Feature selection methods for mining bioinformatics data

Gianluca Bontempi, Benjamin Haibe-Kains

{gbonte,bhaibeka}@ulb.ac.be

Machine Learning Group

Departement d’Informatique

ULB, Université Libre de Bruxelles

Boulevard de Triomphe - CP 212

Bruxelles, Belgium

http://www.ulb.ac.be/di/mlg

Outline

- The ULB Machine Learning Group
- The feature selection problem
- Feature selection and bioinformatics.
- Feature selection as a stochastic optimization problem.
- Some of our bioinformatics applications.
- Future work.

ULB Machine Learning Group (MLG)

- 7 researchers (1 prof, 6 PhD students), 4 graduate students).
- Research topics: Local learning, Classification, Computational statistics, Data mining, Regression, Time series prediction, Sensor networks, Bioinformatics.
- Computing facilities: cluster of 16 processors, LEGO Robotics Lab.
- Website: `www.ulb.ac.be/di/mlg`.
- Scientific collaborations in ULB: IRIDIA (Sciences Appliquées), Physiologie Moléculaire de la Cellule (IBMM), Conformation des Macromolécules Biologiques et Bioinformatique (IBMM), CENOLI (Sciences), Microarray Unit (Hopital Jules Bordet), Service d'Anesthesie (ERASME).
- Scientific collaborations outside ULB: UCL Machine Learning Group (B), Politecnico di Milano (I), Universitá del Sannio (I), George Mason University (US).
- The MLG is part to the "Groupe de Contact FNRS" on Machine Learning.

ULB-MLG: running projects

1. "Integrating experimental and theoretical approaches to decipher the molecular networks of nitrogen utilisation in yeast": ARC (Action de Recherche Concertée) funded by the Communauté FranÄ¿çaise de Belgique (2004-2009). Partners: IBMM (Gosselies and La Plaine), CENOLI.

2. "COMP^{2}SYS" (COMPutational intelligence methods for COMPlex SYStems) MARIE CURIE Early Stage Research Training funded by the European Union (2004-2008). Main contractor: IRIDIA (ULB).

3. "Predictive data mining techniques in anaesthesia": FIRST Europe Objectif 1 funded by the Région wallonne and the Fonds Social Européen (2004-2009). Partners: Service d'anesthesie (ERASME).

4. "AIDAR - Adressage et Indexation de Documents Multimédias Assistés par des techniques de Reconnaissance Vocale": funded by Région Bruxelles-Capitale (2004-2006). Partners: Voice Insight, RTBF, Titan.

Feature selection

- In recent years many applications of data mining (text mining, bioinformatics, sensor networks) deal with a very large number n of features (e.g. tens or hundreds of thousands of variables) and often comparably few samples.
- In these cases, it is common practice to adopt feature selection algorithms [4] to improve the generalization accuracy.
- There are many potential benefits of feature selection:
  - facilitating data visualization and data understanding,
  - reducing the measurement and storage requirements,
  - reducing training and utilization times,
  - defying the curse of dimensionality to improve prediction performance.

Feature selection and bioinformatics

- The availability of massive amounts of experimental data based on genome-wide studies has given impetus in recent years to a large effort in developing mathematical, statistical and computational techniques to infer biological models from data.
- In many bioinformatics problems the number of features is significantly larger than the number of samples (high feature to sample ratio datasets).
- Examples can be found in the following bioinformatics tasks:
  - Breast cancer classification on the basis of microarray data.
  - Network inference on the basis of microarray data.
  - Analysis of sequence/expression correlation.

Breast cancer classification

- Breast cancer is one of the most common malignant tumors affecting women.
- Breast cancer patients with the same stage of disease can have markedly different treatment responses and overall outcome.
- Cancer classification has been based primarily on morphological appearance of the tumor, but with serious limitations. Tumors with similar histopathological appearance can follow significantly different clinical courses and show different responses to therapy. The strongest predictors for metastasis fail to classify accurately breast tumors according to their clinical behavior.
- Cancer classification has been difficult in part because it has historically relied on specific biological insights, rather than systematic and unbiased approaches for recognizing tumor subtypes.

Breast cancer classification (II)

- Chemiotherapy or hormonal therapy reduces the risk of distant metastasis by approximately one-third; however 70-80% of patients receiving this treatment would have survived without it. Also, these therapies frequently have toxic side effects.
- Diagnosis of cancer must be accurate in order for the patient to receive the correct treatment and so have the best chance of survival.
- The cellular and molecular heterogeneity of breast tumors and the large number of genes potentially involved in controlling cell growth, death and differentiation emphasize the importance of studying multiple genetic alterations in concert.
- The development of microarray technology provides the opportunity of correlating genome-wide expressions with the response of tumor cells to chemiotherapy.

Breast cancer classification (III)

- Systematic investigation of expression patterns of thousands of genes in tumors using DNA microarrays and their correlation to specific features of phenotypic variation might provide the basis for an improved taxonomy of cancer.
- It is expected that variations in gene expression patterns in different tumors could provide a “molecular portrait” of each tumor, and that the tumors could be classified into subtypes based solely on the difference of expression patterns.
- In litterature [11] classification techniques have been applied to identify a gene expression signature strongly predictive of a short interval to distant metastases in patients without tumor cells in local lymph nodes at diagnosis.
- In this context the number n of features equals the number of genes (ranging from 6000 to 30000) and the number N of samples is the number of patients under examinations (about hundreds).

Inference of regulatory networks

- Most biological regulatory processes involve intricate networks of interactions and it is now increasingly evident that predicting their behaviour and linking molecular and cellular structure to function are beyond the capacity of intuition.
- The idea is that transcriptional processes of induction and repression are determined through specific interactions, and can be predicted in detail by a logical or a mathematical model.
- The ultimate goal is to know, for each specific gene, what other genes it influences and in what way. In literature two families of techniques have mostly been used to infer network models from large sets of expression data: graphical models (boolean and bayesian networks) and dynamic regulation models.

Inference of regulatory networks (II)

- Dynamic regulation models are black-box prediction models which represent gene activity with continuous values. Examples of dynamic regulation models are linear relationships of the form

  $$x_i(t + \delta t) = \sum_j w_{ji} x_j(t) + b_i$$

  where x_i is the expression level of the i th gene at time t, b_i is a bias term indicating whether the i-th gene is expressed in absence of regulatory inputs and the weight w_{ij} indicates the influence of gene j on the regulation of gene i.

- In this case, inferring a dynamic model boils down at estimating the unknown weights on the basis of expression data.

- Nonlinear version of dynamical regulation models, based on the use of recurrent neural networks, have been proposed in [8, 7].

Inference of regulatory networks (II)

- In general terms, revealing the network of the transcriptional regulation process appears to be a very hard problem for several reasons: noisy data, non linear effects, loose connection of the regulation network, risk of overfitting, dynamic effects (e.g. feedback, stability) to be taken into consideration.
- These problems demand the estimation of a number of predictive models for each gene, where the number of features equals the number of measured genes.

Correlating motifs and expression levels

- These methods consists in directly correlating expression levels and regulatory motif present in presumptive transcription control regions [2, 10].
- Published work adopts a linear regression to model the additive contribution of upstream motifs to the log-expression level of a gene. In [2], the expression of a gene in a single experimental condition is modelled as a linear function $E = a_1 S_1 + a_2 S_2 + \cdots + a_n S_n$ of scores computed for sequence motifs in the upstream control region.
- These sequence motif scores incorporate the number of occurrences of the motifs and their positions with respect to the gene's translation start site.

orrelating motifs and expression levels (I

- In other terms the sequence motifs of a specific gene are considered as explanatory variables (feature inputs) of a statistical model which correlates sequence features and expression of the gene.
- The number of features is $n = 4^m$ for motifs of length m.

The two issues of f.s.

Two main issues make the problem of feature selection a highly challenging task:

Search in a high dimensional space: this is known to be a NP-hard problem.

Assessment on the basis of a small set of samples: this is made difficult by the high ratio between the dimensionality of the problem and the number of measured samples.

Approaches to f.s.

Three are the main approaches to feature selection:

Filter methods: they are preprocessing methods. They attempt to assess the merits of features from the data, ignoring the effects of the selected feature subset on the performance of the learning algorithm. Examples are methods that select variables by **ranking** them through compression techniques (like PCA) or by computing correlation with the output.

Wrapper methods: these methods **assess subsets of variables** according to their usefulness to a given predictor. The method conducts a search for a good subset using the learning algorithm itself as part of the evaluation function. The problem boils down to a problem of stochastic state space search. Example are the stepwise methods proposed in linear regression analysis.

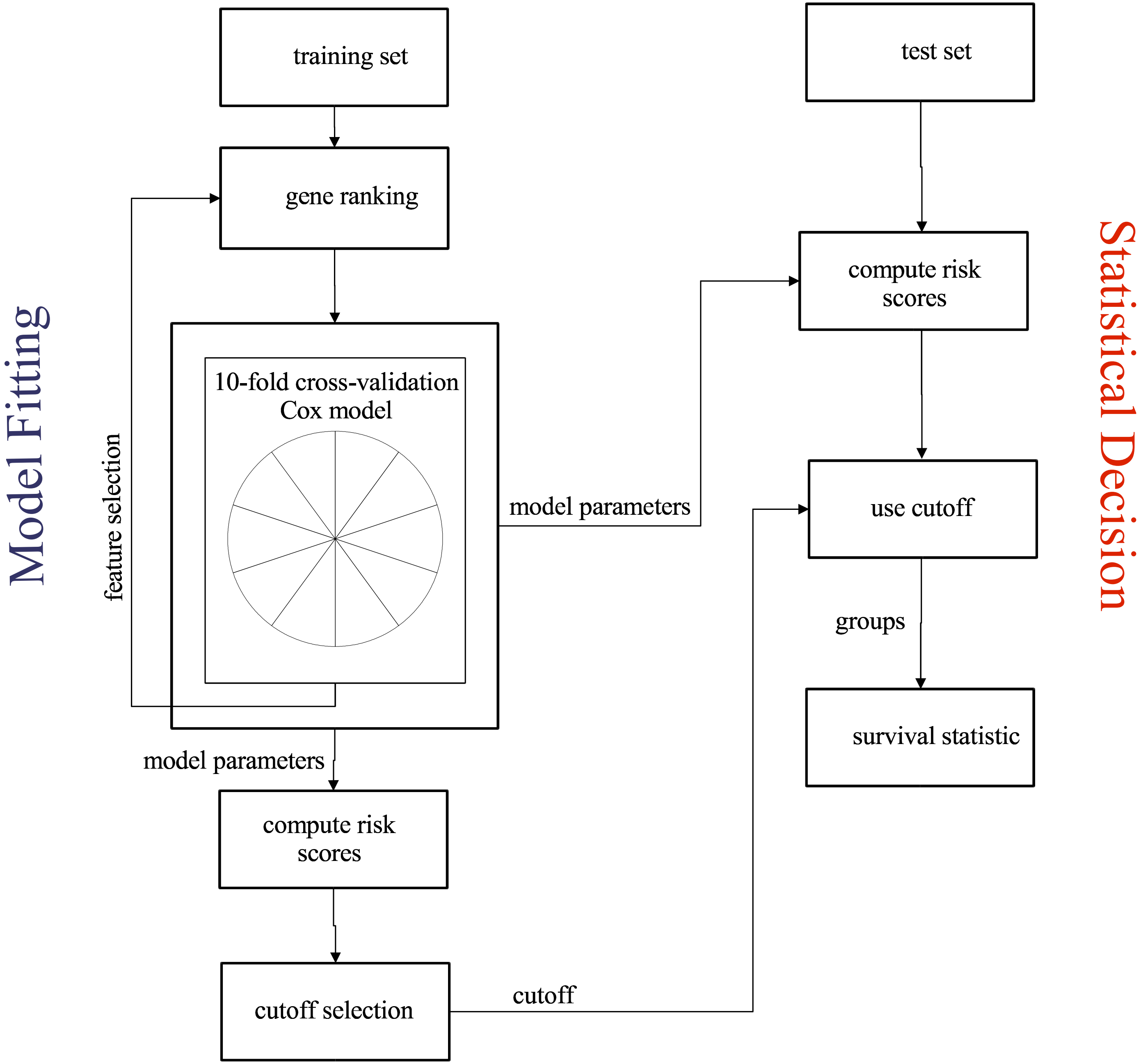
Embedded methods: they perform variable selection as part of the learning procedure and are usually specific to given learning machines. Examples are classification trees, regularization techniques (e.g. lasso).

A ranking example

- Joint project with Microarray Unit of Bordet Hospital (Brussels) headed by Dr. Sotiriou.
- Motivations:
    - majority of early-stage breast cancers express estrogen receptors (ER) and receive the **t** drug in the adjuvant setting.
    - **40%** of these patients will relapse on the **t** drug and develop incurable metastatic disease.
    - the goal of the data mining analysis is to identify those patients at higher risk of **t** resistance on the basis of their genetic profile.

Learning procedure

- Dataset: gene expressions ($n \approx 44,000$ probes) of $N = 166$ patients measured by the AFFYMETRIX$^{©}$ platform.
- Prediction technique: survival analysis with the Cox proportional hazards regression [3]
- Feature filter selection:
  - Gene ranking : univariate Cox regression for each probe on the training set → ranking on the basis of p-value significance.
  - best size selection: selection of the "best" number of probes in a multivariate Cox model by a **10-fold cross-validation** procedure
- The best multivariate Cox model contains only 5 probes.
- Assessment of the **difference in survival** between the low and the high-risk groups with the Kaplan-Meier estimator and the logrank test [9]



Model Fitting

training set

gene ranking

feature selection

10-fold cross-validation
Cox model

model parameters

model parameters

compute risk
scores

cutoff selection

cutoff

Statistical Decision

test set

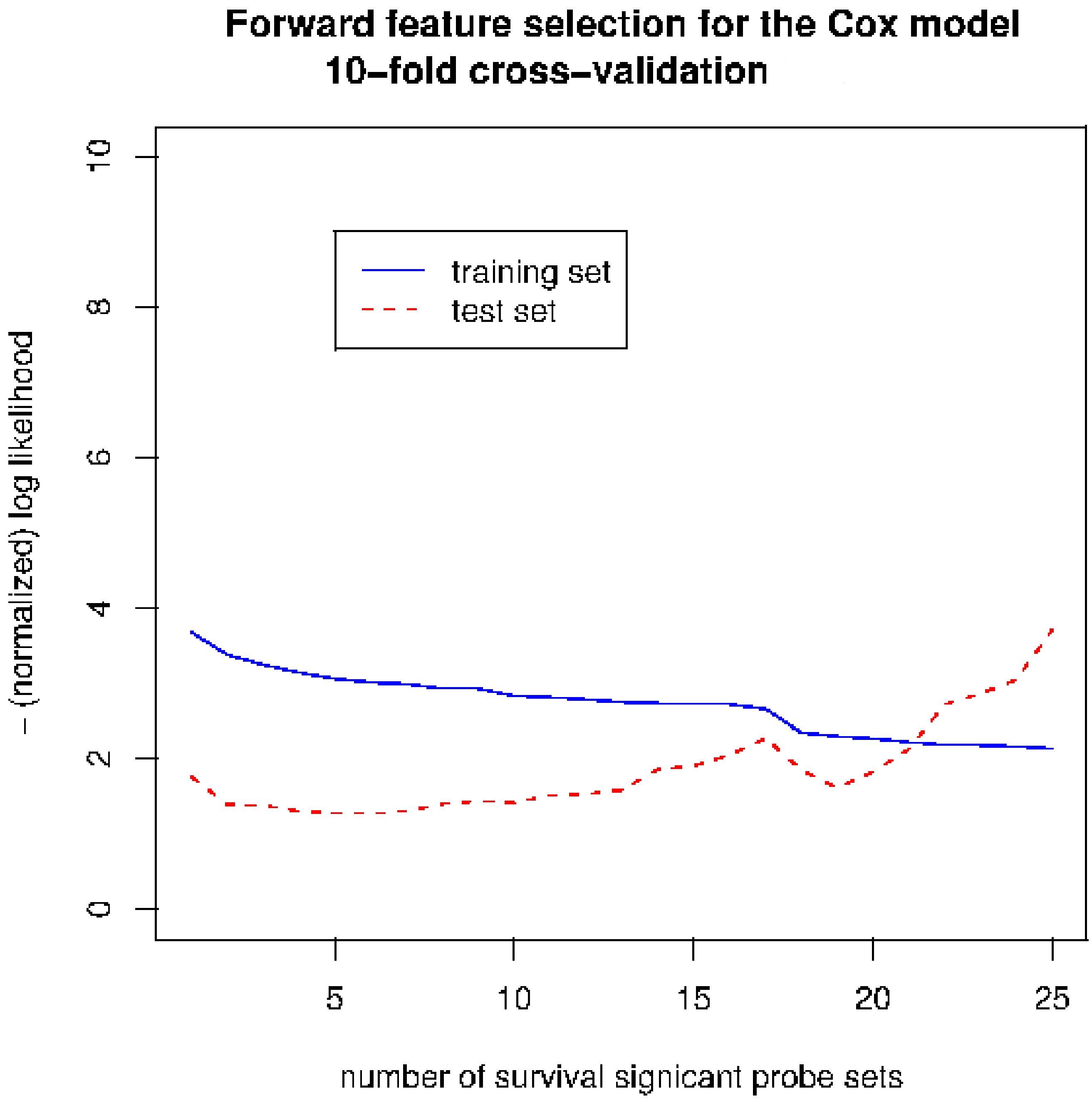
compute risk
scores

use cutoff

groups

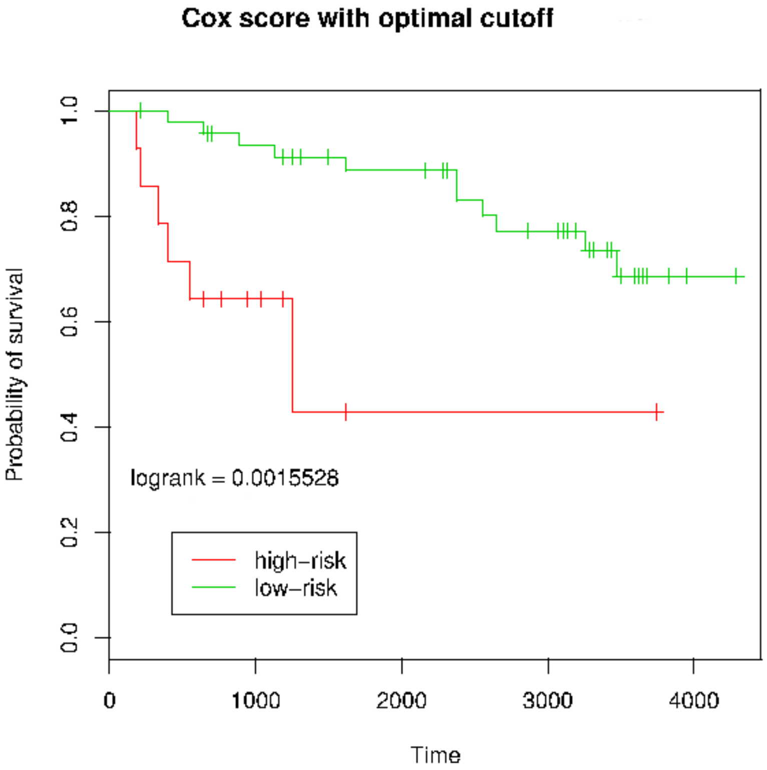
survival statistic

Size selection by 10-fold-cv



Results

- After computing the **risk score**, a threshold is selected to discriminate the low and the high-risk groups
- We assist to a very significant difference in survival on the test set.



The 5 probes are able to discriminate the survival of patients treated by the **t** drug.

Wrapping search

- The wrapper search can be seen as a search in a space $S = \{0,1\}^n$ where a generic vector $s \in S$ is such that

$$s_j = \begin{cases} 0 & \text{if the input } j \text{ does NOT belong to the set of features} \\ 1 & \text{if the input } j \text{ belongs to the set of features} \end{cases}$$

- We look for the optimal vector $s^* \in \{0,1\}^n$ such that

$$s^* = \arg\min_{s \in S} \mathsf{GE}(s)$$

  where GE(s) is the generalization error of the model based on the set of variables described by s.

- the number of vectors in S is equal to 2^n.
- for moderately large n, the exhaustive search is no more possible.

Wrapping greedy strategies

Various methods have been developed for evaluating only a small number of variables by either adding or deleting one variable at a time. We consider here some greedy strategies:

Forward selection: the procedure starts with no variables. The first input selected is the one which allows the lowest generalization error. The second input selected is the one that, together with the first, has the lowest error, and so on, till when no improvement is made.

Backward selection: it works in the opposite direction of the forward approach. We begin with a model that contains all the n variables. The first input to be removed is the one that allows the lowest generalization error.

Stepwise selection: it combines the previous two techniques, by testing for each set of variables, first the removal of features beloning to the set, then the addition of variables not in the set.

Open issues

- The wrapper approach to feature selection requires the assessment of several subset alternatives and the selection of the one which is expected to have the lowest generalization error.
- To tackle this problem, we need to perform a search procedure in a very large space of subsets of features aiming to minimize a leave-one-out or more in general a cross-validation criterion.
- This practice can lead to a strong bias selection in the case of high dimensionality problems.
- In plain words, searching for the best subset in very large spaces is prone to overfitting, even if assessment relies on cross-validations.

Stochastic discrete optimization

Consider the stochastic minimization of the positive function $g(s) = E[\mathbf{G}(s)], s \in S$, that is the expected value function of a random function $\mathbf{G}(s) > 0$. Let $G(s)$ be a realization of $\mathbf{G}(s)$ and

$$\hat{s} = \arg\min_{s \in S} G(s) \tag{1}$$

In general terms, coupling the estimation of an expected value function $g(s)$ with the optimization of the function itself should be tackled very cautiously because of the well-known relation [6]

$$E[\min_{s \in S} \mathbf{G}(s)] = E[\mathbf{G}(\hat{\mathbf{s}})] \leq \min_{s \in S} E[\mathbf{G}(s)] = \min_{s \in S} g(s) = g(s^*) = g^* \tag{2}$$

where g^* is the minimum of $g(s)$ and $\hat{G} = G(\hat{s}) = \min_{s \in S} G(s)$ is the minimum of the resulting “approximation problem” dependent on the realization G of the r.v. $\mathbf{G}$

Stochastic discrete optimization (II)

- The above relation states that the minimum of an expected value is optimistically estimated by the minimum of the corresponding sample function.

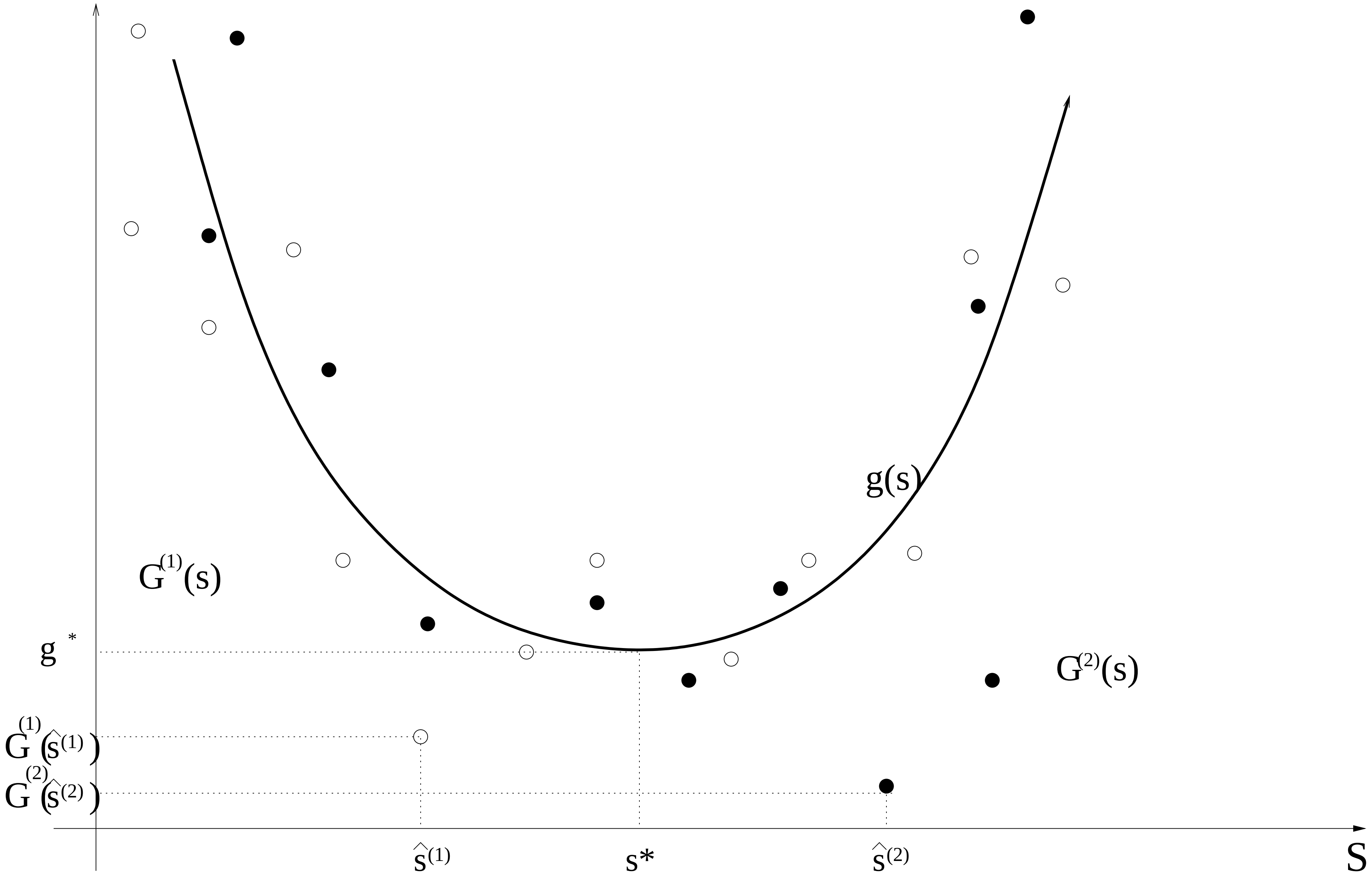
- Also, since $\forall \hat{s}, \min_{s \in S} g(s) \leq g(\hat{s})$, we have $\min_{s \in S} g(s) \leq E[g(\hat{\mathbf{s}})]$ and consequently that

$$E[\min_{s} \mathbf{G}(s)] = E[\mathbf{G}(\hat{\mathbf{s}})] \leq E[g(\hat{\mathbf{s}})]. \tag{3}$$

- This means that the minimum $G(\hat{s})$ of a sample function is a **biased estimate** of the average value of the function $g(\cdot)$ in $\hat{\mathrm{s}}$.

- Note that the value $E[g(\hat{\mathrm{s}})]$ is a measure of the value of the cost function that we are able to obtain once we minimize the *not observable* cost function $g(\cdot)$ on the basis of the observable realizations G.

Stochastic discrete optimization (III)



Supervised learning

Consider a supervised regression problem where the training set $D_N = \{\langle X_1, y_1 \rangle, \langle X_2, y_2 \rangle, \dots, \langle X_N, y_N \rangle\}$ is made of N pairs $\langle X_i, y_i \rangle \in \mathcal{X} \times \mathcal{Y}$ i.i.d. distributed according to the joint distribution $P(\langle X, y \rangle) = P(y|X)P(X)$. Let us define a *learning machine* by the following components:

- a parametric class of *hypothesis* functions $h(s, \alpha^s)$ with $\alpha^s \in \Lambda^s$, where $s \subseteq X$,
- a *cost* function $C(y, h) \geq 0$ such that $C(y, h) = 0$ only if $y = h$,
- an *algorithm* of parametric identification that for a given subset $s \subseteq X$ and a given training set D_N returns a hypothesis function $h(\cdot, \alpha^s_{D_N})$ with $\alpha^s_{D_N} \in \Lambda^s$ such that $\sum_{\langle s,y \rangle \in D_N} C\left(y, h(s, \alpha^s_{D_N})\right) \leq \sum_{\langle s,y \rangle \in D_N} C\left(y, h(s, \alpha^s)\right)$ for all $\alpha^s \in \Lambda^s$.

F.s. and stochastic discrete optimization

We may formulate the feature selection problem as a discrete optimization problem [5]

$$\min_{\mathbf{s}\in S} \mathsf{GE}(\mathbf{s}) =$$
$$= \min_{\mathbf{s}\in S} \left\{ E_{\mathbf{D}_N} \left[\int_{\mathcal{X}} \int_{\mathcal{Y}} C(y, h(s, \alpha^s_{\mathbf{D}_N})) dP(y|s) dP(s) \right] \right\} \quad (4)$$

where the generalization error $\mathsf{GE}(s)$ of the subset $s \subseteq X$ is not observed directly but estimated by the cross-validation measure $\widehat{\mathsf{GE}}(s)$.

Let

$$s^* = \arg\min_{\mathbf{s}\in S} g(s), \quad \mathsf{GE}^* = \min_{\mathbf{s}\in S} g(s) = g(s^*) \quad (5)$$

be the optimal solution of the feature selection problem and the relative optimal generalization error, respectively.

- Unfortunately the GE for a given s is not directly measurable but can only be estimated by the quantity $\widehat{\text{GE}}(s)$ which is an unbiased estimator of $\text{GE}(s)$.
- The feature selection problem may be formulated in terms of a stochastic optimization problem where the selection of the best subset s has to be based on a sample estimate $\widehat{\text{GE}}$.
- The wrapper approach to feature selection aims to return the minimum $\hat{s}$ of a cross-validation criterion $\widehat{\text{GE}}(s)$

$$\hat{s} = \arg\min_{s \in S} \widehat{\text{GE}}(s) = \mathcal{W}(D_N) \tag{6}$$

- This algorithm can be considered as a mapping from the space of datasets of size N to the space S of subsets of X. Since $\mathbf{D}_N$ is a random variable, the variable $\hat{\mathrm{s}}$ is random too.
- The generalisation accuracy of a learner where the feature subset $\hat{s}$ has been selected by feature selection is given by the quantity $E[\text{GE}(\hat{\mathrm{s}})]$.

Open problems

Now two main problems appear:

1. According to the relations above, the quantity $\widehat{\text{GE}}(\hat{s})$, returned by the cross-validation assessment of the wrapper, is a biased estimate both of the minimum $\text{GE}(s^*)$ and of the generalization performance $E[\text{GE}(\hat{\mathbf{s}})]$.
2. Being the state space very large, we could expect a very large variance of $\hat{\mathbf{s}}$ and consequently a very large value of $E[\text{GE}(\hat{\mathbf{s}})]$.

Proposed solutions:

1. An unbiased estimate $\widetilde{\text{GE}}$ of the generalization performance $E[\text{GE}(\hat{\mathbf{s}})]$ can be obtained by using an external cross-validation loop.
2. The issue of large variance of $\hat{\mathbf{s}}$ may be addressed by structuring the search space S.

Our proposal

In order to better control the bias/variance trade-off [1] proposes, accordingly to what is done in model selection tasks, to structure the space S into a nested sequence of spaces $S_1 \subset \cdots \subset S_n = S$ where $S_j = \{s : |s| \leq j\}$.

The approach consists in running in parallel n wrapper strategies $\mathcal{W}_j$, $j = 1, \ldots, n$, each constrained to search in the space S_j. Each wrapper strategy returns a subset $\hat{s}_j \in S_j$, made of $|\hat{s}_j| \leq j$ features.

The expected generalization error of each strategy is measured by $\widetilde{\text{GE}}(\mathcal{W}_j)$.

The outcome of the structured wrapper algorithm can be obtained either by winner-takes-all policy

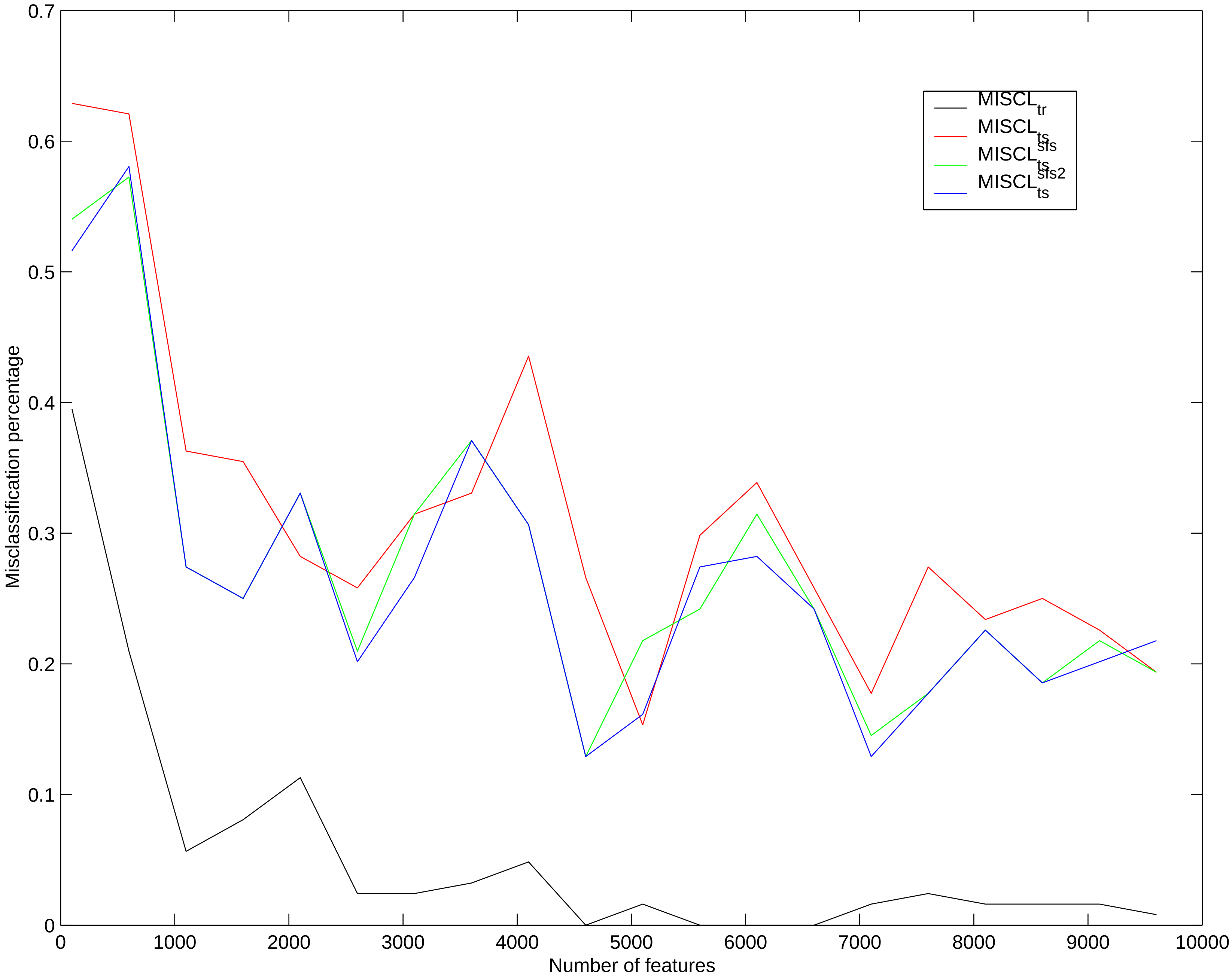
$$\tilde{s} = \hat{s}_{\tilde{j}}, \quad \text{where } \tilde{j} = \arg \min_{j=1,\ldots,n} \widetilde{\text{GE}}(\mathcal{W}_j) \tag{7}$$

or by combining the models associated to the best B subsets, e.g. by using a weighted average of their predictions

Multiclass classification example

- Let us consider the gene expression cancer dataset ALL. This dataset covers six types of acute lymphoblastic leukemia (ALL).
- The number of samples is $N = 248$ and the number of features (i.e. the number of probes) is $n = 12558$.
- We use a Nearest Neighbour classifier.
- We partition the dataset in a training and a test set.
- The training set is used to perform forward feature selection by three-fold cross-validation.
- We compare the results of the conventional forward selection and a structural forward selection

Example



0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1000
2000
3000
4000
5000
6000
7000
8000
9000
10000
Number of features
Misclassification percentage
$MISCL_{tr}$
$MISCL_{ts}$
$MISCL_{ts}^{sfs}$
$MISCL_{ts}^{sfs2}$

Some considerations

- Bioinformatics applications are known to be characterized by highly noisy data.
- The huge size of the feature space compared to the number of samples makes hard the problem in terms of:

  Optimization techniques to explore the feature space.

  Large variance of the resulting model.
- Biologists asks for prediction accuracy but mainly for causual intepretation (gene signature).
- Biologists are scared of unstable feature selection procedures which change the sets of relevant genes simply by adding more observations. Examples are clinical study with different populations of patients.
- Filtering techniques are computational efficient and robust against overfitting. They may introduce bias but may have considerably less variance.
- Literature have been inflationed by over-optimistic results.

Closing remarks

- Let us not forget that any learner is an estimator and as such any outcome it returns, is a **random variable**.
- If the outcome of our feature selection technique is a set of variables, this set is also a random set.
- Data miners are used to return confidence interval on accuracy.
- They should start returning “confidence intervals” also on feature (gene) subsets.

References

[1] G. Bontempi. Structural feature selection for wrapper methods. In *Proceedings of ESANN 2005, European Symposium on Artificial Neural Networks*, 2005.

[2] H.J. Bussemaker, H. Li, and E. D. Siggia. Regulatory element detection using correlation with expression. *Nature Genetics*, 27:167–171, 2001.

[3] D. R. Cox. Regression models and life tables. *Journal of the Royal Statistical Society Series B*, 34:187–220, 1972.

[4] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[5] A. J. Kleywegt, A. Shapiro, and T. Homem de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal of Optimization*, 12:479–502, 2001.

[6] W.K. Mak, D.P. Morton, and R.K. Wood. Monte carlo bounding techniques for determining solution quality in stochastic programs. *Operations Research Letters*, 24:47–56, 1999.

[7] E. Mjolsness, T. Mann, R. Castano, and B. Wold. From coexpression to coregulation: an approach to inferring transcriptional regulation among gene classes from large-scale expression data. In S.A.Solla, editor, *NIPS- Advances in Neural Information Processing Systems 12*, pages 928–934. MIT Press, 2000.

[8] E. Mjolsness, D.H. Sharp, and J. Reinitz. A connectionist model of development. *Journal Theoretical Biology*, 152:429–454, 1991.

[9] T. M. Therneau and P. M. Grambsch. *Modeling Survival Data: Extending the Cox Model*. Springer, 2000. ISBN: 0-387-98784-3.

[10] S. L.M. van der Keles and M. B. Eisen. Identification of regulatory elements using a feature selection method. *Bioinformatics*, 18:1167–1175, 2002.

[11] L. J. van't Veer, H. Dai, and M. J. van de Vijver. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.