

# Bayesian Predictive Classification, Some New Methods

Timo Koski, KTH

Lund 09/11/2011

November 8, 2011



KTH Matematik



This is joint work with Jukka Corander, Yaqiong Cui, and Jukka Sirén, all affiliated with the University of Helsinki.

- J. Corander, Y. Cui, T. Koski & J. Siren: Have I Seen You Before ? Principles of Bayesian Predictive Classification Revisited. *Statistics and Computing*, to appear (2011)
- J. Corander, Y. Cui, & T. Koski: Inductive Inference and Partition Exchangeability in Classification. Proceedings of Roy Solomonoff Conference, Lecture Notes in Computer Science (in press)
- J. Corander, Y. Cui, T. Koski & J. Siren: Predictive Gaussian Classifiers. *submitted*





*Classification* The generic problem of assigning some items into a discrete set of classes using observed characteristics (features) of the items and assessing the uncertainty related to the assignments conditional on all relevant information available.

- *supervised* classification, where all the eligible classes are *a priori* given
- *semi-supervised* classification, where a certain set of eligible classes is *a priori* determined, while the items are not forced to solely be allocated to such classes, but can also form previously unknown groups during the classification task.
- *unsupervised* classification (often referred to as *clustering*), where the classes are only identified by their contents, *i.e.* the items to be classified.



Many probabilistic classifiers may be termed *marginal* as they treat each item separately by considering its marginal probability of assignment to a particular source. In contrast, Seymour Geisser introduced (1964) a *simultaneous* probabilistic supervised classification framework where items are jointly assigned to possible sources.

The components of the new methods (as developed by us) are:

- (i) a model of *simultaneous classification* or *data labeling* defined by generating random urns and assigning data items to them,
- (ii) a predictive model for future data given training data, and
- (iii) an algorithm that can compute the sought predictive probabilities. Algorithms will not be discussed in this talk.



# Some New (?) Methods

The simultaneous predictive supervised classification principle was pioneered by Seymour Geisser (1964, 1966) and fully presented in his monograph *Predictive Inference*, Chapman & Hall, 1993, with a main focus on a Gaussian modeling. Apart from applications in speech recognition and image analysis

- Huo and C-H. Lee. A Bayesian predictive classification approach to robust speech recognition. *IEEE Transactions on Speech and Audio Processing*, 8: 200-204, 2000.
- A. Nádas. Optimal solution of a training problem in speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 33: 326-329, 1985.
- B. D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge, 1996.

seems to have received very limited attention in the statistical or machine learning literature.



By a classification structure  $S$  we refer to an unambiguous allocation of each of  $n$  items in a set  $N$  into one of a finite set of classes. Further, we let  $\mathcal{S}$  denote the space of all such structures.  $p(S)$  represents *a priori* uncertainty about the classification structure in terms of a probability distribution over the space  $\mathcal{S}$ . We shall show in the sequel how to construct  $p(S)$ .



For each of  $n$  items we have finite vectors  $\mathbf{x}_i$  of  $d$  features, such that each element  $x_{ij}$  in  $\mathbf{x}_i$  belongs to a finite alphabet,  $x_{ij} \in \mathcal{X}_j = \{1, \dots, r_j\}$ ,  $r_j \geq 1$ ,  $j = 1, \dots, d$ . The joint feature space is represented by the Cartesian product

$$\mathbf{x}_i \in \mathcal{X} = \prod_{j=1}^d \mathcal{X}_j. \quad (1)$$

The observed data for any subset  $s \subseteq N$  of items is jointly denoted as  $\mathbf{x}^{(s)}$ .



We acknowledge the existing uncertainty about a classification structure  $S$  in a probabilistic fashion, which leads to the predictive prior probability, (we can (in principle) compute this before seeing the data  $\mathbf{x}^{(N)}$ ), according to

$$p(\mathbf{x}^{(N)}) = \sum_{S \in \mathcal{S}} p(\mathbf{x}^{(N)} | S) p(S). \quad (2)$$

Here  $p(\mathbf{x}^{(N)} | S)$  is the prior predictive likelihood of the data given the classification structure  $S$  and  $p(\mathbf{x}^{(N)})$  is the expected predictive likelihood (H.V. Roberts: Probabilistic prediction, JASA, 1965).



# The Predictive Likelihood of the Data given the Classification Structure $S$

The classification structure is a partition of the set  $N$ , *ie.* an unordered collection  $S = (s_1, \dots, s_k)$ ,  $1 \leq k \leq n$ , of subsets or classes of  $N$ , such that  $s_c \cap s_{c^*} = \emptyset$ ,  $c \neq c^*$ ,  $c, c^* = 1, 2, \dots, k$ , and  $\cup_c s_c = N$  and the simultaneous predictive likelihood is

$$p(\mathbf{x}^{(N)}|S) = \int_{\Theta} p(\mathbf{x}^{(N)}|\theta, S)dF(\theta|S), \quad (3)$$

where  $p(\mathbf{x}^{(N)}|\theta, S)$  is a probability mass function for  $\mathbf{x}^{(N)}$  conditionally on a finite-dimensional parameters  $\theta$  and  $S$  and  $dF(\theta|S)$  is a prior distribution.



The observed sequences of features  $\mathbf{x}^{(s)}$  are assumed *unrestrictedly infinitely exchangeable*. This implies that, if we combine any permutation of the item values  $x_{1j}, \dots, x_{|s|j}$ , for a fixed  $j = 1, \dots, d$ , with arbitrary corresponding permutations over the remaining features, the same predictive probability mass function for  $\mathbf{x}^{(s)}$  is obtained. Furthermore, the probabilities of the sequences  $x_{1j}, \dots, x_{|s|j}$  depend on the sufficient statistics  $n_{cjl}$ ,  $l = 1, \dots, r_j$ , where  $n_{cjl}$  represents the number of copies of value  $l$  for feature  $j$  observed among the items in the class  $c$ , with  $c$  referring to an index over considered classes.



The sufficient statistics are  $n_{cjl}$ ,  $l = 1, \dots, r_j$ , where  $n_{cjl}$  represents the number of copies of value  $l$  for feature  $j$  observed among the items in the class  $c$ , with  $c$  referring to an index over considered classes.

There is a general theory of *predictive sufficient statistics*:

**parametric models are limiting forms of predictive distributions, parameters are limiting forms of predictive sufficient statistics**, with lots of technicalities of measure theoretic sort :

S. Fortini, L. Ladelli & E. Regazzini: Exchangeability, predictive distributions, and parametric models. *Sankhya: The Indian Journal of Statistics*, 2000, 62, Series A, Pt. 1, pp. 86–109.



The unique probabilistic characterization of the data under unrestricted infinite exchangeability assumed to hold over the classes  $S = (s_1, \dots, s_k)$  equals (de Finetti -type representation)

$$p(\mathbf{x}^{(N)}|S) = \int_{\Theta} \prod_{c=1}^k \prod_{j=1}^d \prod_{l=1}^{r_j} \theta_{cjl}^{n_{cjl}} dF(\theta|S), \quad (4)$$

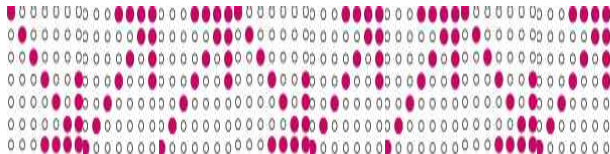
where  $\theta$  denotes jointly the probabilities  $(\theta_{c11}, \dots, \theta_{c1r_1}, \dots, \theta_{cd1}, \dots, \theta_{cdr_d})_{c=1, \dots, k}$  and  $F(\theta|S)$  is a probability measure over the space  $\Theta$ , which can be interpreted to represent our prior beliefs about the limits of the relative frequencies  $\theta_{cjl}$  of observing certain feature values in the different classes.



A statistical model is obtained- rather than imposed - by invariance judgements of predictive opinion. Only strictly observable events are considered relative to inference. Example: the pair  $(\mathbf{x}^{(N)}, S)$  defining the sufficient statistics.



# ExchangeAbility:



Five Logotypes of the International Erasmus Student Network & Three Times Rotated

<http://exchangeability.esn.org/content/what-exchangeability>



KTH Matematik

The distribution in (3) is a (prior) predictive model for the observed items in  $N$  (=the marginal likelihood). By using the so-called *sufficientness* postulate, the prior beliefs about the relative frequencies for a feature can be explicitly expressed in terms of the Dirichlet distribution. When this is combined with the product form of (3) over features and classes arising from the generalized exchangeability, we can write the joint prior as

$$F(\theta|S) \propto \prod_{c=1}^k \prod_{j=1}^d \prod_{l=1}^{r_j} \theta_{cjl}^{\lambda_{cjl}-1}, \quad (5)$$

which is the product Dirichlet distribution where the hyperparameter  $\lambda_{cjl} > 0$ , for all index values.



# The Explicit Conditional Predictive Probability

Then, the explicit conditional predictive probability for the data derived in (3) equals

$$p(\mathbf{x}^{(N)}|S) = \prod_{c=1}^k \prod_{j=1}^d \frac{\Gamma(\sum_{l=1}^{r_j} \lambda_{cjl})}{\Gamma(n_c + \lambda_{cjl})} \prod_{l=1}^{r_j} \frac{\Gamma(n_{cjl} + \lambda_{cjl})}{\Gamma(\lambda_{cjl})}, \quad (6)$$

where  $\Gamma(\cdot)$  is the gamma function.  $n_c = \sum_{l=1}^{r_j} n_{cjl}$ , and we make use of a well-known reference prior (Perks):

$$\lambda_{cjl} = r_j^{-1}, l = 1, \dots, r_j; j = 1, \dots, d; c = 1, \dots, k. \quad (7)$$



# Posterior Distribution of Unsupervised Classification Structures

Thus, by Bayes' formula, the posterior distribution for the classification structure  $S$  is obtained as

$$p(S|\mathbf{x}^{(N)}) = \frac{p(\mathbf{x}^{(N)}|S)p(S)}{\sum_{S \in \mathcal{S}} p(\mathbf{x}^{(N)}|S)p(S)}. \quad (8)$$

The exact form of the prior distribution depends on the classification context. In unsupervised classification, the classes of  $S$  are identifiable only in terms of their contents.

To specify a prior distribution for the classification structure  $S$ , we will utilize a stochastic urn model due to A.J. Stam (1983).



# The Prior Distribution for Unsupervised Classification Structures: the Urn Model

Let  $u \in \mathbb{Z}^+$  be a stochastic number of urns into which the items in  $N$  are randomly allocated. The classification structure  $S$  is now taken to be the contents of the  $k$  nonempty urns and the prior probability of  $S$  can be expressed as

$$p(S) = \sum_{u=1}^{\infty} \pi(S|u)p(u), \quad (9)$$

where  $\pi(S|u)$  is the conditional probability of  $S$  given  $u$ .



# The Prior Distribution for Unsupervised Classification Structures: the Urn Model

The conditional probability  $\pi(S|u)$  of a partition  $S$  with  $k$  classes, given the number of urns  $u$ , equals  $u^{(k)}u^{-n}$ , for  $u \geq n$ , and zero for  $u < k$ , where  $u^{(0)} = 1$  and  $u^{(k)} = u(u-1) \cdots (u-k+1)$ . Stam noted that that a uniform distribution over the space  $\mathcal{S}$  of partitions is obtained by defining the distribution for  $u$  as

$$p(u) = |\mathcal{S}|^{-1} e^{-1} \frac{u^n}{u!},$$

where  $|\mathcal{S}|$  is the cardinality of  $\mathcal{S}$  and is given by the  $n$ th Bell number

$$|\mathcal{S}| = B_n = e^{-1} \sum_{k=1}^{\infty} \frac{k^n}{k!}, \quad (10)$$



Assume now that in addition to  $N$ , another set  $M$  of  $m$  items is available, and that the items in  $M$  are classified into  $k$  classes, using some external information. Let  $T$  denote the classification of  $M$ . Similarly as with  $N$ , let  $\mathbf{z}^{(s)}$  denote the **training data** consisting of  $d$ -dimensional feature vectors  $\mathbf{z}_i^T$ ,  $i \in s$ ,  $s \subset M$ .



We define the probability of the whole data  $(\mathbf{z}^{(M)}, \mathbf{x}^{(N)})$  as

$$\begin{aligned} & p(\mathbf{x}^{(N)}, \mathbf{z}^{(M)}, T) \\ = & \sum_{S \in \mathcal{S}} p(\mathbf{x}^{(N)} | \mathbf{z}^{(M)}, S, T) p(\mathbf{z}^{(M)} | S, T) p(S | T) p(T). \quad (11) \end{aligned}$$



We may also write, for clarity of thought,

$$p(S|\mathbf{z}^{(M)}, \mathbf{x}^{(N)}, T) = \frac{p(\mathbf{x}^{(N^{(1)})}|\mathbf{z}^{(M)}, S^{(1)}, T)p(\mathbf{x}^{(N^{(2)})}|S^{(2)})p(S^{(2)}|T)}{\sum_{S \in \mathcal{S}} p(\mathbf{x}^{(N^{(1)})}|\mathbf{z}^{(M)}, S^{(1)}, T)p(\mathbf{x}^{(N^{(2)})}|S^{(2)})p(S^{(2)}|T)},$$

where  $S^{(1)}, S^{(2)}$  are the subvectors of the joint labeling  $S$  corresponding to the  $k_1$  and  $k_2$  classes, respectively,  $N^{(1)}, N^{(2)}$  are the corresponding subsets of  $N$ , and  $p(\mathbf{x}^{(N^{(2)})}|S^{(2)})$  is the joint prior predictive probability of the data in all previously unknown classes.



# Supervised Classification Structures: the Posterior

In supervised classification the classified training data represents all possible classes. Thus, the number of urns  $u$  is fixed to  $k$  and the urns can be identified through the items of the set  $M$ . The conditional probability  $\pi(S|u, T)$  of the classification structure  $S$  is now  $u^{-n}$  for every  $S$ , as different allocations lead to different structures  $S$ , and we have

$$p(S|T) = \sum_{u=1}^{\infty} \pi(S|u, T)p(u|T) = \pi(S|k, T) = k^{-n}. \quad (12)$$



## Definition

Let  $i \in N$  be an item for which the feature vector  $x_i$  is observed. The marginal predictive supervised classifier assigns for each  $i$  the following posterior probabilities on the  $k$  possible classes from which the training data is available:

$$p(S_i = c | \mathbf{z}^{(M)}, \mathbf{x}_i) = \frac{p(\mathbf{x}_i | \mathbf{z}^{(M)}, S_i = c, T) p(S_i = c)}{\sum_{c=1}^k p(\mathbf{x}_i | \mathbf{z}^{(M)}, S_i = c, T) p(S_i = c)}, \quad (13)$$

where  $S_i = c$ ,  $c = 1, \dots, k$ , denotes the event of assigning item  $i$  into class  $c$  representing the training items and  $p(S_i = c)$  is the prior probability of this assignment. The classifier and the rule maximizing it will be abbreviated as  $p_{\text{PRED}_1}$  and  $\hat{p}_{\text{PRED}_1}$ , respectively.

## Definition

A simultaneous predictive supervised classifier assigns the posterior probability on the classes jointly for all items in  $N$  according to:

$$p(S|\mathbf{z}^{(M)}, \mathbf{x}^{(N)}, T) = \frac{p(\mathbf{x}^{(N)}|\mathbf{z}^{(M)}, S, T)p(S|T)}{\sum_{S \in \mathcal{S}} p(\mathbf{x}^{(N)}|\mathbf{z}^{(M)}, S, T)p(S|T)}. \quad (14)$$

*The classifier and the rule maximizing it will be abbreviated as  $p_{\text{PRED}_2}$  and  $\hat{p}_{\text{PRED}_2}$ , respectively.*



The idea of a simultaneous supervised classification underlying  $PPRED_2$  was originally proposed in Seymour Geisser 1964 and 1966, however, without the operationalization due to random classification structures proposed here.



## Definition

A marginalized predictive supervised classifier assigns the posterior probability for each item  $i \in N$  on the classes by marginalization of the joint posterior distribution over the classification structures in  $\mathcal{S}$ :

$$p(S_i = c | \mathbf{z}^{(M)}) = \sum_{\{S \in \mathcal{S} : S_i = c\}} p(S | \mathbf{z}^{(M)}, \mathbf{x}^{(N)}), c = 1, \dots, k. \quad (15)$$

The classifier and the rule maximizing it will be abbreviated as  $p_{\text{PRED}_3}$  and  $\hat{p}_{\text{PRED}_3}$ , respectively.



Classifiers  $p_{PPRED_2}$  and  $p_{PPRED_3}$  can be interpreted to be more honest than  $p_{PPRED_1}$  in probabilistic sense, as they compare the data predictions after updating knowledge about the class-conditional distributions given the putative classifications of the new items in  $N$ . For any particular item  $i \in N$  the marginal uncertainty about the classification represented by  $p_{PPRED_3}$  follows from the application of the law of total probability on  $p_{PPRED_2}$ . Such a classifier can be interpreted as the most honest item-wise representation of classification uncertainty among the considered alternatives.



Is there something like a general statistical learning theory (SLT) underlying this ?



Jackson, Kalai & Smorodinsky (1999) launch the following definition:

## Definition

Let  $P$  and  $Q$  be two probability measures,  $\{X_n\}_{n \geq 1}$  a sequence of discrete random variable, and let  $A$  be an event in the sigma-field  $\sigma(\{X_i\}_{i \geq n+1}^{n+l})$ .  $P^m$  and  $Q^m$  are the respective conditional distributions given the sigma-field  $\sigma(\{X_l\}_{l \geq 1}^m)$ . Then  $P$  **merges with**  $Q$ , if for all  $\epsilon > 0$ , and all  $l > 0$  it holds that

$$\max_{n \geq m, A \in \sigma(\{X_i\}_{i \geq n+1}^{n+l})} |P^m(A) - Q^m(A)| < \epsilon$$

with  $Q$  -probability one.

Jackson, Kalai & Smorodinsky (1999)

$$\max_{n \geq m, A \in \sigma(\{X_i\}_{i \geq n+1})} |P^m(A) - Q^m(A)| < \epsilon$$

with  $Q$  -probability one. Let us think that  $Q$  is the 'true probability' and  $P$  is our model. Then the finite horizon event predictions at arbitrary times in the future will approach the true forecasts provided by  $Q$ .



It follows as in D. Blackwell & L. Dubins (1962) (under a stronger version of merging) that if  $P$  and  $Q$  are two measures for a sequence of discrete random variables  $\{X_n\}_{n \geq 1}$  and  $Q \ll P$ , then  $P$  merges with  $Q$ . Let us think that  $Q$  is the 'true probability' and  $P$  is our model.  $Q \ll P$  means that  $Q(A) > 0 \Rightarrow P(A) > 0$  i.e., we cannot be surprised by an event that actually happens.



If  $P$  corresponds to the probability mass distribution

$$p(\mathbf{x}^{(N)}|S) = \int_{\Theta} p(\mathbf{x}^{(N)} | \theta, S) dF(\theta|S),$$

and if  $Q$  corresponds to the conditional probability mass distribution

$$q(\mathbf{x}^{(N)}|S) = \int_{\Theta} p(\mathbf{x}^{(N)} | \theta, S) dQ(\theta|S),$$

and if  $dQ|S \ll dF|S$  on  $\Theta$ , then  $Q \ll P$ .



# Merging & the Classification Structure $S$ (II)

Then with

$$P^m \leftrightarrow p(\mathbf{x}^{(N)} | S, T, \mathbf{z}^{(M)}) = \int_{\Theta} p(\mathbf{x}^{(N)} | \theta, S) dP(\theta | T, \mathbf{z}^{(M)})$$

$$Q^m \leftrightarrow q(\mathbf{x}^{(N)} | S, T, \mathbf{z}^{(M)}) = \int_{\Theta} p(\mathbf{x}^{(N)} | \theta, S) dQ(\theta | T, \mathbf{z}^{(M)})$$

we have

$$\max_{n \geq m, A \in \sigma(\{X_i\}_{i \geq n+1}^{n+l})} |P^m(A) - Q^m(A)| < \epsilon$$

or, the finite horizon event predictions at arbitrary times in the future will approach the true forecasts provided by  $Q$ , where  $P$ -predictions are conditioned on  $S, T, \mathbf{z}^{(M)}$ .

This is a 'sufficient for prediction' property of  $P$  in supervised classification using simultaneous predictive probabilities.



Another form of learnability related to  $Q^m$  given by the *posterior predictive probability*

$$q(\mathbf{x}^{(N)} | S, T, \mathbf{z}^{(M)}) = \int_{\Theta} p(\mathbf{x}^{(N)} | \theta, S) dQ(\theta | T, \mathbf{z}^{(M)}),$$

is obtained by exchangeability, since by DeFinetti  $Q$  merges with the measure  $Q_\theta$  given by any  $p(\mathbf{x}^{(N)} | \theta, S)$  chosen from  $Q(\theta | S)$ . For prediction and supervised classification we need not estimate the (randomly chosen) 'true'  $Q_\theta$ . We can regard  $Q_\theta$  as a tool for making infinite horizon event predictions, too.



Under the current assumptions yielding Merging & Learnability one can show, since  $\mathcal{S}$  is finite the following:

## Theorem

$$\left| \max_{S \in \mathcal{S}} p(\mathbf{x}^{(N)} | S, T, \mathbf{z}^{(M)}) - \max_{S \in \mathcal{S}} q(\mathbf{x}^{(N)} | S, T, \mathbf{z}^{(M)}) \right|$$

*becomes arbitrarily small for large  $m$ .*



The logarithm of the data predictive probability for the simultaneous classifier  $p_{\text{PRED}_2}$  which equals under the previous derivations:

$$\begin{aligned} & \log p(\mathbf{x}^{(N)} | \mathbf{z}^{(M)}, \mathcal{S}, \mathcal{T}) && (16) \\ = & \sum_{c=1}^k \sum_{j=1}^d \log \frac{\Gamma(m_c + 1)}{\Gamma(n_c + m_c + 1)} \\ & + \sum_{c=1}^k \sum_{j=1}^d \sum_{l=1}^{r_j} \log \frac{\Gamma(n_{cjl} + m_{cjl} + \lambda_{cjl})}{\Gamma(m_{cjl} + \lambda_{cjl})}. \end{aligned}$$



The expression for the logarithm of the data predictive probability of the marginal classifier  $p_{PRED_1}$  by considering:

$$\begin{aligned} & \log p(\mathbf{x}_1, \dots, \mathbf{x}_n | \mathbf{z}^{(M)}, S, T) & (17) \\ &= \sum_{i=1}^n \log p(\mathbf{x}_i | \mathbf{z}^{(M)}, i, T) \\ &= \sum_{c=1}^k \sum_{S_j=c} \log p(\mathbf{x}_j | \mathbf{z}^{(M)}, T, i), \end{aligned}$$

where  $S$  now refers to the classification structure resulting from the assignment of the  $n$  items into the classes one by one.



Using Stirling's approximation to the gamma function terms for any particular classification structure  $S$ , the difference in the logarithms of the data predictive probability between the simultaneous and marginal classifiers tends to zero as the amount of training data increases, i.e.

## Theorem

$$\log p(\mathbf{x}^{(N)} | \mathbf{z}^{(M)}, S, T) - \log p(\mathbf{x}_1, \dots, \mathbf{x}_n | \mathbf{z}^{(M)}, S, T) \xrightarrow{m \rightarrow \infty} 0,$$

*when the sampling process of training data from each class is infinitely exchangeable and when all classes in  $T$  are infinitely persistent.*

The limits of the relative frequencies of feature values exist and we assume that they are strictly positive. Consequently, the classifier rules  $\hat{p}_{PRED_1}$  and  $\hat{p}_{PRED_2}$  will become equivalent with increasing training dataset size, as the optimal classification according to one rule will coincide with that of the other rule, which is formally stated in the theorem below.



In semi-supervised classification the training set information in  $T$  is considered to represent only a subset of the eligible classes. The classification structure  $S$  can then contain classes that are present in  $T$  and new classes that are identifiable only in terms of their contents.



# Semisupervised Classification: A Cartoon

© Original Artist

Reproduction rights obtainable from  
www.CartoonStock.com



KTH Matematik

# Semisupervised Classification structures

The number of such classification structures can be computed by considering the known and the new classes separately. Let  $r$ ,  $0 \leq r \leq n$ , be the number of items of  $N$  allocated to the  $k_1$  known classes. The  $r$  items can be chosen from  $N$  in  $\binom{n}{r}$  ways and allocated to the  $k_1$  classes in  $k_1^r$  ways. The remaining  $n - r$  items are allocated analogously to the unsupervised classification strategy discussed above and form a partition of the corresponding item set into  $k_2$  new classes. The number of classification structures is consequently obtained by summing over the possible values of  $r$  as

$$|\mathcal{S}| = \sum_{r=0}^n \binom{n}{r} k_1^r B_{n-r} = \sum_{r=0}^n \binom{n}{r} k_1^r e^{-1} \sum_{k_2=1}^{\infty} \frac{k_2^{n-r}}{k_2!}. \quad (18)$$



# The Prior for Semisupervised Classification Structures

This is an extension of random partitions by random urns due to Stam.

We draw the number  $u$  of urns from a distribution  $p(u|T)$  and then given  $u$ , each item of  $N$  is allocated to an urn corresponding to a class in  $T$  with probability  $k_1/u$ . Then  $\pi(S|u, T)$  is the conditional probability of a classification structure  $S$  with  $r$  items in the  $k_1$  known classes and  $n - r$  items in  $k_2$  new classes.

The probability of a classification structure  $S$  is now conditional to the classification  $T$  of the training data. Using the stochastic urn model it can be expressed as

$$p(S|T) = \sum_{u=1}^{\infty} \pi(S|u, T)p(u|T). \quad (19)$$



The conditional probability  $\pi(S|u, T)$  of a classification structure  $S$  with  $r$  items in the  $k_1$  known classes and  $k_2$  new classes can be computed by considering the  $r$  and  $n - r$  items separately and has the expression

$$\begin{aligned} & \pi(S|u, T) \\ &= \left( \left( \frac{k_1}{u} \right)^r \left( \frac{1}{k_1} \right)^r \right) \left( \left( \frac{u - k_1}{u} \right)^{n-r} \frac{(u - k_1)^{(k_2)}}{(u - k_1)^{n-r}} \right) \\ &= \frac{(u - k_1)^{(k_2)}}{u^n}, \end{aligned} \tag{20}$$

for  $u \geq k_1$ , where  $u^{(k)} = u(u - 1) \dots (u - k + 1)$ .



Given the urns, some semi-supervised classifier rules can be obtained by using expressions analogous to  $p_{PRED_1}$  and  $p_{PRED_2}$ , except that the space of classification structures  $\mathcal{S}$  is extended accordingly.



$i \in N$  is an item for which the feature vector  $\mathbf{x}_i$  is observed. The marginal predictive semi-supervised classifier for each  $i$  is based on the probability:

$$p(S_i = c | \mathbf{z}^{(M)}, \mathbf{x}_i) = \frac{p(\mathbf{x}_i | \mathbf{z}^{(M)}, S_i = c, T) p(S_i = c)}{\sum_{c=1}^{k_1} p(\mathbf{x}_i | \mathbf{z}^{(M)}, S_i = c, T) p(S_i = c)},$$

if the item is assigned in one of  $k_1$  known classes harboring the training data,  $c = 1, \dots, k_1$ ,



and alternatively on the probability:

$$p(S_i = c' | \mathbf{x}_i) = \frac{p(\mathbf{x}_i | S_i = c')p(S_i = c')}{\sum_{c'=1}^{k'} p(\mathbf{x}_i | S_i = c')p(S_i = c')}, \quad (21)$$

if the single item is assigned in one of  $k_2$  unknown classes  $c'$  lacking training data,  $c' = 1, \dots, k_2$ . Notice that  $k_2$  is not known or fixed a priori. The classifier and the rule maximizing it will be abbreviated as  $p_{pred_1}^{semi}$  and  $\hat{p}_{pred_1}^{semi}$ , respectively.



A simultaneous predictive semi-supervised classifier assigns the posterior probability on the classes jointly for all items in  $N$  according to :

$$p(S|\mathbf{z}^{(M)}, \mathbf{x}^{(N)}) = \frac{p(\mathbf{x}^{(K)}|\mathbf{z}^{(M)}, S, T)p(\mathbf{x}^{(U)}|S)p(S|T)}{\sum_{S \in \mathcal{S}} p(\mathbf{x}^{(K)}|\mathbf{z}^{(M)}, S, T)p(\mathbf{x}^{(U)}|S)p(S|T)}, \quad (22)$$

where  $\mathbf{x}^{(K)}$  represents the items assigned in known classes and  $\mathbf{x}^{(U)}$  denotes those assigned in unknown classes according to the structure  $S$ . The classifier and the rule maximizing it will be abbreviated as  $p_{pred_2}^{semi}$  and  $\hat{p}_{pred_2}^{semi}$ , respectively.



A marginalized predictive semi-supervised classifier assigns the posterior probability for each item  $i \in N$  on the classes by marginalization of the joint posterior distribution over the classification structures in  $\mathcal{S}$ :

$$p(S_i = c | \mathbf{z}^{(M)}, \mathbf{x}_i) = \sum_{\{S \in \mathcal{S} : S_i = c\}} p(S | \mathbf{z}^{(M)}, \mathbf{x}^{(N)}), \quad (23)$$

if the item is assigned in one of  $k_1$  known classes  $s_c$   $c = 1, \dots, k_1$ ,



and alternatively the probability

$$p(S_i = c' | \mathbf{x}_i) = 1 - \sum_{c=1}^k \sum_{\{S \in \mathcal{S} : S_i = c\}} p(S | \mathbf{z}^{(M)}, \mathbf{x}^{(N)}), \quad (24)$$

if the single item is assigned in any of unknown classes  $s_{c'}$ ,  $c' = 1, \dots, k_2$ . The classifier and the rule maximizing it will be abbreviated as  $p_{pred_3}^{semi}$  and  $\hat{p}_{pred_3}^{semi}$ , respectively.



# The Explicit Conditional Predictive Probability for the Semi-Supervised

Following the two above subsections, the corresponding predictive probability for the semi-supervised situation is a trivial extension

$$\begin{aligned} & p(\mathbf{x}^{(N)} | \mathbf{z}^{(M)}, S, T) \\ &= \prod_{c=1}^{k_1} \prod_{j=1}^d \frac{\Gamma(m_c + 1)}{\Gamma(n_c + m_c + 1)} \prod_{l=1}^{r_j} \frac{\Gamma(n_{cjl} + m_{cjl} + \lambda_{cjl})}{\Gamma(m_{cjl} + \lambda_{cjl})} \\ & \cdot \prod_{c'=1}^{k_2} \prod_{j=1}^d \frac{\Gamma(1)}{\Gamma(n_{c'} + 1)} \prod_{l=1}^{r_j} \frac{\Gamma(n_{c'jl} + \lambda_{c'jl})}{\Gamma(\lambda_{c'jl})}, \end{aligned} \quad (25)$$

where  $k_1$  is the number of previously known (labeled) classes and  $k_2$  is the number of novel, previously unknown classes under the structure  $S$ , respectively. The parameters in this expression are defined analogously to the above.



We now compare for a real dataset. We only compare  $\hat{p}_{PRED_1}$  and  $\hat{p}_{PRED_2}$  to assess the relative effect of simultaneous inductive learning. The real data consist of 912 bacterial isolates representing  $k = 15$  distinct classes which are summarized in Table 2. The data contain binary profiles over  $d = 47$  features and have earlier been considered by Corander et al. (2009). In particular, in Corander et al. (2009) approximately 130 classes were obtained from a database of 5313 bacterial isolates. Among these we picked at random a subset of 15 classes with sizes varying between 34 and 124 samples. An image of the class-wise data is shown in Figure 1.



# Performance of Prediction

Paul Klee: *Angelus Novus* 1920, seen (c.f., W. Benjamin) as the Angel of Prediction: 'back against the future'



KTH Matematik



# Performance of the Marginal and Simultaneous Classifiers

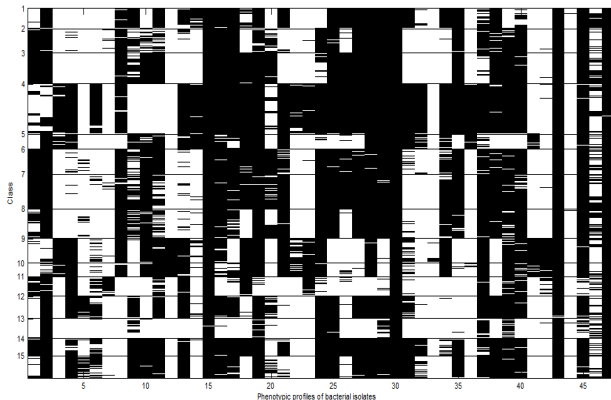


Image of the bacterial data. Each column represents a binary phenotypic characteristic of an isolate coded with black/white colors and each row is an observed sample. The black horizontal lines indicate class boundaries.

# A look at the data format

<i>genus species</i>	<i>strain</i>	<i>biochem. profile</i>
BUDV AQUA	0442-84	01001100000000111010010010101000000001000010100
BUDV AQUA	2574-80	01001000001000111010000010101000000110000010100
BUDV AQUA	0444-84	01001100000000111010000010101000000100000010110
BUDV AQUA	0443-84	01001100000000111010000010101000000111000010100
ERWI HERB	8060-83	00000000000100100010100010101100010000100010101
ESCH COLI	0783-83	01000000000000101011100010111100001011100010111
ESCH COLI	4367-83	00000000000000100010000010110100000001000010001
ESCH COLI	1442-77	11000001000000101000100010111100011011100010101
ESCH COLI	1536-79	01000000001000111000100010011100010100100010110
ESCH COLI	1022-78	01000000000000111001100110111100010000000010101
ESCH COLI	3270-74	10000001000000110000100110111100011101100110001



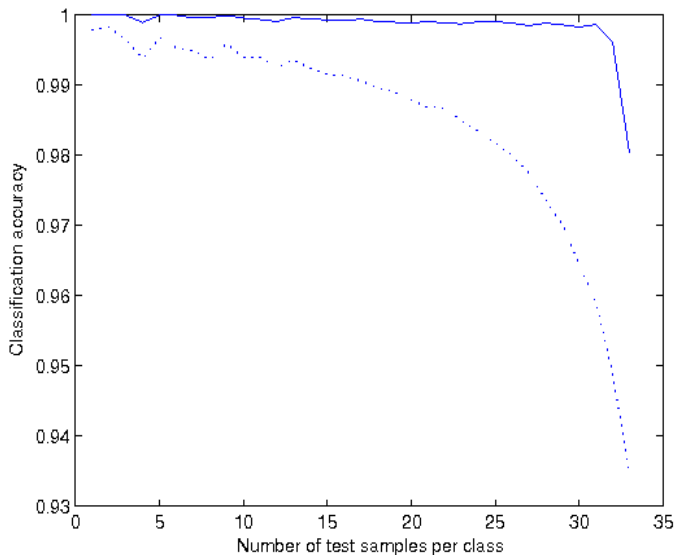
To compare the abilities of the marginal and simultaneous classifiers to identify the origin of a bacterial sample in the presence of varying amount of training and test data, we sampled without replacement  $j$  bacterial isolates, for  $j = 1, \dots, 33$ , from each of the 15 classes and comprised the test dataset  $N$  from these  $15j$  samples, while the remainder of the data were assigned to  $M$  given the fixed classification from the earlier unsupervised analysis. Thus, the smallest and largest test datasets contained 15 and 495 bacterial isolates, respectively.



A total of 100 random replicates of the test dataset were generated for each value of  $j$  and these were classified using the greedy optimization algorithms implementing  $\hat{p}_{PRED_1}$  and  $\hat{p}_{PRED_2}$ . Figure 2 shows the item-wise correct classification rates for the two classifiers as a function of  $j$ . The asymptotic equivalence of the classifiers is well reflected by the nearly identical results for small values of  $j$ , while the difference of the two methods increases steadily as a function of the size of the test datasets. This demonstrates in a realistic supervised setting that gains in classification accuracy can be obtained by using the simultaneous classifier instead of a marginal approach, even if it is only implemented using the deterministic greedy optimization.



# Performance of the Marginal and Simultaneous Classifiers



KTH Matematik

Figure 2. The item-wise correct classification rates as a function of test dataset size. The curves correspond to the two classifiers  $\hat{p}_{PRED_1}$  (lower, dashed curve) and  $\hat{p}_{PRED_2}$  (upper, continuous curve). Results are based on random splits of the bacterial data described in the text and Table 2.



# The Semi-Supervised Classifier

Item-wise correct classification rates for different semi-supervised classification rules based on 1,000 replications of each simulation

	$\hat{p}_{pred_1}$	$\hat{p}_{pred_2}$	$\hat{p}_{pred_3}$
setting. 1. Setup 1: 2 + 2 + 1 :	0.5782	0.5996	0.6018
Setup 2: 2 + 1 + 2 :	0.4592	0.5952	0.6174

the first setup (2+2+1), two test samples were generated from each of two known source distributions and one sample was from an unknown source distribution. In the second experiment (2+1+2), two and a single test samples were simulated from the two known classes, respectively, and additionally two samples were from an common unknown source. In these two cases there are 674 distinct simultaneous classification structures based on the urn models.



It was demonstrated that the standard marginal supervised classifiers are asymptotically equivalent to a simultaneous classifier, however, as the numerical examples illustrated, the latter approach can in practice provide higher correct classification rates. In contrast to the standard supervised classifiers, the simultaneous and marginalized classifiers considered here enable coherent statistical learning about source distributions to take place also *after* the training data have been introduced. The cost of such coherence is the increased computational challenge of using such classifiers, which is beyond the scope of this lecture.

