–the–art bounds really the theoretical limit?[5]
- Does an Optimal splitting procedure exist?[6]
  - Size of the different sets
  - Randomized splitting
  - Stratified splitting
  - Nearly Homogeneous Splitting
- Why the methods work so well in practice?
  - No theoretical justification
  - The error of the Bootstrap procedure is not completely understood
  - Leave One Out works well in practice for Model Selection. For Error Estimation it is completely unreliable.
- Bootstrap or Cross Validation?[7]

---

[4] Anguita D. Ghio A. Greco N. Oneto L. Ridella S. Model selection for support vector machines: Advantages and disadvantages of the machine learning theory. IEEE International Joint Conference Neural Networks. 2010

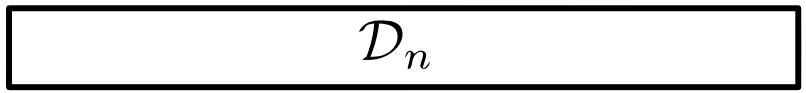
[5] Anguita D. Ghelardoni L. Ghio A. Ridella S. Test error bounds for classifiers: A survey of old and new results. IEEE Symposium on Foundations of Computational Intelligence. 2011

[6] Anguita D. Ghelardoni L. Ghio A. Oneto L. Ridella S. The 'K' in K-fold Cross Validation. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. 2012

[7] Anguita D. Ghio A. Oneto L. Ridella S. In–sample and out-of-sample model selection and error estimation for support vector machines. IEEE Transactions on Neural Networks and Learning Systems. 2012

Outline

In–Sample Methods[8,9] (I)



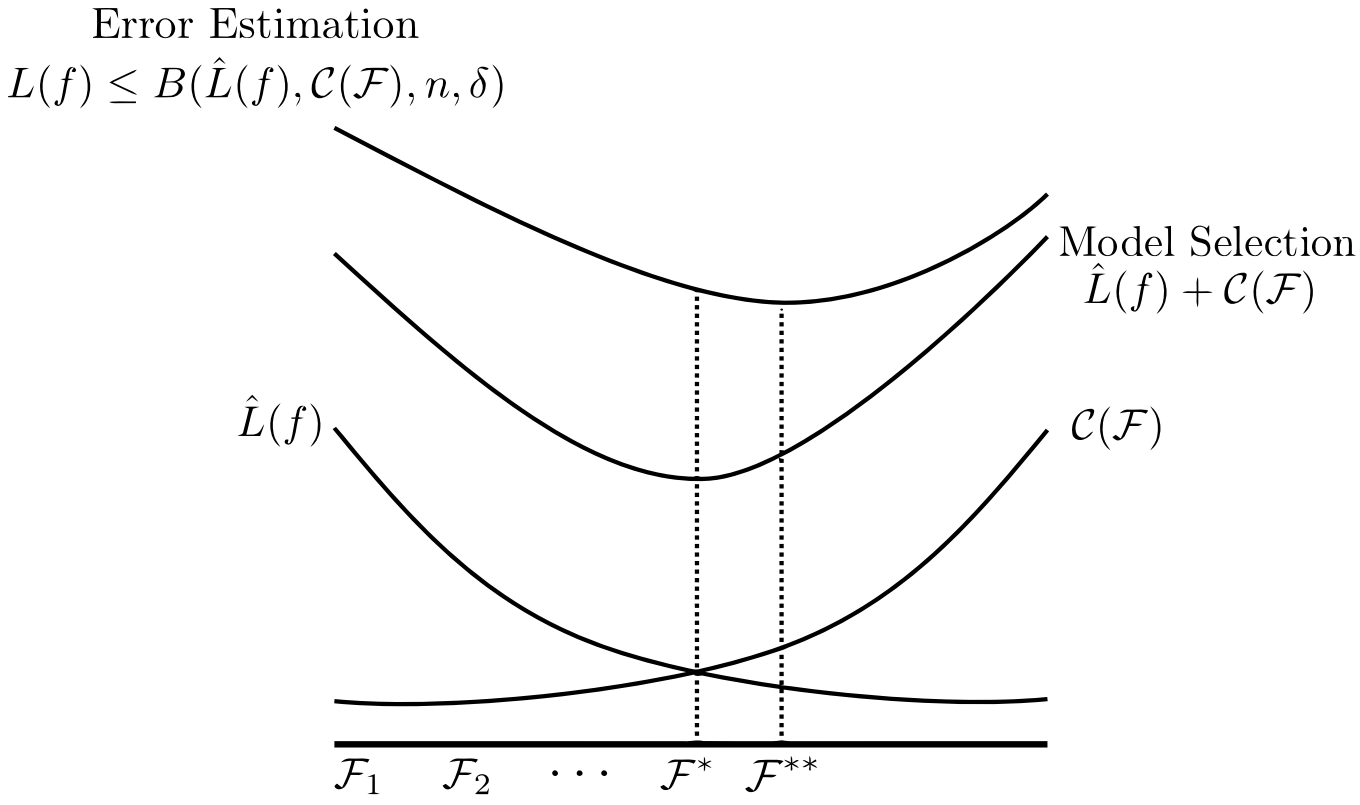
train
model selection
error estimation

Problem:
Since we use the data for different purposes, the errors over the different samples are not I.I.D. anymore even if the the data in $\mathcal{D}_n$ are I.I.D.

[8] Vapnik V. N. Statistical learning theory.1998

[9] Poggio T. Rifkin R. Mukherjee S. Niyogi P. General conditions for predictivity in learning theory. Nature. 2004

In–Sample Methods (II)



$\mathcal{F}$ **can be:** the number of hidden neuron in MLP, the depth of a Decision Three, a space of Functions, etc.

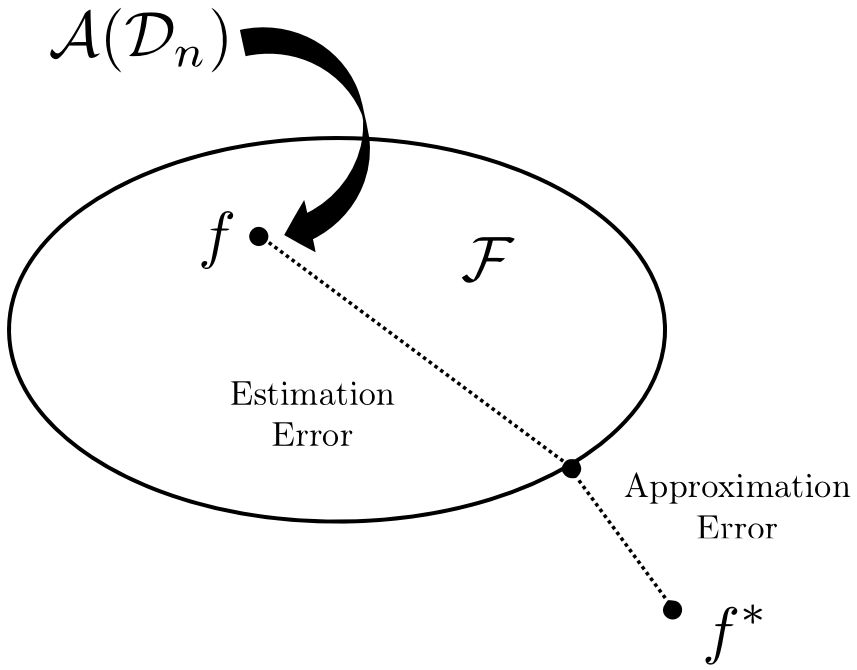
Outline

Hypotheses Space–Based Methods[10]

We need to define the hypotheses space from which the algorithm will choose the model.

NOTE: For many algorithms it may become computationally intractable:

- K–NN
- Decision Tree
- SVM
- etc.

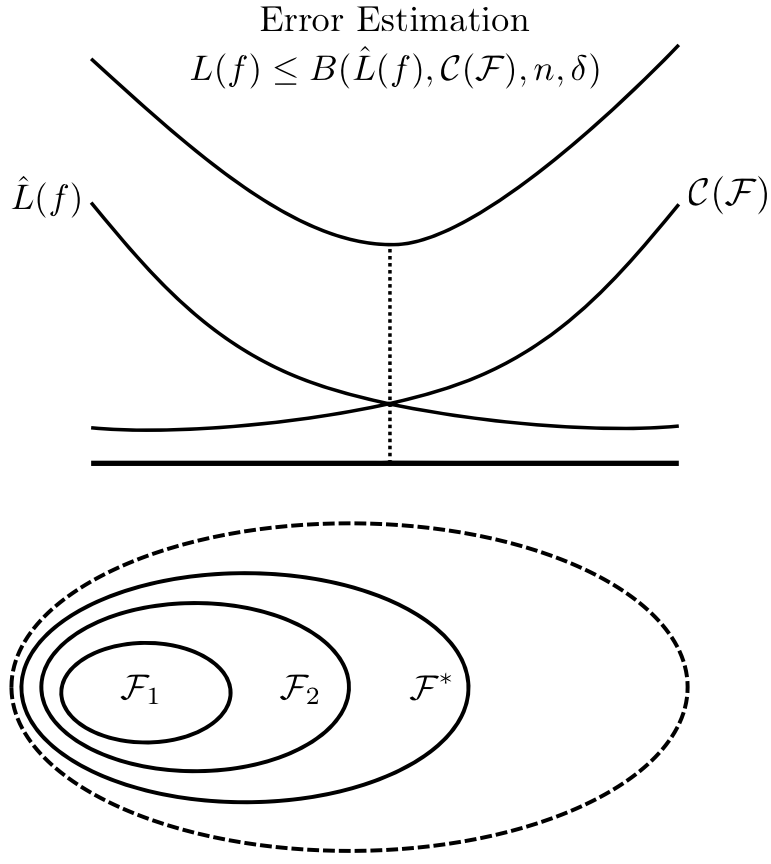


[10] Vapnik V. N. Statistical learning theory.1998

Structural Risk Minimization (SRM):

Different Approach to estimate $\mathcal{C}(\mathcal{F})$

- Vapnik–Chervonenkis dimension
  - Data–independent
  - Takes into account functions that will be never piked up during the learning procedure
- Rademacher Complexity
  - Data–dependent (on $\mathbb{P}(\mathcal{X})$ not on $\mathbb{P}(\mathcal{Y}|\mathcal{X})$)
  - Takes into account functions that will be never piked up during the learning procedure
- Local Rademacher Complexity
  - Data–dependent (on $\mathbb{P}(\mathcal{X})$ and on $\mathbb{P}(\mathcal{Y}|\mathcal{X})$)
  - Takes into account functions with small empirical error
- PAC Bayes
  - Data–dependent (on $\mathbb{P}(\mathcal{X})$ and on $\mathbb{P}(\mathcal{Y}|\mathcal{X})$)
  - Takes into account functions with small empirical error



Open Problems (And Some Solutions)

- Does not work for all the algorithms (Decision Three, k–NN, etc.[11])
- We need to improve the Concentration Inequalities[12]
- Hard to compute in practice[13]
- Still loose in practice

Can we exploit these methods for optimizing the amount of resources needed for learning?

LINK - VIDEO

[11] Bousquet O. Andre E. Stability and generalization. The Journal of Machine Learning Research. 2002

[12] Boucheron S. Gabor L. Pascal M.. Concentration Inequalities: A nonasymptotic theory of independence. Oxford University Press. 2013

[13] Anguita D. Ghio A. Oneto L. Ridella S. A Deep Connection Between the Vapnik–Chervonenkis Entropy and the Rademacher Complexity. IEEE Transaction on Neural Networks and Learning Systems. 2014

Resource Aware Learning (I)[14]

- Reproducing kernel Hilbert space

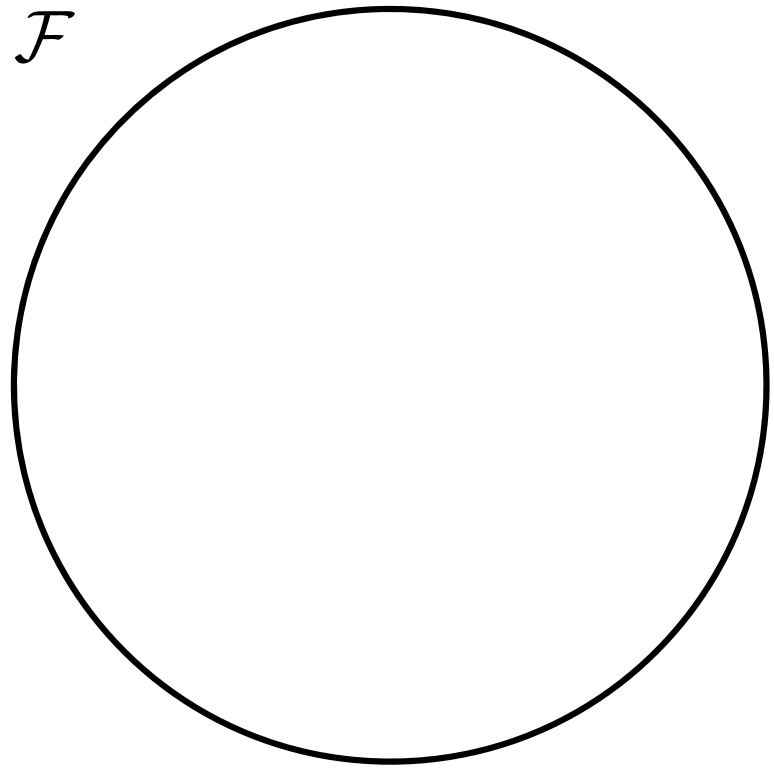
$$f(\boldsymbol{x}) = \boldsymbol{w}^T \phi(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha K(\boldsymbol{x}_i, \boldsymbol{x})$$

- Regularization (limiting the space of functions)

$$\|\boldsymbol{w}\|_2^2 \leq w_{\text{MAX}}$$

- Empirical Risk Minimization

$$\inf_{f \in \mathcal{F}} \hat{L}_n(f)$$



[14] L. Oneto, A. Ghio, D. Anguita, and S. Ridella. Learning resource-aware models for mobile devices: from regularization to energy efficiency. Neurocomputing, 2014.

Resource Aware Learning (II)[15]

- Bound on the True Error

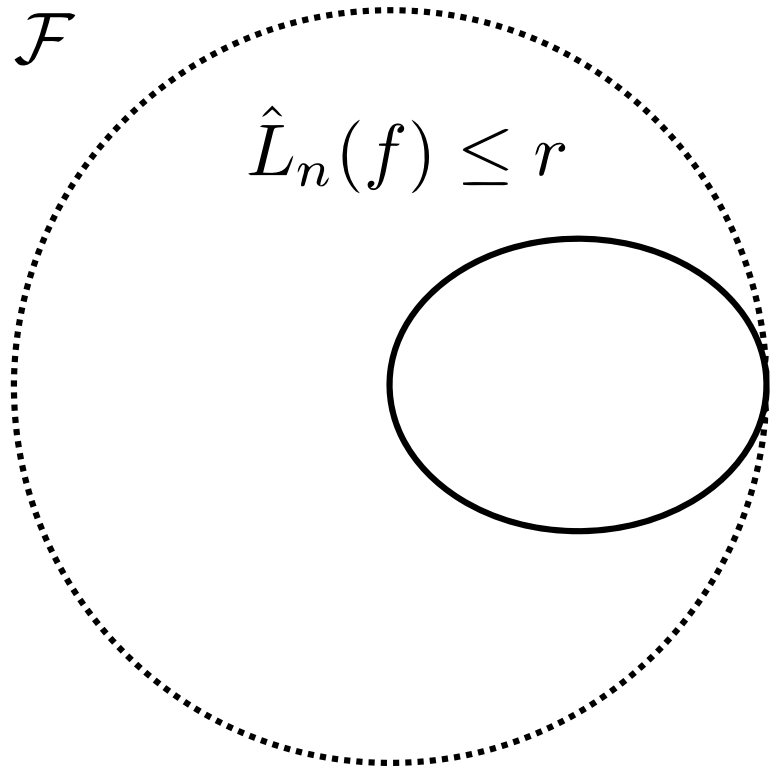
$$L(f)$$

- Local Rademacher Complexity

$$\hat{L}_n(f) \leq \rho$$

- SRM with Local Rademacher Complexity

$$\mathcal{F}^* \in \{\mathcal{F}_1, \mathcal{F}_2, \ldots\}$$

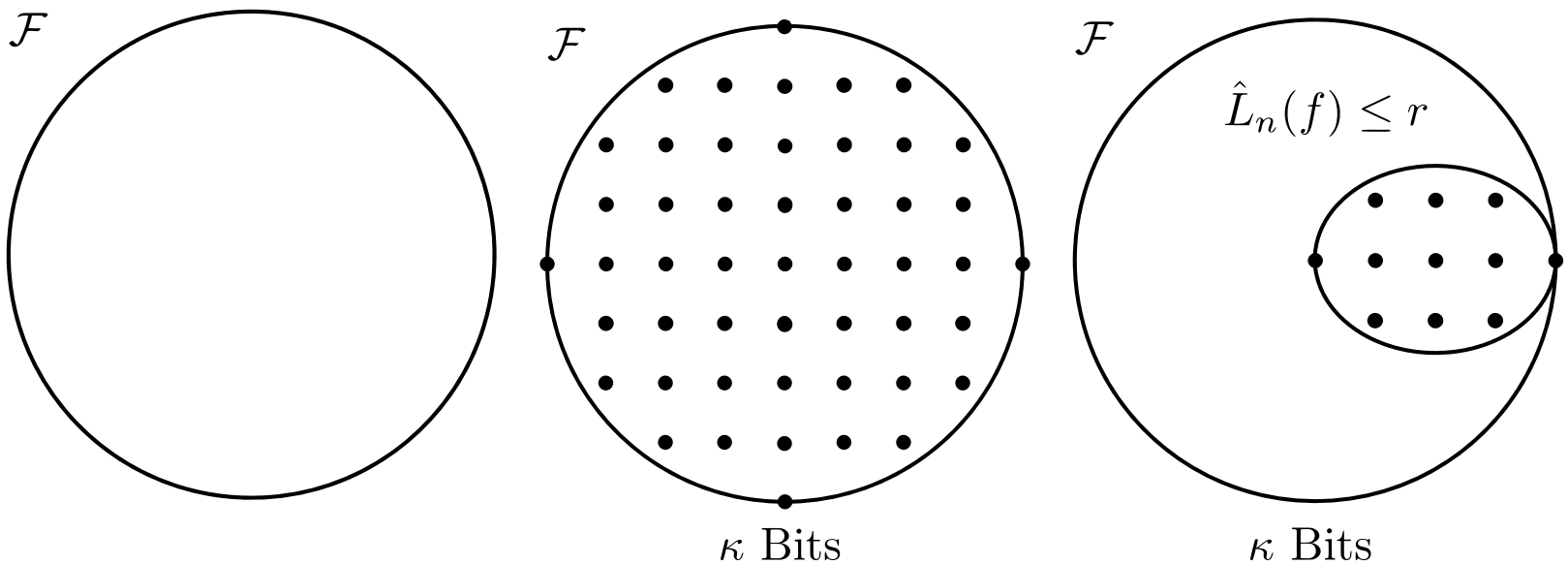


[15] L. Oneto, A. Ghio, D. Anguita, and S. Ridella. Learning resource-aware models for mobile devices: from regularization to energy efficiency. Neurocomputing, 2014.

Resource Aware Learning (III)[16]

Reducing the number of bits for representing $\boldsymbol{w}$

$$w_j \in \frac{w_{\text{MAX}}}{2^\kappa - 1} \{-2^\kappa + 1, \ldots, 2^\kappa - 1\}$$

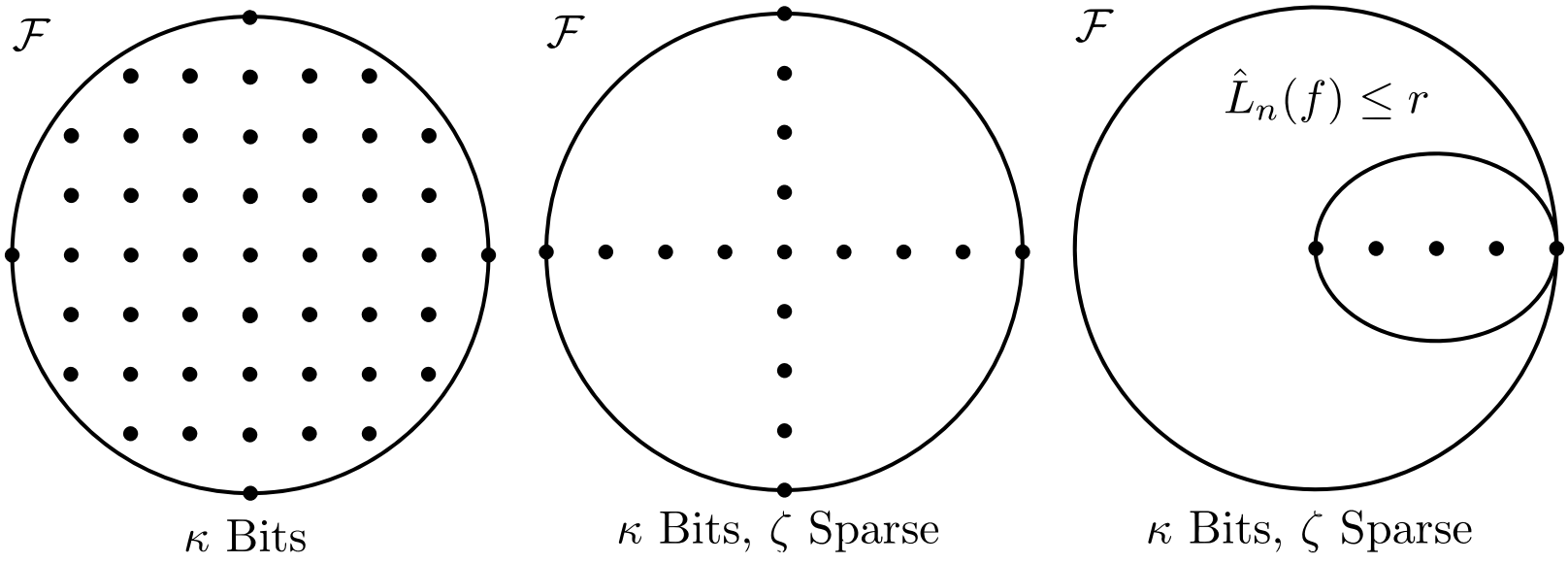


[16] L. Oneto, A. Ghio, D. Anguita, and S. Ridella. Learning resource-aware models for mobile devices: from regularization to energy efficiency. Neurocomputing, 2014.

Resource Aware Learning (IV)[17]

Increasing the sparsity of $\boldsymbol{w}$

$$\sum_{i=1}^{D} [w_j \neq 0] \leq D\zeta$$



[17] L. Oneto, A. Ghio, D. Anguita, and S. Ridella. Learning resource-aware models for mobile devices: from regularization to energy efficiency. Neurocomputing, 2014.

Resource Aware Learning (V)[18]

Human Activity Recognition Using Smartphones Data Set

Download: Data Folder, Data Set Description

Abstract: Human Activity Recognition database built from the recordings of 30 subjects performing activities of daily living (ADL) while ca

Data Set Characteristics:	Multivariate, Time-Series	**Number of Instances:**	10299	**Area:**	Computer
Attribute Characteristics:	N/A	**Number of Attributes:**	561	**Date Donated**	2012-12-10
Associated Tasks:	Classification, Clustering	**Missing Values?**	N/A	**Number of Web Hits:**	131283

[18] D. Anguita, A. Ghio, L. Oneto, X. Parra, J. L. Reyes-Ortiz, A Public Domain Dataset for Human Activity Recognition using Smartphones, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), pp. 437-442, Bruges, Belgium, 24-26 Apr. 2013.

Resource Aware Learning (VI)[19]

Standing vs Sitting

n	ζ	κ	*Error*
25	0.5	1	21.3
100	0.5	1	11.7
225	0.5	4	7.7
400	0.5	8	7.2
625	0.5	8	7.1
900	0.5	8	7.0

[19] D. Anguita, A. Ghio, L. Oneto, X. Parra, J. L. Reyes-Ortiz, A Public Domain Dataset for Human Activity Recognition using Smartphones, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), pp. 437-442, Bruges, Belgium, 24-26 Apr. 2013.

Outline

Algorithm–Based Methods

These methods do not require the knowledge of the Space of Function $\mathcal{F}$ from which the Algorithm $\mathcal{A}$ will choose.

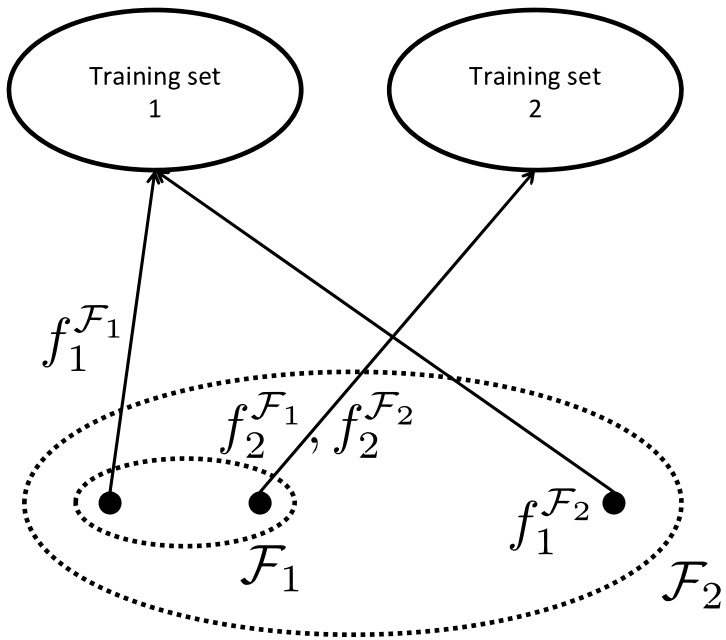
$$\mathcal{A} : \mathcal{D}_n \rightarrow f$$

Different Approach to estimate $\mathcal{C}(\mathcal{A})$

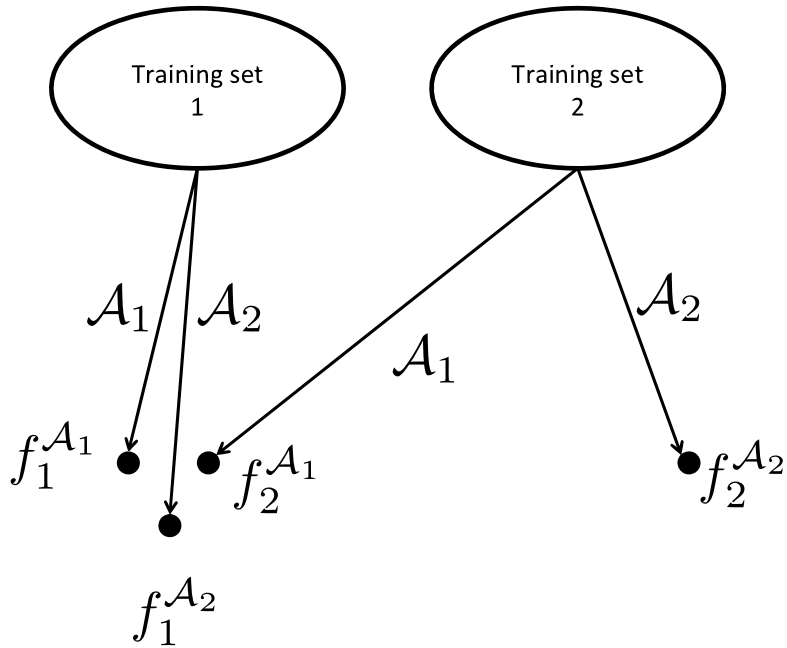
- Algorithmic Stability: how much the solution found by the algorithm is affected by changing of just one sample?
- Compression Bound: how much data can be removed from the training set without affecting the performance of the learning algorithms?

Algorithmic Stability[20,21] (I)

Structural Risk Minimization



Stability



[20] Bousquet O. Andre E. Stability and generalization. The Journal of Machine Learning Research. 2002

[21] Poggio T. Rifkin R. Mukherjee S. Niyogi P. General conditions for predictivity in learning theory. Nature. 2004

Algorithmic Stability (II)

Approach 1: Hypothesis Stability

$$H_{\text{emp}}\left(\mathcal{A}_{\mathcal{F}}, n\right) = \mathbb{E}_{\mathcal{D}_n, Z_i'}\left|\ell\left(\mathcal{A}_{(\mathcal{D}_n,\mathcal{F})}, Z_i\right) - \ell\left(\mathcal{A}_{\left(\mathcal{D}_n^i,\mathcal{F}\right)}, Z_i\right)\right| \leq \beta_{\text{emp}}$$

$$H_{\text{loo}}\left(\mathcal{A}_{\mathcal{F}}, n\right) = \mathbb{E}_{\mathcal{D}_n, Z}\left|\ell\left(\mathcal{A}_{(\mathcal{D}_n,\mathcal{F})}, Z\right) - \ell\left(\mathcal{A}_{\left(\mathcal{D}_n^{\setminus i},\mathcal{F}\right)}, Z\right)\right| \leq \beta_{\text{loo}}$$

Bound on the Generalization Error

$$L\left(\mathcal{A}_{(\mathcal{D}_n,\mathcal{F})}\right) \leq \hat{L}_{\text{emp}}\left(\mathcal{A}_{(\mathcal{D}_n,\mathcal{F})}, \mathcal{D}_n\right) + \sqrt{\frac{1}{2n\delta} + \frac{3\beta_{\text{emp}}}{\delta}}$$

$$L\left(\mathcal{A}_{(\mathcal{D}_n,\mathcal{F})}\right) \leq \hat{L}_{\text{loo}}\left(\mathcal{A}_{(\mathcal{D}_n,\mathcal{F})}, \mathcal{D}_n\right) + \sqrt{\frac{1}{2n\delta} + \frac{3\beta_{\text{loo}}}{\delta}}$$

Algorithmic Stability (III)

Approach 2: Uniform Stability

$$U^i\left(\mathcal{A}_{\mathcal{F}}, n\right) = \left|\ell\left(\mathcal{A}_{(\mathcal{D}_n,\mathcal{F})}, \cdot\right) - \ell\left(\mathcal{A}_{\left(\mathcal{D}_n^i,\mathcal{F}\right)}, \cdot\right)\right|_\infty \leq \beta^i$$

$$U^{\setminus i}\left(\mathcal{A}_{\mathcal{F}}, n\right) = \left|\ell\left(\mathcal{A}_{(\mathcal{D}_n,\mathcal{F})}, \cdot\right) - \ell\left(\mathcal{A}_{\left(\mathcal{D}_n^{\setminus i},\mathcal{F}\right)}, \cdot\right)\right|_\infty \leq \beta^{\setminus i}$$

Bound on the Generalization Error

$$L\left(\mathcal{A}_{(\mathcal{D}_n,\mathcal{F})}\right) \leq \hat{L}_{\text{emp}}\left(\mathcal{A}_{(\mathcal{D}_n,\mathcal{F})}, \mathcal{D}_n\right) + 2\beta^i + \left(4n\beta^i + 1\right)\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}},$$

$$L\left(\mathcal{A}_{(\mathcal{D}_n,\mathcal{F})}\right) \leq \hat{L}_{\text{loo}}\left(\mathcal{A}_{(\mathcal{D}_n,\mathcal{F})}, \mathcal{D}_n\right) + \beta^{\setminus i} + \left(4n\beta^{\setminus i} + 1\right)\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2n}}.$$

Open Problems (And Some Solutions)

- We need to improve and Concentration Inequalities[22,23]
- Loose in practice

Uniform Stability is data independent and (almost) algorithmic independent.
Hypothesis Stability is unknown for kernel-based algoritms.

[22] Hoeffding W. Probability inequalities for sums of bounded random variables. Journal of the American statistical association. 1963

[23] Boucheron S. Gabor L. Pascal M.. Concentration Inequalities: A nonasymptotic theory of independence. Oxford University Press. 2013

Fully Empirical Algorithmic Stability Based Bound (I)[24]

Hypothesis

$$H^{\text{loo}}(\mathcal{A}_{\mathcal{F}}, n) \leq H^{\text{loo}}(\mathcal{A}_{\mathcal{F}}, \sqrt{n}/2)$$

We point out that the above property is a desirable requirement for any learning algorithm: in fact, the impact on the learning procedure of removing samples from $\mathcal{D}_n$ should decrease, on average, as n grows.

[24] L. Oneto, A. Ghio, D. Anguita, and S. Ridella. Fully empirical and data-dependent stability-based bounds. IEEE Transactions on on Cybernetics, 2014.

Fully Empirical Algorithmic Stability Based Bound (II)[25]

Thanks to this hypothesis we can estimate the Hypothesis Stability based on the empirical data.

$$\hat{H}^{\text{loo}}(\mathcal{A}_{(\mathcal{D}_{\sqrt{n}/2},\mathcal{F})},\mathcal{D}_{\sqrt{n}/2})=\frac{8}{n\sqrt{n}}\sum_{k=1}^{\sqrt{n}/2}\sum_{j=1}^{\sqrt{n}/2}\sum_{i=1}^{\sqrt{n}/2}\left|\ell(\mathcal{A}_{(\check{\mathcal{D}}^k_{\sqrt{n}/2},\mathcal{F})},\check{\boldsymbol{z}}^k_j)-\ell(\mathcal{A}_{((\check{\mathcal{D}}^k_{\sqrt{n}/2})^{\setminus i},\mathcal{F})},\check{\boldsymbol{z}}^k_j)\right|$$

where:

$$\check{\mathcal{D}}^k_{\sqrt{n}/2}:\left\{\boldsymbol{z}_{(k-1)\sqrt{n}+1},\ldots,\boldsymbol{z}_{(k-1)\sqrt{n}+\sqrt{n}/2}\right\},\ \check{\boldsymbol{z}}^k_j:\boldsymbol{z}_{(k-1)\sqrt{n}+\sqrt{n}/2+j},\ k\in\{1,\ldots,\sqrt{n}/2\}$$

then:

$$H^{\text{loo}}(\mathcal{A}_{\mathcal{F}},\sqrt{n}/2)\leq\hat{H}^{\text{loo}}(\mathcal{A}_{(\mathcal{D}_{\sqrt{n}/2},\mathcal{F})},\mathcal{D}_{\sqrt{n}/2})+\sqrt{\log(1/\delta)/\sqrt{n}}$$

[25] L. Oneto, A. Ghio, D. Anguita, and S. Ridella. Fully empirical and data-dependent stability-based bounds. IEEE Transactions on on Cybernetics, 2014.

Fully Empirical Algorithmic Stability Based Bound (III)[26]

Bound on the true error

$$L(f) \leq \hat{L}_n^{\text{loo}}(f) + \sqrt{\frac{2}{\delta}\left[\frac{1}{2n} + 3\left(\hat{H}^{\text{loo}}(\mathcal{A}_{(\mathcal{D}_{\sqrt{n}/2},\mathcal{F})}, \mathcal{D}_{\sqrt{n}/2}) + \sqrt{\frac{\log(\frac{2}{\delta})}{\sqrt{n}}}\right)\right]}.$$

Method for comparing the performance of an algorithm or hyperparameters

$$\mathcal{A}^*_{(\mathcal{D}_n,\mathcal{F})} = \arg \min_{\mathcal{A}_{\mathcal{F}} \in \{\mathcal{A}^1_{\mathcal{F}^1_1}, \mathcal{A}^1_{\mathcal{F}^1_2}, \ldots, \mathcal{A}^2_{\mathcal{F}^2_1}, \mathcal{A}^2_{\mathcal{F}^2_2}, \ldots\}} E(\mathcal{A}_{(\mathcal{D}_n,\mathcal{F})}),$$

$$E(\mathcal{A}_{(\mathcal{D}_n,\mathcal{F})}) = \hat{L}_n^{\text{loo}}(f) + \sqrt{\frac{2}{\delta}\left[\frac{1}{2n} + 3\left(\hat{H}^{\text{loo}}(\mathcal{A}_{(\mathcal{D}_{\sqrt{n}/2},\mathcal{F})}, \mathcal{D}_{\sqrt{n}/2}) + \sqrt{\frac{\log(\frac{2}{\delta})}{\sqrt{n}}}\right)\right]}.$$

[26] L. Oneto, A. Ghio, D. Anguita, and S. Ridella. Fully empirical and data-dependent stability-based bounds. IEEE Transactions on on Cybernetics, 2014.

Fully Empirical Algorithmic Stability Based Bound (IV)[27]

Covertype

$n = 25$				$n = 50$				$n = 100$				$n = 200$				$n = 400$			
KCV	LOO	STAB	MARG	KCV	LOO	STAB	MARG	KCV	LOO	STAB	MARG	KCV	LOO	STAB	MARG	KCV	LOO	STAB	MARG
7	**7**	**21**	**4**	**9**	**7**	**20**	**2**	**8**	**8**	**17**	**6**	**9**	**7**	**21**	**5**	**12**	**8**	**17**	**6**

Mnist OVO

$n = 25$				$n = 50$				$n = 100$				$n = 200$				$n = 400$			
KCV	LOO	STAB	MARG	KCV	LOO	STAB	MARG	KCV	LOO	STAB	MARG	KCV	LOO	STAB	MARG	KCV	LOO	STAB	MARG
16	**13**	**45**	**9**	**8**	**11**	**43**	**8**	**10**	**11**	**41**	**5**	**19**	**10**	**43**	**0**	**18**	**11**	**35**	**4**

Not-Mnist OVO

$n = 25$				$n = 50$				$n = 100$				$n = 200$				$n = 400$			
KCV	LOO	STAB	MARG	KCV	LOO	STAB	MARG	KCV	LOO	STAB	MARG	KCV	LOO	STAB	MARG	KCV	LOO	STAB	MARG
16	**14**	**45**	**15**	**13**	**13**	**45**	**11**	**12**	**12**	**42**	**11**	**15**	**14**	**42**	**11**	**21**	**16**	**36**	**11**

[27] L. Oneto, A. Ghio, D. Anguita, and S. Ridella. Fully empirical and data-dependent stability-based bounds. IEEE Transactions on on Cybernetics, 2014.

Outline

Final Remarks[28,29]

New results:

- SRM $\rightarrow$ Energy Aware Learning
- Fully Empirical Algorithmic Stability

Open Problems:

- A long list...

[28] L. Oneto, A. Ghio, D. Anguita, and S. Ridella. Learning resource-aware models for mobile devices: from regularization to energy efficiency. Neurocomputing, 2014.

[29] L. Oneto, A. Ghio, D. Anguita, and S. Ridella. Fully empirical and data-dependent stability-based bounds. IEEE Transactions on on Cybernetics, 2014.