A selecting-the-best method for budgeted model selection

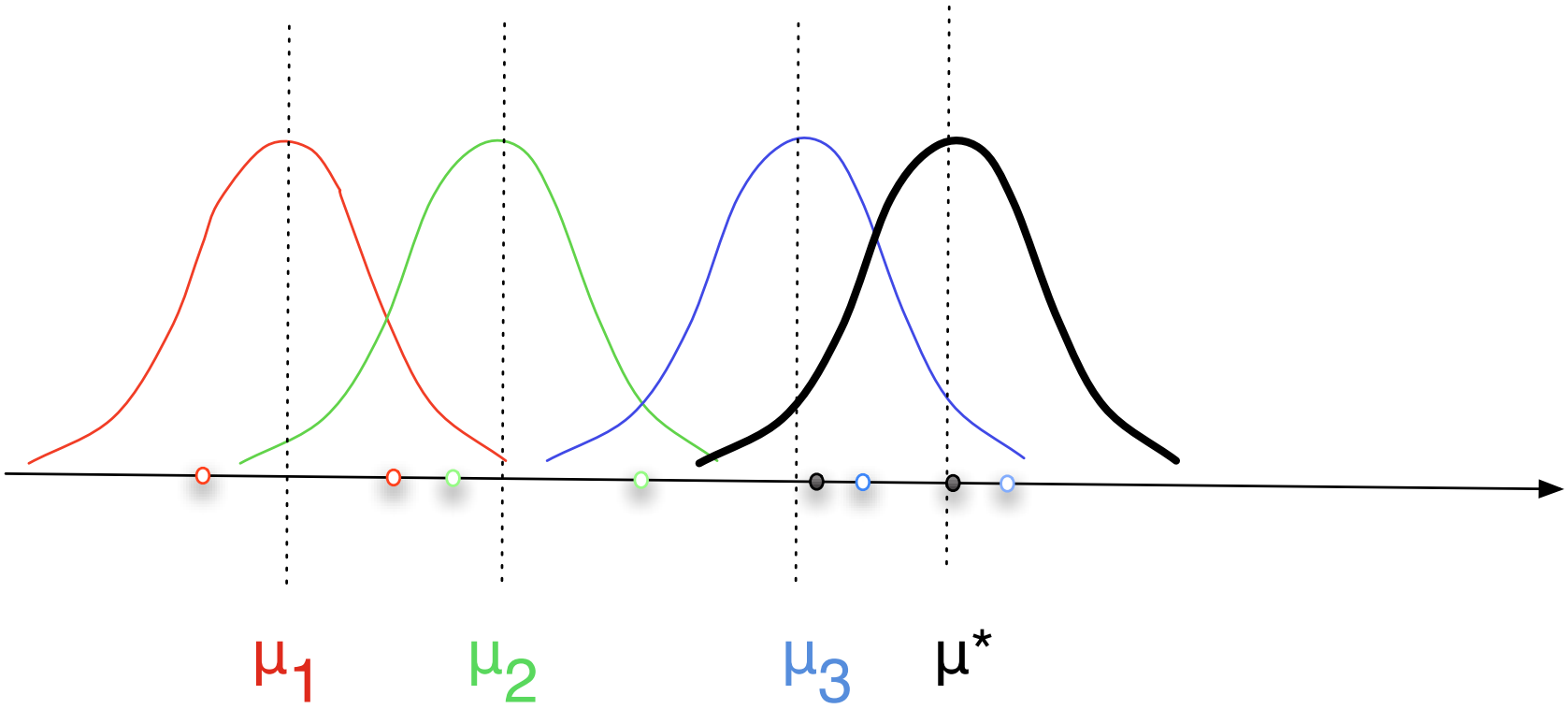
Gianluca Bontempi and Olivier Caelen

Machine Learning Group, Département d'Informatique,
Faculté des Sciences, Université Libre de Bruxelles, Belgium

Context and motivation

- A priori no learning algorithm or model structure is the best one.
- This makes of model selection a necessary step in statistical modeling and machine learning.
- Nowadays we are confronted with the need of using model selection in contexts where the number of alternatives is huge and sometimes much larger than the number of samples, for instance in
  - supervised learning, where the model designer is confronted with a huge number of model families and paradigms
  - feature selection (e.g. in bioinformatics or text mining) where the problems of selecting the best subset among a large set of inputs can be cast in the form of a model selection problem
- In this setting it interesting to address the "budgeted active model selection" problem [3] where we want to use a fixed budget of assessments (e.g. leave-one-out errors) to identify which of a given set of model alternatives has the highest expected accuracy.

Selection in a stochastic setting



4 alternatives. We want to select the one with the largest expected value on the basis of sampled means.
Greedy selection after spending a budget of $L = 8$ samples returns the blue alternative instead of the best one.

Multi-armed bandit and selecting-the-best

- The multi-armed bandit problem is an exploration/exploitation problem in which a player has to decide which arm of a slot machine to pull to maximize the total reward in a series of rounds. Each arms returns a random reward unknown to the player.
- However, the paradigm of the bandit problem is not the the most adequate manner of interpreting a model selection problem.
- The difference between a budgeted model selection problem and a bandit problem is that in model selection there is a pure exploration phase (characterised by spending the budget to assess the different alternatives) followed by a pure exploitation step (i.e. the selection of the best model). This is not the case of typical bandit problem where there is an immediate exploitation (and consequent reward) following each exploration action.
- We focus on another computation framework extensively dealt with by the community of stochastic simulation: the "selecting-the-best" problem [2, 1].

Contribution

- We propose an approach based on the notion of *probability of correct selection*, a notion borrowed from the domain of Monte Carlo stochastic approximation.
- We estimate from data the expected gain of a greedy selection and define a sampling rule which maximises such quantity.
- Analytical results in the case of two alternatives are extended to a larger number of alternatives by using the Clark's approximation of the maximum of a set of random variables.
- Preliminary results on synthetic and real model selection tasks show that the technique is competititive with state-of-the-art algorithms, like the bandit UCB.

Formal setting

- Let us consider a set $\{\mathbf{z}_k\}$ of K model alternatives. We model their performance (e.g. accuracy) by independent random variables $\mathbf{z}_k$ with mean μ_k and standard deviation σ_k.
- The model selection goal is to select the model with the highest accuracy

$$k^* = \arg\max \mu_k, \qquad \mu^* = \max \mu_k \tag{1}$$

- Let us consider a sequential budgeted setup where at most L assessments are allowed, $L/K > 1$ is small and, at each round l, one alternative $k_l \in \{1, \ldots, K\}$ is chosen and assessed.
- We wish to design a strategy of samplings k_l such that, if we select the alternative with the highest sampled mean when the L assessments are done, the probability of selecting k^* is maximal.

- Let $\mathbf{z}_k^i$ be the ith observation of the kth alternative and $n_k(l)$ the number of times that the kth alternative was selected during the l first rounds.
- A greedy selection at the lth step returns the alternative

$$\hat{\mathbf{k}}_g^l = \arg\max \hat{\boldsymbol{\mu}}_k^l \quad , \text{ where } \hat{\boldsymbol{\mu}}_k^l = \frac{\sum_{i=1}^{n_k(l)} \mathbf{z}_k^i}{n_k(l)}.$$

- Let

$$p_k^l = \text{Prob}\left\{\hat{\mathbf{k}}_g^l = k\right\}, \qquad k = 1, \ldots, K$$

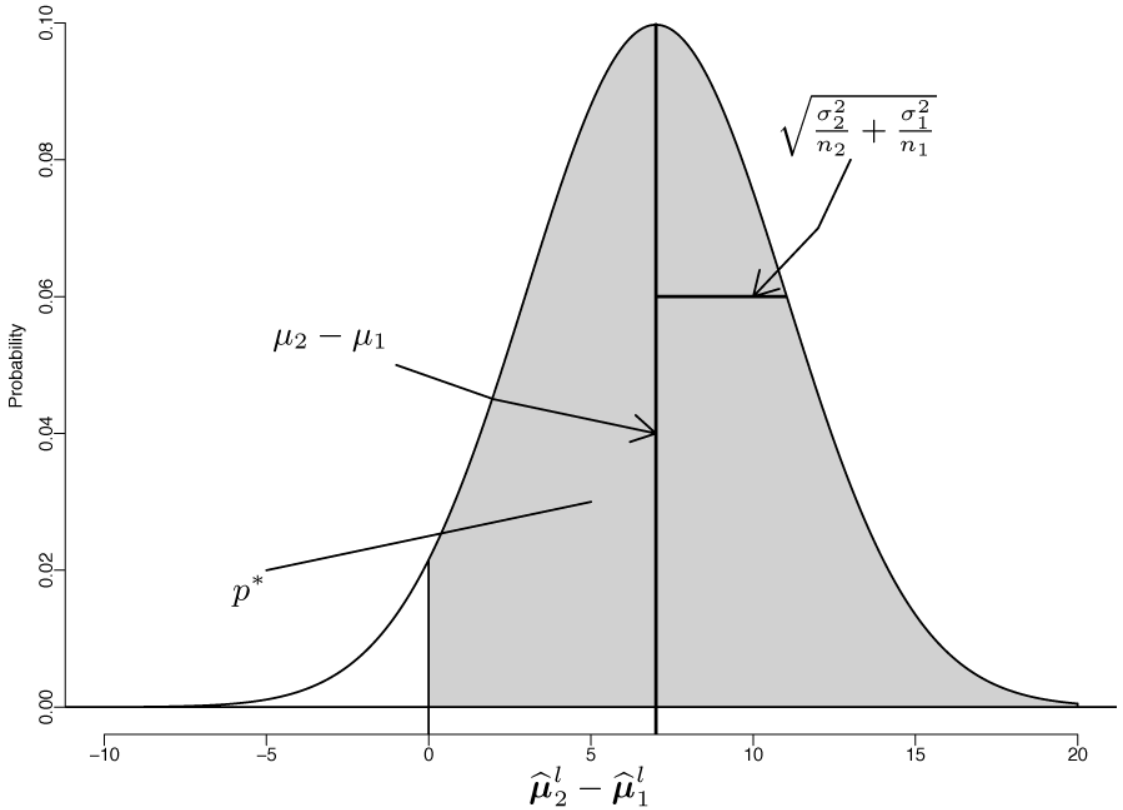
be the probability that the greedy selection returns the kth alternative at the lth step and

$$p_{k^*}^l = \text{Prob}\left\{\hat{\mathbf{k}}_g^l = k^*\right\}$$

the probability of making the correct selection at the lth step.

The $K = 2$ Gaussian case

Let $K = 2$ and $\mu_2 > \mu_1$. The selection is correct if $\hat{\mu}_2^l > \hat{\mu}_1^l$



Distribution of $\hat{\mu}_2^l - \hat{\mu}_1^l$. The grey area represents the probability of correct selection at the lth step.

- Since the estimation of p_k^I is difficult for $K > 2$, we prefer to focus on the expected gain of a greedy selection

$$\mu_g^I = \sum_{k=1}^{K} p_k^I \mu_k \tag{2}$$

- We wish to design an exploration strategy which maximises the expected gain of a greedy selection.
- We need to estimate both the terms p_k and μ_k. While the plug-in estimation of μ_k is trivial, deriving p_k from data is much more complex and requires either a numeric or Monte Carlo procedure.
- We avoid having recourse to a multivariate estimation by approximating the $K > 2$ variate problem with a bivariate problem, since for $K = 2$ Gaussian alternatives the analysis of the term μ_g^I is much easier and some interesting properties can be derived analytically.

The $K = 2$ Gaussian case

If $\mathbf{z}_1 \sim \mathcal{N}(\mu_1, \sigma_1)$ and $\mathbf{z}_2 \sim \mathcal{N}(\mu_2, \sigma_2)$ are two independent Gaussian distributed alternatives and μ_g^l the expected greedy gain at step l then

Theorem

Let $\mu_g^{l+1}(k)$ denote the value of the expected greedy gain at step $l+1$ if the kth alternative is tested at step l. Then

$$\forall k \in \{1, 2\}, \quad \mu_g^{l+1}(k) \geq \mu_g^l$$

Theorem

Let $n_1 = n_1(l)$, $n_2 = n_2(l)$ the respective number of collected observations at step l. The sampling rule

$$k_l = \begin{cases} 1 & \text{if } N_\Delta < 0 \\ 2 & \text{if } N_\Delta > 0 \\ \text{random} & \text{if } N_\Delta = 0 \end{cases} \tag{3}$$

where

$$N_\Delta = n_1(n_1+1)(\sigma_2^2 - \sigma_1^2) + \sigma_1^2(n_2 + n_1 + 1)(n_1 - n_2) \tag{4}$$

maximises the value of μ_g at the step $l+1$.

The $K = 2$ optimal sampling strategy

- The first theorem implies that when only two alternatives are in competition, testing one of them leads invariably to an increase of the expected gain.
- The second theorem states that, though sampling any of the two alternatives brings to an increase of the expected gain, this increase is not identical.
- It is possible then to define an optimal sampling strategy for $K = 2$
- SRule($\mu_1, \sigma_1, n_1, \mu_2, \sigma_2, n_2$)

  Input: μ_1, σ_1: parameters of the first alternative
  μ_2, σ_2: parameters of the second alternative
  Compute N_Δ by Equation (4)
  if $N_\Delta \leq 0$ **then**
  Sample alternative 1
  end if
  if $N_\Delta > 0$ **then**
  Sample alternative 2
  end if

The $K > 2$ case

- For $K = 2$ Gaussian alternatives we may control the evolution of the expected gain .
- This is no longer the case for $K > 2$ where the evolution of μ_g is much less predictable.
- We transform the problem of selection among $K > 2$ alternatives into an approximate problem where only two configurations are given.
- If we have to select the best alternative among $K > 2$ configurations, the problem can be reformulated as the problem of choosing between one configuration and the maximum of all the others. By doing in this way we reduce the problem to a two-configuration setting where one of the alternatives is assessed against the maximum of the remaining $K - 1$ ones.
- We adopt the Clark approximation, which is a fast and elegant way to approximate the maximum of a set of Gaussian random variables without having recourse to analytical or numerical integration.

Clark approximation

- The method approximates the distribution of the maximum of K normal variables. Consider the mean μ_m and the variance σ_m^2 of $\mathbf{z}_m = \max\{\mathbf{z}_1, \mathbf{z}_2\}$.
- Let

$$a^2 = \sigma_1^2 + \sigma_2^2 - 2\sigma_1\sigma_2\rho_{12}, \qquad z = \frac{\mu_1 - \mu_2}{a}$$

where ρ_{12} is the correlation coefficient between $\mathbf{z}_1$ and $\mathbf{z}_2$, it is possible to show that

$$\begin{aligned}\mu_m &= \mu_1\Phi(z) + \mu_2\Phi(-z) + a\phi(z)\\ \sigma_m^2 &= [(\mu_1^2 + \sigma_1^2)\Phi(z) + (\mu_2^2 + \sigma_2^2)\Phi(-z) + \\ &\quad + (\mu_1 + \mu_2)a\phi(z)] - \mu_m^2\end{aligned}$$

where ϕ is the standard normal density function and Φ is the associated cumulative distribution.

- Consider a third variable $\mathbf{z}_3 \sim \mathcal{N}(\mu_3, \sigma_3^2)$. The distribution of $\mathbf{z}_M = \max\{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3\} \approx \max\{\mathbf{z}_m, \mathbf{z}_3\}$ is obtained by using iteratively the procedure sketched above for 2 variables.

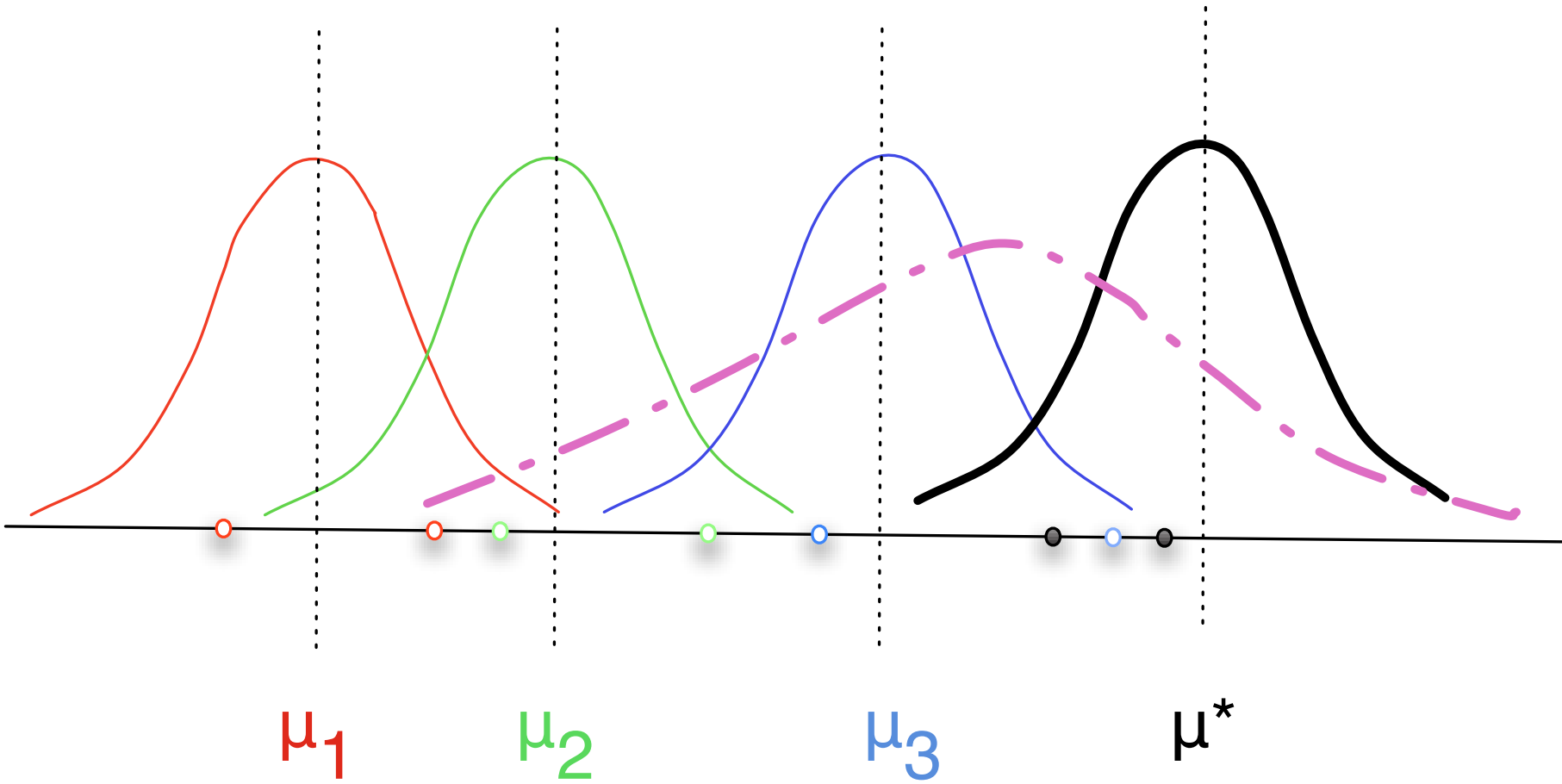
$K > 2$ sampling strategy

SELBEST: Budgeted model selection with Clark approximation

Input: K: number of alternatives,
$\{Z_k\}$: the observations of K random variables,
I: number of inititalization steps

Initialization sampling of the K alternatives
for $l = 1$ **to** L **do**
  for $k = 1$ **to** K **do**
    Compute $\hat{\mu}_k, \hat{\sigma}_k$
  end for
  $[k] \leftarrow$ decreasing order of $\hat{\mu}_k$
  for $k = 1$ **to** $K - 1$ **do**
    Compute $\hat{\mu}_m, \hat{\sigma}_m$ by Clark approximation
    where $\hat{\mathbf{z}}_m \approx \max\{\hat{\mathbf{z}}_{[k+1]}, \ldots, \hat{\mathbf{z}}_{[K]}\}$
    $s \leftarrow$SRule$(\mu_{[k]}, \sigma_{[k]}, n_{[k]}, \hat{\mu}_m, \hat{\sigma}_m, \max_{j=k+1}^{K} n_{[j]})$
    if $s = 1$ **then**
      $k_l \leftarrow [k]$
      break;
    end if
    if $(s = 2)$ AND $(k = K - 1)$ **then**
      $k_l \leftarrow [K]$
    end if
  end for
  Sample z_{k_l}
  $n_{[k_l]} \leftarrow n_{[k_l]} + 1$
end for
return $\arg\max_k \hat{\mu}_k$

The $K > 2$ case



Dashed line is the Clark approximation of $\max\{\mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_3\}$.

Experiments

We compared the SELBEST algorithm in synthetic and real sequential selection problems with three reference methods:

1. a greedy algorithm which for each l samples the model with the highest estimated performance

$$k_l = \arg\max \hat{\mu}_k$$

2. an interval estimation algorithm which samples the model with the highest upper value of the confidence interval

$$k_l = \arg\max \left(\hat{\mu}_k + t_{0.05, n_k(l)-1} \hat{\sigma}_k \right)$$

where $t_{0.05,n}$ is the upper 95% upper critical point of the Student distribution with n degrees of freedom and

3. the UCB bandit algorithm which implements the following sampling strategy

$$\hat{k}^l = \arg\max \left(\hat{\mu}_k + \sqrt{\frac{2 \log l}{n_k(l)}} \right)$$

Synthetic experiments

The synthetic experimental setting consists of 5000 experiments of three types:

1. K independent Gaussian alternatives where K is uniformly sampled in the interval $[10, 200]$. $\mu_k \sim \mathcal{U}(0, 1)$ and $\sigma_k \sim \mathcal{U}(0.5, 1)$, $k = 1, \ldots, K$.
2. like the first one with the only difference that standard deviations $\sigma_k, k = 1, \ldots, K$ are obtained by sampling the distributions $\mathcal{U}(1, 2.5)$.
3. a non Gaussian configuration, where the K alternatives have a chi-squared distribution of degree thirty.The number K is uniformly sampled within the interval $[10, 200]$ and $\mu_k \sim \mathcal{U}(1, 2)$.

Synthetic experiments: average regrets over 5000 experiments

L	SELBEST	GREEDY	IE	UCB
20	0.125	**0.137**	**0.136**	0.12
40	0.115	**0.129**	**0.13**	0.115
60	0.108	**0.123**	**0.125**	**0.111**
80	0.102	**0.119**	**0.121**	**0.108**
100	0.097	**0.114**	**0.118**	**0.105**
120	0.093	**0.11**	**0.115**	**0.102**
140	0.09	**0.106**	**0.112**	**0.1**
160	0.087	**0.103**	**0.109**	**0.098**
180	0.084	**0.1**	**0.106**	**0.096**
200	0.082	**0.097**	**0.104**	**0.094**

L	SELBEST	GREEDY	IE	UCB
20	0.267	**0.282**	**0.284**	0.26
40	0.254	**0.272**	**0.275**	0.249
60	0.243	**0.266**	**0.269**	0.242
80	0.233	**0.26**	**0.264**	0.235
100	0.225	**0.254**	**0.258**	**0.229**
120	0.218	**0.25**	**0.254**	**0.224**
140	0.212	**0.245**	**0.25**	**0.219**
160	0.206	**0.24**	**0.246**	**0.215**
180	0.201	**0.236**	**0.242**	**0.212**
200	0.196	**0.232**	**0.238**	**0.208**

L	SELBEST	GREEDY	IE	UCB
20	0.082	**0.092**	**0.089**	0.082
40	0.075	**0.086**	**0.084**	**0.079**
60	0.07	**0.082**	**0.08**	**0.075**
80	0.066	**0.078**	**0.077**	**0.073**
100	0.063	**0.075**	**0.074**	**0.071**
120	0.06	**0.072**	**0.072**	**0.069**
140	0.058	**0.069**	**0.07**	**0.067**
160	0.056	**0.067**	**0.068**	**0.066**
180	0.054	**0.065**	**0.066**	**0.065**
200	0.053	**0.063**	**0.064**	**0.064**

Real data experiments

- The real data experimental setting consists of a set of feature selection problems applied to 26 UCI regression tasks.
- Each alternative corresponds to a feature set with 5 inputs and the assessment is done by using a specific learner (i.e. a KNN) as a measure of the performance of the feature set.
- We partition each dataset in two parts. The first part is used for assessment and the second is used to compute what we consider a reliable estimate of the generalization accuracy $\mu_k, k = 1, \ldots, K$ of the K alternatives.
- The number L of assessments takes values in the set $\{20, 40, \ldots, 200\}$, $I = 20$ and the number of alternatives is uniformly sampled in the range $K \in [60, 100]$.

Real data experiment

L	SELBEST	GREEDY	IE	UCB
20	19.07	**19.68**	**19.78**	**19.59**
40	18.56	**19.56**	**19.74**	**19.56**
60	18.17	**19.24**	**19.65**	**19.42**
80	17.94	**19.27**	**19.68**	**19.27**
100	17.55	**18.99**	**19.59**	**19.12**

Average relative regrets over the 26 datasets and over 500 repetitions for L total assessments. Bold notation is used for regrets which are statistically significantly worse (paired permutation test pv<0.05) than the SELBEST performance.

Discussion

- comparison wrt GREEDY and IE: the SELBEST technique is significantly better for all ranges of L.
- comparison wrt UCB: SELBEST and UCB are comparable for low values of L, while SELBEST is better when L increases.
  - Interpretation: the bandit technique has not been designed for budgeted model selection tasks and its effectiveness is reduced when the exploration phase is sufficiently large to take advantage of selecting-the-best approaches.

Conclusion and perspectives

- Model selection is more and more confronted with tasks characterised by a huge number of alternatives with respect to the amount of learning data and the time allowed for assessment (e.g. bioinformatics, text mining, large-scale optimization problems , Monte Carlo tree searches in games).
- New challenges ask for techniques able to compare stochastic configurations more rapidly and with a smaller budget of observations than conventional leave-one-out techniques.
- Bandit techniques have not been conceived for problems where assessment and selection are sequential and not intertwined like in multi-armed problems.
- We proposed an alternative family of approaches which relies on notions of stochastic optimisation and low variate approximation of a large number of alternatives.
- The promising results open the way to additional validations in real configurations like feature selection in large feature to sample ratio dataset, notably in bioinformatics and text mining.


S. Kim and B. Nelson.
Handbooks in Operations Research and Management Science: Simulation, chapter Selecting the Best System.
Elsevier Science, 2006.


A.M. Law and W.D. Kelton.
Simulation Modeling & analysis.
McGraw-Hill International, second edition, 1991.


O. Madani, D. Lizotte, and R. Greiner.
Active model selection.
In *Proceedings of the Proceedings of the Twentieth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-04)*, pages 357–365. AUAI Press, 2004.