Model Selection and Automatic Model Selection for Statistical Learning: A Comparative Study on Local Factor Analysis

Lei Xu & Lei Shi, Dept Comp Sci & Eng, Chinese Univ Hong Kong lxu@cse.cuhk.edu.hk

Given a paremetric model, the task of statistical learning consists of a parameter learning part for determining unknown parameters and a model selection part for selecting an appropriate scale for a model that accommodates these parameters. Typically, the two tasks are implemented in a two-phase procedure. First, a number of models of a same architecture but in different scales are enumerated, with the unknown parameters estimated via the maximum likelihood (ML). Second, one of typical learning theories, being different from a ML principle, is applied to select the best model. There are four major types of theories are available in the literatures, including (a) AIC and extensions (Akaike,1974; Bozdogan&Ramirez,1988; Cavanaugh, 1997), (b) Bayesian approach related criteria, i.e., BIC (Schwarz, 1978), or equivalently MML (Wallace, 1966, 1999) and MDL (Rissanen, 1986, 1989), (c) the cross validation based criteria (Stone, 1978; Rivals & Personnaz, 1999), and (d) Vapnik SRM based error bound (Vapnik, 1977, 1995).

A two-phase implementation is very computaionally extensive and thus impractical in many real applications. Alternatively, efforts have been made on seeking model selection during parameter learning. One type is incremental approaches, i.e., as the scale increases from k to k+1, parameter learning is made incrementally with the parts already learned kept or partially adjusted such that redundant computing can be saved. The increamntal process is stoped by a criterion. It usually leads to a suboptimal performance because not only those newly added parameters but also the old parameter set have to be relearned. Oppositely, making learning decremnetally may also be a choice. However, decreaing the scale from k to k-1 can not be made by simply discarding those extra parameters while all the remaining parameters have to be re-learned, i.e., an entire process of parameter learning has to be implemented at the scale k-1. That is, it has no difference from a two-phase implementation.

Another direction to explore is that model selection can be implemented automatically during parameter learning, in a sense that parameter learning (on a model with its scale large enough to include the correct one) will not only determine parameters but also automatically shrink its scale to an appropriate one, while those extra substructures are discarded during learning. One effort is Rival Penalized Competitive Learning, which was heuristically proposed on a bottom level (i.e., a level of learning dynamics or updating rule), which is quite different two-phase implementing approaches that uses a learning theory to guide model selection in a top-down manner. Bayesian Ying-Yang (BYY) harmony learning is such a global level theory that guides various statistical learning tasks with model selection achieved automatically during parameter learning.

The above approaches have been studied on this or that specific task in certain specific cases. However, there is seldomly a systematic comparative study on all these approaches though it is important for applications and further development of model selection studies. One reason is the difficulty of getting a benchmark model such that not only it is typical in the literatures and practical to real world applications but also the criteria and/or algorithms for implementing the approaches are either available already or easy to be developed. A quite popular topic in the past decade, namely local factor analysis (LFA) or a mixture of factor analysis, is choosen as this benchmark task here. Either directly applying the existing criteria and/or algorithms for LFA or further extending those from factor analysis, we are ready to make a systematic comparative experimental study.

Topic: learning theory Preference: oral/poster

Considering the approaches for this study, we include AIC and its modification consistent AIC (CAIC), BIC or equivalently MDL, cross-validation (CV) (mainly 5 fold and 10 fold). Moreover, effort has also been made on comparing with a VC-dimension based SRM error bound. After an extensive search of the existing literature, only one criterion has been found for selecting k on a Gaussian mixture (Wang&Feng, 2005), while there is no criterion available for LFA yet. We extend the criterion for LFA. Also, comparisons have made with two typical incremental approaches, namely an incremental mixture of factor analyzer (IMoFA) (Salah & Alpaydin, 2004) and Variational Bayes. Furthermore, we implement BYY-C (i.e., BYY harmony learning via a two stage implementation to link with those criteria) and BYY-A (i.e., the BYY learning with automatic model selection to link with IMoFA and Variational Bayes). Comparisons are made from the perspectives of both performances and computing times. Some examples are list below for an illustration. Many other experiments and applications on the widely used handwritten digits database MNIST are referred to the site http://appsrv.cse.cuhk.edu.hk/~shil/research_res/LFA.pdf Results on variational Bayes are not available yet but will be ready at the workshop.

	case I						case II						case III					
criteria & methods	m	k					m	k					m	k				
		1	2	3*	4	5		1	2	3*	4	5		1	2	3*	4	5
AIC	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
	2	0	0	0	0	0	2	0	0	0	0	0	2	0	0	0	1	0
	3*	0	0	76	8	3	3*	0	0	67	10	0	3*	0	0	46	15	9
	4	0	0	9	2	1	4	0	0	11	2	5	4	0	0	2	4	16
	5	0	0	0	0	1	5	0	0	0	4	1	5	0	0	0	8	0
CAIC	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
	2	0	0	3	0	0	2	0	0	7	0	0	2	0	31	18	0	0
	3*	0	4	90	0	0	3*	0	4	81	0	0	3*	0	13	66	0	0
	4	2	0	0	0	0	4	0	3	2	2	0	4	3	4	0	0	0
	5	1	0	0	0	0	5	1	0	0	0	0	5	1	0	0	0	0
BIC	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
	2	0	0	0	1	0	2	0	0	4	0	0	2	0	0	0	0	0
	3*	0	2	94	1	0	3*	0	3	83	0	1	3*	0	5	75	2	0
	4	1	0	0	1	0	4	0	5	0	4	0	4	0	4	10	0	1
	5	0	0	0	0	0	5	0	0	0	0	0	5	0	0	0	2	0
SRM	-	0	1	85	14	0	-	0	8	80	9	3	-	0	12	73	16	1
CV-5	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
	2	0	0	0	0	0	2	0	0	0	0	0	2	0	0	0	0	0
	3*	0	3	87	5	0	3*	0	3	72	8	2	3*	0	0	71	14	0
	4	0	0	3	1	0	4	0	0	7	2	5	4	0	0	5	9	1
	5	0	0	0	1	0	5	0	0	1	0	0	5	0	0	0	0	0
CV-10	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
	2	0	1	0	0	0	2	0	0	0	0	0	2	0	0	0	0	0
	3*	0	2	88	5	0	3*	0	0	76	12	0	3*	0	0	68	11	0
	4	0	0	1	1	2	4	0	0	9	0	3	4	0	0	4	12	2
	5	0	0	0	0	0	5	0	0	0	0	0	5	0	0	0	1	2
BYY-C	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
	2	0	0	0	1	0	2	0	0	1	0	0	2	0	0	2	0	0
	3*	0	1	94	2	0	3*	0	2	88	2	0	3*	0	1	86	0	0
	4	0	0	1	1	0	4	0	4	2	0	1	4	0	4	3	1	0
	5	0	0	0	0	0	5	0	0	0	0	0	5	0	3	0	0	0
IMoFA	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
	2	0	0	0	2	0	2	0	0	10	0	0	2	0	0	21	0	0
	3*	0	1	91	3	0	3*	0	3	78	0	0	3*	0	14	64	0	0
	4	0	3	0	0	0	4	0	4	2	2	0	4	0	0	0	1	0
	5	0	0	0	0	0	5	0	0	0	0	0	5	0	0	0	0	0
BYY-A	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
	2	0	0	0	1	0	2	0	0	6	0	0	2	0	0	0	0	0
	3*	0	0	93	2	0	3*	0	1	84	2	0	3*	0	0	81	12	0
	4	0	0	1	1	0	4	0	0	0	6	0	4	0	0	4	1	0
	5	0	0	0	0	2	5	0	0	1	0	0	5	0	0	0	2	0

Computational Time Cost (in minutes)					
Criteria	CV-5	CV-10	BYY-C	IMoFA	BYY-A
10.5 ± 5.7	53.7 ± 28.1	108.9 ± 62.4	13.8 ± 7.7	2.9 ± 2.6	2.8 ± 1.2

Datasets Description				
Datasets	Training	Testing	Dimensions	Classes
PEN	7,494	3,498	16	10
OPT	2,880	1,797	64	10
SEG	700	1,610	14	7
WAVE	300	4,700	21	3
ORL	400	CV10	256	2
VIS	2,700	910	169	10
YEAST	208	CV10	79	5
LVQ	1,929	1,929	20	16

CPU Time (in minutes)								
Methods	PEN	OPT	SET	WAVE	ORL	VIS	YEAST	LVQ
Criteria	171	246	232	154	98	255	192	288
ML-CV10	1692	2304	2421	1427	846	1991	1874	2412
IMoFA	26	49	45	26	14	61	38	65
BYY-A	33	41	43	25	26	49	43	35

PEN (Pendigits), OPT (Optdigits), SEG (Segment) and WAVE (Waveform) from UCI repository http://www.ics.uci.edu/mlearn/MLRepository.html
ORL from the Olivetti Research Lab http://www.cam-orl.co.uk/facedatabase.html
Vistex from MIT Media Lab (http://www.white.media.mit.edu/vismod/imageryVisionTexture/vistex.html)
Yeast gene data from (http://www.soe.ucsc.edu/research/compbio/genex/expressdata.html)
LVQ from http://www.cis.hut.fi/research/lvqpak/
CV10 generated via the 10-fold cross-validation.

Classification Accuracy								
Methods	PEN	OPT	SEG	WAVE	ORL	VIS	YEAST	LVQ
ML-AIC	94.28 ± 0.14	92.76 ± 0.32	72.48 ± 2.31	71.28 ± 1.24	98.40 ± 2.13	63.71 ± 1.39	88.93 ± 4.87	88.28 ± 0.92
ML-CAIC	95.18 ± 0.16	96.98 ± 0.49	84.08 ± 3.89	75.85 ± 2.31	98.67 ± 3.90	62.59 ± 3.02	92.46 ± 3.01	87.61 ± 0.83
ML-BIC	97.77 ± 0.13	97.82 ± 0.56	78.61 ± 2.57	82.74 ± 2.04	99.08 ± 1.57	68.68 ± 2.97	92.39 ± 6.25	90.18 ± 0.51
ML-CV10	95.22 ± 0.11	96.93 ± 0.37	82.13 ± 2.28	75.89 ± 1.87	99.04 ± 1.06	62.53 ± 3.48	92.37 ± 4.72	87.58 ± 0.22
IMoFA	97.90 ± 0.20	92.92 ± 0.89	86.11 ± 3.90	81.88 ± 2.99	98.54 ± 0.89	69.64 ± 2.12	91.88 ± 5.06	89.57 ± 0.27
BYY-A	98.87 ± 0.12	97.90 ± 0.43	88.68 ± 3.18	84.75 ± 1.99	99.17 ± 1.24	71.12 ± 1.64	95.69 ± 2.99	90.13 ± 0.32