Noname manuscript No.
(will be inserted by the editor)

Statistical comparison of classifiers through Bayesian hierarchical modelling

Giorgio Corani · Alessio Benavoli · Janez Demšar · Francesca Mangili · Marco Zaffalon

Received: date / Accepted: date

Abstract Usually one compares the accuracy of two competing classifiers using null hypothesis significance tests. Yet such tests suffer from important shortcomings, which can be overcome by switching to Bayesian hypothesis testing. We propose a Bayesian hierarchical model that jointly analyzes the cross-validation results obtained by two classifiers on multiple data sets. The model estimates more accurately the difference between classifiers on the individual data sets than the traditional approach of averaging, independently on each data set, the cross-validation results. It does so by jointly analyzing the results obtained on all data sets, and applying shrinkage to the estimates. The model eventually returns the posterior probability of the accuracies of the two classifiers being practically equivalent or significantly different.

1 Introduction

The statistical comparison of learning algorithms is fundamental in machine learning; it is typically carried out through hypothesis testing. In this paper we assume that one is interested in comparing the accuracy of two learning algorithms for classification (referred to as *classifiers* in the following). However our discussion readily applies to any other measure of performance.

Assume that two classifiers have been assessed via cross-validation on a single data set. The recommended approach for comparing them is the correlated t-test (Nadeau & Bengio, 2003). If instead one aims at comparing two classifiers on multiple data sets the recommended test is the signed-rank test (Demšar, 2006). Both tests are based on the frequentist framework of the null-hypothesis significance tests (nhst), which has severe drawbacks.

First, the nhst computes the probability of getting the observed (or a larger) difference in the data if the null hypothesis was true. It does *not* compute the probability of interest, which is the probability of one classifier being more accurate than another given the observed results.

Second, the claimed statistical significances do not necessarily imply practical significance, since null hypotheses can be easily rejected by increasing the number of observations (Wasserstein & Lazar, 2016). Thus for instance the signed-rank can reject the null hypothesis when dealing with

G. Corani, A. Benavoli F. Mangili and M. Zaffalon are with
Istituto Dalle Molle di studi sull'Intelligenza Artificiale (IDSIA), Manno, Switzerland
J. Demšar is with
Faculty of Computer and Information Science, University of Ljubljana, Slovenia
E-mail:
giorgio{alessio,francesca,zaffalon}@idsia.ch
janez.demsar@fri.uni-lj.si

two classifiers whose accuracies are nearly equal, but which have been compared on a large number of data sets.

Third, when the null hypothesis is not rejected, we cannot assume the null hypothesis to be true (Kruschke, 2015, Chap. 11). Thus nhst tests cannot recognize equivalent classifiers.

These issues can be overcome by switching to Bayesian hypothesis testing (Kruschke, 2015, Sec. 11), which are recently being applied also in machine learning (Lacoste et al., 2012; Corani & Benavoli, 2015; Benavoli et al., under review).

Let us denote by δ_i the actual difference of accuracy between the two classifiers on the i-th data set. Usually δ_i is estimated via cross-validation. We propose the first model that represents both the distribution $p(\delta_i)$ across the different data sets and the distribution of the cross-validation results on the i-th data set given δ_i.

Following Kruschke (2013) we analyze the results by adopting a region of practical equivalence (rope). In particular we consider two classifiers to be practically equivalent if their difference of accuracy belongs to the interval $(-0.01, 0.01)$. This mitigates the risk of claiming significance because of a thin difference of accuracy in simulation, which is likely to be swamped by other sources of uncertainty when the classifier is adopted in practice (Hand et al., 2006). There are however no correct rope limits; thus other researchers might set the rope differently. Based on the rope we compute the posterior probability of the two classifiers being practically equivalent or significantly different. Such probabilities convey meaningfully information even when they do not exceed the 95% threshold: this is a more informative outcome than that of a nhst.

Moreover, the hierarchical model estimates the δ_i's more accurately than the traditional approach of computing, independently on each data set, the mean of the cross-validation differences. It does so by jointly estimating the δ_i's and *shrinking* them towards each other. We prove theoretically that such shrinkage yields lower estimation error than the traditional approach.

2 Existing approaches

Let us introduce some notation. We have a collection of q data sets; the actual mean difference of accuracy between the two classifiers on the i-th data set is δ_i. We can think of δ_i as the average difference of accuracy that we would obtain by repeating many times the procedure of sampling from the actual distribution as many instances as there are in the actually available data set, train the two classifiers and measure the difference of accuracy on a large test set.

Usually δ_i is estimated via cross-validation. Assume that we have performed m runs of k-fold cross-validation on each data set, using the same folds for both classifiers. The *differences of accuracy* on each fold of cross-validation are $\boldsymbol{x_i} = \{x_{i1}, x_{i2}, \ldots, x_{in}\}$, where $n = mk$. The mean and the standard deviation of the results on the i-th data set are $\bar{x}_i$ and s_i. The mean of the cross-validation results is also the maximum likelihood estimator (MLE) of δ_i.

The values $\boldsymbol{x_i} = \{x_{i1}, x_{i2}, \ldots, x_{in}\}$ are correlated because of the overlapping training sets built during cross-validation. In particular, there is a) correlation among folds within the same run of cross-validation and b) correlation among folds from different cross-validation runs. Nadeau & Bengio (2003) proposed $\rho = \frac{1}{k}$ (k is the number of folds) as an approximated estimation of the correlation when repeated random train/test splits are adopted. This validation method is slightly different from cross-validation, but such heuristic is generally used (Witten et al., 2011, Chap. 5.5) also for the modeling the correlations (a) and (b) of cross-validation, given the lack of better options. The statistic of the correlated t-test is thus:

$$t = \overline{x}_i / \sqrt{\hat{s}_i^2 (\frac{1}{n} + \frac{\rho}{1-\rho})}. \tag{1}$$

The denominator of the statistic is the standard error, which is informative about the accuracy of $\bar{x}_i$ as an estimator of δ_i. The standard error of the correlated t-test accounts for the correlation of

cross-validation results. The statistic of Eqn.(1) follows a t distribution with n-1 degrees of freedom. When the statistic exceeds the critical value, the test claims δ_i to be significantly different from zero. This is the standard approach for comparing two classifiers on a single data set.

The signed-rank test is instead the recommended method (Demšar, 2006) to compare two classifiers on a collection of q different data sets. It is usually applied after having performed cross-validation on each data set. The test analyzes the mean differences measured on each data set $(\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_q)$ assuming them to be i.i.d.. This is a simplistic assumption: the $\bar{x}_i$'s are not i.i.d. since they are characterized by different uncertainty; indeed their standard errors are typically different.

The test statistic is:

$$T^+ = \sum_{\{i:\ \bar{x}_i \geq 0\}} r_i(|\bar{x}_i|) = \sum_{1 \leq i \leq j \leq n} T^+_{ij},$$

where $r_i(|\bar{x}_i|)$ is the rank of $|\bar{x}_i|$ and

$$T^+_{ij} = \begin{cases} 1 \ \textit{if} \ \bar{x}_i \geq \bar{x}_j, \\ 0 \ \textit{otherwise.} \end{cases}$$

For a large enough number of samples (e.g., q>10), the statistic under the null hypothesis is normally distributed. When the test rejects the null hypothesis, it claims that the median of the population of the δ_i's is different from zero.

The two tests discussed so far are null-hypothesis significance test (nhst) and as such they suffer from the drawbacks discussed in the Introduction.

Let us now consider the Bayesian approaches. Kruschke (2013) presents a Bayesian t-test for i.i.d. observations, which is thus not suitable for analyzing the correlated cross-validation results. The Bayesian correlated t-test (Corani & Benavoli, 2015) is instead suitable. It computes the posterior distribution of δ_i on a *single* data set, assuming the cross-validation observations to be sampled from a multivariate normal distribution whose components have the same mean δ_i, the same standard deviation σ_i and are equally cross-correlated with correlation $\rho = \frac{1}{k}$.

As for the analysis of multiple data sets, Lacoste et al. (2012) models each data set as an independent Bernoulli trial. The two possible outcomes of the Bernoulli trial are the first classifier being more accurate than the second or vice versa. This approach yields the posterior probability of the first classifier being more accurate than the second classifier on more than half of the q data sets. A shortcoming is that its conclusions apply only to the q available data sets without generalizing to the whole population of data sets.

3 The hierarchical model

We propose a Bayesian hierarchical model for comparing two classifiers. Its core assumptions are:

$$\delta_1, ..., \delta_q \sim t(\delta_0, \sigma_0, \nu), \tag{2}$$

$$\sigma_1, ..., \sigma_q \sim \text{unif}(0, \bar{\sigma}), \tag{3}$$

$$\mathbf{x}_i \sim MVN(\mathbf{1}\delta_i, \mathbf{\Sigma_i}). \tag{4}$$

The i-th data set is characterized by the mean difference of accuracy δ_i and the standard deviation σ_i. Thus we model each data set as having its own estimation uncertainty. Notice that instead the signed-rank test simplistically assumes the $\bar{x}_i$'s to be i.i.d.

The δ_i's are assumed to be drawn from a Student distribution with mean δ_0, scale factor σ_0 and degrees of freedom ν. The Student distribution is more flexible than the Gaussian, thanks to the additional parameter ν. When ν is small, the Student distribution has heavy tails; when ν is above 30, the Student distribution is practically a Gaussian. A Student distribution with low degrees of

freedom robustly deals with outliers and for this reason is often used for robust Bayesian estimation (Kruschke, 2013).

We assume σ_i to be drawn from a uniform distribution over the interval $(0, \bar{\sigma})$. This prior (Gelman, 2006) yields inferences which are insensitive to the value of $\bar{\sigma}$ if $\bar{\sigma}$ is large enough. We adopt $\bar{\sigma} = 1000\bar{s}$, where $\bar{s}$ is the mean standard deviation observed on the different data sets ($\bar{s} = \sum_i^q s_i/q$).

Equation (4) models the fact that the cross-validation measures $\boldsymbol{x_i} = \{x_{i1}, x_{i2}, \ldots, x_{in}\}$ of the i-th data set are generated from a multivariate normal whose components have the same mean (δ_i), the same standard deviation (σ_i) and are equally cross-correlated with correlation ρ. Thus the covariance matrix $\mathbf{\Sigma_i}$ is patterned as follows: each diagonal elements equals σ_i^2; each non-diagonal element equals $\rho\sigma_i^2$. Such assumptions are borrowed from the Bayesian correlated t-test (Corani & Benavoli, 2015).

We complete the model with the prior on the parameters δ_0, σ_0 and ν of the high-level distribution. We assume δ_0 to be uniformly distributed within 1 and -1. This choice works for all the measures bounded within ± 1, such as accuracy, AUC, precision and recall. Other type of indicators might require different bounds.

For the standard deviation σ_0 we adopt the prior $unif(0, \bar{s_0})$, with $\bar{s_0} = 1000 s_{\bar{x}}$, where $s_{\bar{x}}$ is the standard deviation of the $\bar{x}_i$'s.

As for the prior $p(\nu)$ on the degrees of freedom, there are two proposals in the literature. Kruschke (2013) proposes an exponentially shaped distribution which balances the prior probability of nearly normal distributions (ν >30) and heavy tailed distributions (ν <30). We re-parameterize this distribution as a Gamma(α,β) with α=1, β= 0.0345. Juárez & Steel (2010) proposes instead $p(\nu) = \text{Gamma}(2, 0.1)$, assigning larger prior probability to normal distributions, as shown in Tab. 1.

We have no reason for preferring a prior over another, but the hierarchical model shows some sensitivity on the choice of $p(\nu)$. We model this uncertainty by representing the coefficients α and β of the Gamma distribution as two random variables (hierarchical prior). In particular we assume $p(\nu) = \text{Gamma}(\alpha, \beta)$, with $\alpha \sim \text{unif}(\underline{\alpha}, \bar{\alpha})$ and $\beta \sim \text{unif}(\underline{\beta}, \bar{\beta})$, setting $\underline{\alpha}$=0.5, $\bar{\alpha}$=5, $\underline{\beta}$=0.05, $\bar{\beta}$=0.15. The mean and standard deviation of the limiting Gamma distribution are given in Table 1; they encompass a wide range of different prior beliefs. In this way the model becomes more stable, showing only minor variations when the limiting ranges of α and β are modified. Being more expressive it also fits better the data as we show in the experimental section.

	α	β	mean	sd	p(ν <30)
Juárez & Steel (2010)	2	0.1	20	14	0.80
Kruschke (2013)	1	0.0345	29	29	0.64
	0.5	0.05	10	14	0.92
	0.5	0.15	3	5	0.99
	5	0.05	100	45	0.02
	5	0.15	33	15	0.47

Table 1 Characteristics of the Gamma distribution for different values of α and β. The last four rows show the characteristic of the extreme distributions assumed by our hierarchical model. The hierarchical model however contains all the priors corresponding to intermediate values of α and β.

The priors for the parameters of the high-level distribution are thus:

$$\begin{aligned}
&\delta_0 \sim \text{unif}(-1, 1),\\
&\sigma_0 \sim \text{unif}(0, \bar{\sigma_0}),\\
&\nu \sim \text{Gamma}(\alpha, \beta),\\
&\alpha \sim \text{unif}(\underline{\alpha}, \bar{\alpha}),\\
&\beta \sim \text{unif}(\underline{\beta}, \bar{\beta}).
\end{aligned}$$

3.1 The region of practical equivalence

Our knowledge about a parameter is fully represented by the posterior distribution. Yet it is handy to summarize the posterior in order to take decisions. In (Corani & Benavoli, 2015) we summarized the posterior distribution by reporting the probability of positiveness and negativeness; however in this way we considered only the sign of the differences, neglecting their magnitude.

A more informative summary of the posterior is obtained introducing a region of practical equivalence (rope), constituted by a range of parameter values that are practically equivalent to the null difference between the two classifiers. We thus summarize the posterior distribution by reporting how much probability lies within the rope, at its left and at its right. The limits of the rope are established by the analyst based on his experience; thus there are no uniquely correct limits for the rope (Kruschke, 2015, Chap. 12). In this paper we consider two classifiers to be practically equivalent if their mean difference of accuracy lies within (-0.01,0.01).

The rope yields a realistic null hypothesis that can be verified. If a large mass of posterior probability lies within the rope, we claim the two classifiers to be practically equivalent. A sound approach to detect equivalent classifiers could be very useful in online model selection (Krueger et al., 2015) where one should quickly discard algorithms that are practically equivalent.

3.2 The inference of the test

We focus on estimating the posterior distribution of the difference of accuracy between the two classifiers on a *future unseen data set*. We compute the probability of left, rope and right being the most probable outcome on the next data set.

Thus we compute the probability by which $p(left) > \max(p(rope), p(right))$ or $p(right) > \max(p(rope), p(left))$ or $p(rope) > \max(p(left), p(right))$. This is similar to the inference carried out by the Bayesian signed-rank test (Benavoli et al., 2014).

To compute such inference, we proceed as follows:

1. initialize the counters $n_{left} = n_{rope} = n_{right} = 0$;
2. for $i = 1, 2, 3, \ldots, N_s$ repeat
   - sample μ_0, σ_0, ν from their posteriors;
   - define the posterior of the mean difference accuracy on the next dataset, i.e., $t(\delta_{next}; \delta_0, \sigma_0, \nu)$;
   - from $t(\delta_{next}; \delta_0, \sigma_0, \nu)$ compute the three probabilities $p(left)$ (integral on $(-\infty, r]$)), $p(rope)$ (integral on $[-r, r]$) and $p(right)$ (integral on $[r, \infty)$; notice that $-r$ and r denote the lower and upper bound of the rope);
   - determine the highest among $p(left), p(rope), p(right)$ and increment the respective counter $n_{left}, n_{rope}, n_{right}$;
3. compute $P(left) = n_{left}/N_s$, $P(rope) = n_{rope}/N_s$ and $P(right) = n_{right}/N_s$;
4. decision: when $P(rope) > 1 - \alpha$ (α is the size of the test) declare the two classifiers to be *practically equivalent*; when $P(left) > 1 - \alpha$ or $P(right) > 1 - \alpha$, declare the two classifiers to be significantly different.

3.3 The shrinkage estimator

The δ_i's of the hierarchical model are independent given the parameters of the higher-level distribution. If such parameters were known, the δ_i's would be conditionally independent and they would be independently estimated. Instead such parameters are unknown, causing the δ_0 and the δ_i's to be jointly estimated. The hierarchical model jointly estimates the δ_i's by applying shrinkage to the $\bar{x}_i$'s, namely it pulls the estimates close to each other. It is known that the shrinkage estimator achieves a lower error than MLE in case of uncorrelated data; see (Murphy, 2012, Sec 6.3.3.2) and the references therein. However there is currently no analysis of shrinkage with correlated data, such as those yielded by cross-validation. We study this problem in the following.

To this end, we assume the cross-validation results on the q data sets to be generated by the hierarchical model:

$$\begin{aligned} &\delta_i \sim p(\delta_i), \\ &\mathbf{x}_i \sim MVN(\mathbf{1}\delta_i, \boldsymbol{\Sigma}), \end{aligned} \tag{5}$$

where for simplicity we assumed the variances σ_i^2 of the individual data sets to be equal to σ^2 and known. Thus all data sets have the same covariance matrix $\boldsymbol{\Sigma}$, which is defined as follows: all variances are σ^2 and all correlations equal ρ. Note that Eqn. (5) coincides with (4). This is a general model that makes no assumptions about the distribution $p(\delta_i)$. We denote the first two moments of $p(\delta_i)$ as $E[\delta_i] = \delta_0$ and $\mathrm{Var}[\delta_i] = \sigma_0^2$.

We study the MAP estimates of the parameters $\delta_1, \ldots, \delta_m, \delta_o, \sigma_o^2$, which asymptotically tend to the Bayesian estimates. A hierarchical model is being fitted to the data. Such model is a simplified version of that presented in Sec. 3. In particular $p(\delta_i)$ is Gaussian for analytical tractability.

$$P(\bar{\boldsymbol{x}}, \boldsymbol{\delta}, \delta_0, \sigma_0^2) = \prod_{i=1}^{q} N(\mathbf{x}_i; \mathbf{1}\delta_i, \boldsymbol{\Sigma}) N(\delta_i; \delta_o, \sigma_o^2) p(\delta_o, \sigma_o^2). \tag{6}$$

This model is misspecified since $p(\delta_i)$ is generally not Gaussian. Nevertheless, it correctly estimates the mean and variance of $p(\delta_i)$, as we show in the following.

Proposition 1 *The derivatives of the logarithm of $P(\bar{\boldsymbol{x}}, \boldsymbol{\delta}, \delta_0, \sigma_0^2)$ are:*

$$\begin{aligned} \frac{d}{d\delta_i} \ln(P(\cdot)) &= \frac{\delta_o - \delta_i}{\sigma_o^2} + \frac{\bar{x}_i - \delta_i}{\sigma_n^2}, \\ \frac{d}{d\delta_o} \ln(P(\cdot)) &= \frac{-q\delta_o + \sum_{i=1}^{q} \delta_i}{\sigma_o^2} + \frac{d}{d\delta_o} \ln(p(\delta_o, \sigma_o^2)), \\ \frac{d}{d\sigma_o} \ln(P(\cdot)) &= \frac{q\delta_o^2 + \sum_{i=1}^{q} \delta_i^2 - 2\delta_o \sum_{i=1}^{q} \delta_i - q\sigma_o^2}{\sigma_o^3} + \frac{d}{d\sigma_o} \ln(p(\delta_o, \sigma_o^2)). \end{aligned}$$

If we further assume that $p(\delta_o, \sigma_o^2) \approx$ constant (flat prior), by equating the derivatives to zero, we derive the following consistent estimators:

$$\sigma_o^2 = \frac{1}{q} \sum_{i=1}^{q} (\hat{\delta}_i - \hat{\delta}_o)^2, \tag{7}$$

$$\hat{\delta}_i = \frac{\hat{\sigma}_o^2 \bar{x}_i + \sigma_n^2 \frac{1}{q} \sum_{i=1}^{q} \bar{x}_i}{\hat{\sigma}_o^2 + \sigma_n^2} = w\bar{x}_i + (1-w)\frac{1}{q} \sum_{i=1}^{q} \bar{x}_i, \tag{8}$$

where $w = \hat{\sigma}_o^2 / (\hat{\sigma}_o^2 + \sigma_n^2)$ and, to keep a simple notation, we have not explicited the expression $\hat{\sigma}_o$ as a function of $\bar{x}_i, \sigma_n^2$. Notice that the estimator $\hat{\delta}_i$ shrinks the estimate towards $\frac{1}{q} \sum_{i=1}^{q} \bar{x}_i$ that

is an estimate of δ_0. Hence, the Bayesian hierarchical model consistently estimates δ_0 and σ_0^2 from data and converges to the shrinkage estimator $\hat{\delta}_i(\mathbf{x}_i) = w\bar{x}_i + (1-w)\delta_0$.

Consider the generative model (5). The likelihood regarding the i-th data set is:

$$p(\mathbf{x}_i|\delta_i, \boldsymbol{\Sigma}) = N(\mathbf{x}_i; \mathbf{1}\delta_i, \boldsymbol{\Sigma}) = \frac{\exp(-\frac{1}{2}(\mathbf{x}_i - \mathbf{1}\delta_i)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \mathbf{1}\delta_i))}{(2\pi)^{n/2}\sqrt{|\boldsymbol{\Sigma}|}}. \tag{9}$$

Let us denote by $\boldsymbol{\delta}$ the vector of the δ_i's. The joint probability of data and parameters is:

$$P(\boldsymbol{\delta}, \mathbf{x}_1, \ldots, \mathbf{x}_q) = \prod_{i=1}^{q} N(\mathbf{x}_i; \mathbf{1}\delta_i, \boldsymbol{\Sigma}) p(\delta_i).$$

Let us focus on the i-th group, denoting by $\hat{\delta}_i(\mathbf{x}_i)$ an estimator of δ_i. The mean squared error (MSE) of the estimator w.r.t. the true joint model $P(\delta_i, \mathbf{x}_i)$ is:

$$\iint \left(\delta_i - \hat{\delta}_i(\mathbf{x}_i)\right)^2 N(\mathbf{x}_i; \mathbf{1}\delta_i, \boldsymbol{\Sigma}) p(\delta_i) d\mathbf{x}_i d\delta_i. \tag{10}$$

Proposition 2 *The MSE of the maximum likelihood estimator is:*

$$\begin{aligned} \mathrm{MSE}_{\mathrm{MLE}} &= \iint (\delta_i - \bar{x}_i)^2 N(\mathbf{x}_i; \mathbf{1}\delta_i, \boldsymbol{\Sigma}) p(\delta_i) d\mathbf{x}_i d\delta_i \\ &= \frac{1}{n^2} \mathbf{1}^T \boldsymbol{\Sigma} \mathbf{1}, \end{aligned}$$

which we denote in the following also as $\sigma_n^2 = \frac{1}{n^2}\mathbf{1}^T\boldsymbol{\Sigma}\mathbf{1}$.

Now consider the shrinkage estimator $\hat{\delta}_i(\mathbf{x}_i) = w\bar{x}_i + (1-w)\delta_0$ with $w \in (0,1)$, which pulls the MLE estimate $\bar{x}_i$ towards the mean δ_0 of the upper-level distribution.

Proposition 3 *The MSE of the shrinkage estimator is:*

$$\begin{aligned} \mathrm{MSE}_{\mathrm{SHR}} &= \iint (\delta_i - w\bar{x}_i - (1-w)\delta_0)^2 N(\mathbf{x}_i; \mathbf{1}\delta_i, \boldsymbol{\Sigma}) p(\delta_i) d\mathbf{x}_i d\delta_i \\ &= w^2\sigma_n^2 + (1-w)^2\sigma_0^2. \end{aligned}$$

As we have seen, the hierarchical model converges to the shrinkage estimator with $w = \sigma_0^2/(\sigma_0^2+\sigma_n^2)$. The shrinkage estimator achieves a smaller mean squared error than the MLE since:

$$\begin{aligned} \mathrm{MSE}_{\mathrm{SHR}} = w^2\sigma_n^2 + (1-w)^2\sigma_0^2 &= \frac{\sigma_0^4 + \sigma_n^2\sigma_0^2}{(\sigma_0^2+\sigma_n^2)^2}\sigma_n^2 \\ &= \frac{\sigma_0^2}{(\sigma_0^2+\sigma_n^2)}\sigma_n^2 < \sigma_n^2 = \mathrm{MSE}_{\mathrm{MLE}}. \end{aligned}$$

3.4 Implementation and code availability

We implemented the hierarchical model in Stan (Carpenter et al., 2017), a language for Bayesian inference. In order to improve the computational efficiency, we exploit a quadratic matrix form to compute simultaneously the likelihood of the q data sets. This provides a speedup of about one order of magnitude compared to the naive implementation in which the likelihoods are computed separately on each data set. Inferring the hierarchical model on the results of 10 runs of 10-folds cross-validation on 50 data sets (a total of 5000 observations) takes about three minutes on a standard laptop. For the sake of completeness we recall that the computation of the much simpler signed-rank test is instead immediate.

The Stan code is available from `https://github.com/BayesianTestsML/tutorial/tree/master/hierarchical`. The same repository provides the R code of all the simulations of Sec. 4.

4 Experiments

4.1 Estimation of the δ_i's under misspecification of p(δ_i)

According to the proofs of Sec. 3, the shrinkage estimator of the δ_i's has lower mean squared error than the maximum likelihood estimator, constituted by the arithmetic mean of the cross-validation results. This result holds even if the $p(\delta_i)$ of the hierarchical model is misspecified: it only requires the hierarchical model to reliably estimate the first two moments of $p(\delta_i)$.

To verify this theoretical result we design the following experiment. We consider these numbers of data sets: $q = \{5, 10, 50\}$. For each value of q we repeat 500 experiments consisting of:

- sampling of the δ_i's $(\delta_1, \delta_2, \ldots, \delta_q)$ from the *bimodal mixture*

$$p(\delta_i) = \pi_1 N(\delta_i|\mu_1, \sigma_1) + \pi_2 N(\delta_i|\mu_2, \sigma_2),$$

  with k=2, μ_1=0.005, μ_2=0.02, σ_1=σ_2=σ=0.001, $\pi_1 = \pi_2 = 0.5$.
- For each δ_i:
  - implement two classifiers whose actual difference of accuracy is δ_i, following the procedure given in Appendix;
  - perform 10 runs of 10-folds cross-validation with the two classifiers;
  - measure the mean of the cross-validation results $\bar{x}_i$ (MLE).
- infer the hierarchical model using the results referring to the q data sets;
- obtain the shrinkage estimates of each δ_i;
- measure MSE_{MLE} and MSE_{SHR} as defined in Sec. 3.3.

q	*Mean Squared Error*	
	MLE	Shrinkage
5	.00036	.00017
10	.00036	.00014
50	.00036	.00012

Table 2 Estimation error of the δ_i's.

As reported in Tab. 2, MSE_{SHR} is at least 50% lower than MSE_{MLE} for every value of q. This confirms our theoretical findings. It also shows that the mean of the cross-validation estimates is a quite noisy estimator of δ_i, even if 10 repetitions of cross-validation are performed. The problem is that all such results are correlated and thus they have limited informative content.

Interestingly, the MSE of the shrinkage estimator decreases with q. Thus the presence of more data sets allows to better estimate the moments of $p(\delta_i)$, improving the shrinkage estimates as well. Instead the error of the MLE does not vary with q since the parameters of each data set are independently estimated.

4.2 Comparison of equivalent classifiers

In this section we adopt a Cauchy distribution as $p(\delta_i)$; this is an idealized situation in which the hierarchical model can recover the actual $p(\delta_i)$. We will relax this assumption in Sec. 4.7.

We simulate the null hypothesis of the signed-rank test by setting the median of the Cauchy to $\delta_0 = 0$. We set the scale factor of the distribution to 1/6 of the rope length; this implies that 80% of the sampled δ_i's lies within the rope, which is the most probable outcome.

We consider the following numbers of data sets: $q = \{10, 20, 30, 40, 50\}$. For each value of q we repeat 500 experiments consisting of:

– sampling the δ_i's $(\delta_1, \delta_2, \ldots, \delta_q)$ from $p(\delta_i)$;
– for each δ_i:
  – implement two classifiers whose actual difference of accuracy is δ_i, following the procedure given in Appendix;
  – perform 10 runs of 10-fold cross-validation with the two classifiers;
– analyze the results through the signed-rank and the hierarchical model.

The signed-rank test (α=0.05) rejects the null hypothesis about 5% of the times for each value of q. It is thus correctly calibrated. Yet, it provides no valuable insights. When it does not reject H_0 (95% of the times), it does *not* allow claiming that the null hypothesis is true. When it rejects the null (5% of the times), it draws a *wrong* conclusion since δ_0=0.

The hierarchical model draws more sensible conclusions. The posterior probability $p(rope)$ increases with q (Fig. 1): the presence of more data sets provides more evidence that they are equivalent. For q=50 (the typical size of a machine learning study), the average p(rope) reported in simulations is larger than 90%. Fig. 1 shows also the *equivalence recognition*, which is the proportion of simulations in which p(rope) exceeds 95%. Equivalence recognition increases with q, reaching about 0.7 for q=50.

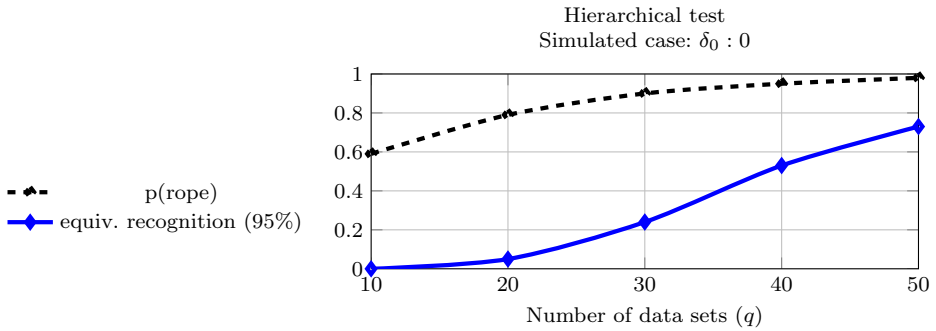

Fig. 1 Behavior of the hierarchical classifier when dealing with two actually equivalent classifiers.

Moreover in our simulations the hierarchical model never estimated p(left)>95% or p(right)>95%, so it made no Type I errors. In fact nsht commits a rate α of Type I errors under the null hypothesis, while Bayesian estimation with rope typically makes less Type I errors (Kruschke, 2013).

Running the signed-rank twice? We *cannot* detect practically equivalent classifiers by running twice the signed-rank test, e.g., once with null hypothesis δ_0 =0.01 and once with the null hypothesis δ_0 =-0.01. Even if the signed-rank test does not reject the null in both cases, we still cannot affirm that the two classifiers are equivalent, since non-rejection of the null does not allow claiming that the null is true.

4.3 Comparison of practically equivalent classifiers

We now simulate two classifiers whose actual difference of accuracy is practically irrelevant but different from zero. We consider two classifiers whose average difference is δ_0=0.005, thus within the rope.

We consider $q = \{10, 20, 30, 40, 50\}$. For each value of q we repeat 500 experiments as follows:

– set $p(\delta_i)$ as a Cauchy distribution with δ_0=0.005 and the same scale factor as in previous experiments (the rope remains the most probable outcome for the sampled δ_i's);
– sample the δ_i's $(\delta_1, \delta_2, \ldots, \delta_q)$ from $p(\delta_i)$;

– implement for each δ_i two classifiers whose actual difference of accuracy is δ_i and perform 10 runs of 10-fold cross-validation;
– analyze the cross-validation results through the signed-rank and the hierarchical model.

The signed-ranked test is more likely to reject the null hypothesis as the number of data sets increases (Fig. 2). When 50 data sets are available, the signed-rank rejects the null in about 25% of the simulations, despite the trivial difference between the two classifiers. Indeed one can reject the null of the signed-rank test when comparing two almost equivalent classifiers, by comparing them on enough data sets. As reported in the ASA statement on p-value (Wasserstein & Lazar, 2016), even a tiny effect can produce a small p-value if the sample size is large enough.

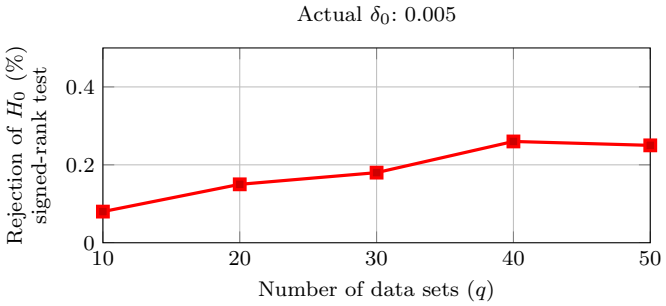

Fig. 2 When faced with two practically equivalent classifiers, the signed-rank rejects more often H_0 as q increases.

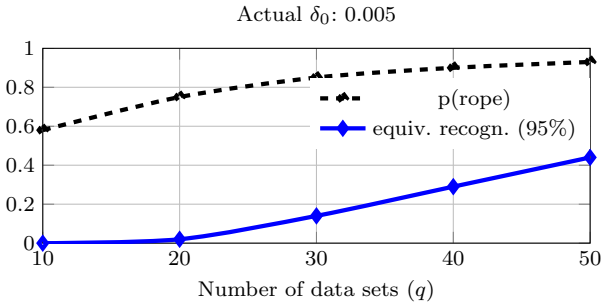

Fig. 3 When faced with two practically equivalent classifiers, the hierarchical test responds to an increase of q by increasing the probability of rope.

The behavior of the hierarchical test is way more sensible. The hierarchical test increases the posterior probability of rope (Fig. 3) when the number of data sets in which the classifiers show similar performance increases. It is slightly less effective in recognizing equivalence than in the previous experiment since δ_0 is now closer to the limit of the rope. When q=50, it declares equivalence detection with 95% confidence in about 40% of the simulated cases.

The hierarchical test thus effectively detects classifiers that are practically equivalent; this is instead impossible for the signed-rank test.

The hierarchical model is more conservative as it rejects the null hypothesis less easily than the signed rank test. The price to be paid is that it might be less powerful at claiming significance when comparing two classifiers whose accuracies are truly different. We investigate this setting in the next section.

4.4 Simulation of practically different classifiers

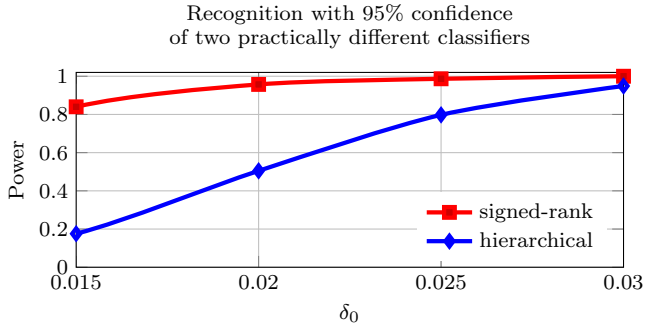

Fig. 4 The signed-rank test is generally more powerful than the hierarchical test in the detection of significant differences between classifiers. Yet the two tests have similar power when δ_0 is far enough from the rope.

We now simulate two classifiers which are significantly different. We consider different values of δ_0: $\{0.015, 0.02, 0.025, 0.03\}$. We set the scale factor of the Cauchy to σ_0=0.01 and the number of data sets to q=50.

We repeat 500 experiments for each value of δ_0, as in the previous sections. We then check the *power* of the two tests for each value of δ_0. The power of the signed-rank is the proportion of simulations in which it rejects the null hypothesis (α=0.05). The power of the hierarchical test is the proportion of simulations in which it estimates p(right)>0.95.

As expected, the signed-rank test is indeed more powerful in this setting than the hierarchical model, especially when δ_0 lies just slightly outside the rope (Fig. 4). The two tests have however similar power when δ_0 is larger than 0.02.

4.5 Discussion

The main experimental findings so far are as follows. First, the shrinkage estimator of the δ_i's yields a lower mean squared error than the MLE estimator, even under misspecification of $p(\delta_i)$.

Second, the hierarchical model effectively detects equivalent classifiers, unlike the nhst test.

However, it is also less powerful than the signed-rank when comparing two significantly different classifiers. The difference in power is however not necessarily large, as shown in the previous simulation.

In the next section we discuss how the probabilities returned by the hierarchical model can be interpreted in a more meaningful way than simply checking if they are larger than 1-α.

4.6 Interpreting posterior odds

The ratio of posterior probabilities (*posterior odds*) shows the extent to which the data support one hypothesis over the other. For instance we can compare the support for left and right by computing the posterior odds $o(\text{left}, \text{right}) = \frac{p(left)}{p(right)}$. When $o(\text{left}, \text{right}) > 1$ there is evidence in favor of left; when $o(\text{left}, \text{right}) < 1$ there is evidence in favor of right. Rules of thumb for interpreting the amount of evidence corresponding to posterior odds are discussed by Raftery (1995) and summarized in Tab. 3:

Posterior Odds	Evidence
1–3	weak
3–20	positive
>20	strong

Table 3 Grades of evidence corresponding to posterior odds.

Thus even if none of the three probabilities exceeds the 95% threshold, we can still draw meaningful conclusions by interpreting the posterior odds. We will adopt this approach in the following simulations.

The p-values cannot be interpreted in a similar fashion, since they are affected both by sample size and effect size. In particular (Wasserstein & Lazar, 2016) show that smaller p-values do not necessarily imply the presence of larger effects and larger p-values do not imply a lack of effect. A tiny effect can produce a small p-value if the sample size is large enough, and large effects may produce unimpressive p-values if the sample size is small.

4.7 Experiments with Friedman's functions

The results presented in the previous sections refer to conditions in which the actual $p(\delta_i)$ (misspecified or not) is known. In this section we perform experiments in which the δ_i's are not sampled from an analytical distribution; rather, they are due to different settings of sample size, noise etc. This is a challenging setting for the hierarchical model, whose $p(\delta_i)$ is unavoidably misspecified.

We generate data sets via the three functions ($F\#1$, $F\#2$ and $F\#3$) proposed by Friedman (1991).

Function $F\#1$ contains ten features $x_1, \ldots, x_{10}$, each uniformly distributed over $[0, 1]$. Only five features are used to generate the response y:

$$F\#1: \; y = 10sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \epsilon_1,$$

where $\epsilon_1 \sim N(0, 1)$. We turn this regression problem into a classification one by discretizing y in two bins, delimited by the median of y (which we estimate on a sample of 10,000 instances).

Functions $F\#2$ and $F\#3$ have four features $x_1, \ldots, x_4$ uniformly distributed over the ranges:

$$\begin{aligned} 0 \leq x_1 \leq 100, \\ 40\pi \leq x_2 \leq 560\pi, \\ 0 \leq x_3 \leq 1, \\ 1 \leq x_4 \leq 11. \end{aligned}$$

The functions are:

$$\begin{aligned} F\#2: \; y &= (x_1^2 + (x_2 x_3 - (1/x_2 x_4))^2)^{0.5} + \epsilon_2, \\ F\#3: \; y &= \arctan\left(\frac{x_2 x_3 - (1/x_2 x_4)}{x_1}\right) + \epsilon_3, \end{aligned}$$

where $\epsilon_2 \sim N(0, \sigma_{\epsilon_2}^2)$ and $\epsilon_3 \sim N(0, \sigma_{\epsilon_3}^2)$. The original paper sets σ_{ϵ_2}=125 and σ_{ϵ_3}=0.1. Also in this case we turn the problem into a classification one by discretizing the response variable in two bins, around the median of y.

We consider 18 settings for each function, obtained by varying the sample size (n), the standard deviation of the noise (σ_ϵ) and either considering only the original features or adding further twenty

Function type	σ_ϵ	n	random feats	Tot settings
F#1	{0.5,1,2}	{30,100,1000}	{0,20}	$3 \cdot 3 \cdot 2 = 18$
F#2	{62.5,125,250}	{30,100,1000}	{0,20}	$3 \cdot 3 \cdot 2 = 18$
F#3	{0.05,0.1,0.2}	{30,100,1000}	{0,20}	$3 \cdot 3 \cdot 2 = 18$

Table 4 Settings of the Friedman functions.

Gaussian features, all independent of the class (*random features*). We have overall 54 settings: 18 settings for each function. They are summarized in Table 4.

As a pair of classifiers we consider linear discriminant analysis (*lda*) and classification trees (*cart*), as implemented in the `caret` package for R, without any hyper-parameter tuning. As first step we need to measure the actual δ_i between two given classifiers in each setting, which then allows us to know the population of the δ_i's.

Our second step will be to check the conclusions of the signed-rank test and of the hierarchical model when they are provided with cross-validation results referring to a subset of settings.

Measuring δ_i

We start by measuring the actual difference of accuracy between *lda* and *cart* in each setting. In the i-th setting we estimate δ_i as follows:

- for j=1:500
  - sample training data according to the specifics of the i-th setting: <function type, n, σ_ϵ, number of random features;
  - fit *lda* and *cart* on the generated training data;
  - sample a large test set (5000 instances) and measure the difference of accuracy d_{ij} between *cart* and *lda*;
- set $\delta_i \simeq 1/500 \sum_j d_{ij}$.

Our procedure yields accurate estimates since each repetition is performed on independently generated data characterized by large test sets.

For instance if two classifiers have mean difference of accuracy $\bar{x}$=0.09, with standard deviation s=0.06; the 95% confidence interval of their difference is tight:

$$\bar{x} \pm 1.96 \cdot \frac{s}{\sqrt{n}} = 0.09 \pm 1.96 \cdot \frac{0.06}{\sqrt{500}} = [0.085 - 0.095].$$

If instead we had performed 500 runs of 10-folds cross-validation obtaining the same value of $\bar{x}$ and s, the confidence interval of our estimates would be about 3.5 times larger, as the standard error would be $s\sqrt{\frac{1}{n} + \frac{\rho}{1-\rho}}$ instead of $\frac{s}{\sqrt{n}}$, as shown in Eqn.(1).

Ground-truth

We compute the δ_i of each setting using the above procedure. The ground-truth is that lda is significantly more accurate than cart. More in detail, 65% of the δ_i's belong to the region to the right of the rope (lda being significantly more accurate than cart). Thus right is the most probable outcome of the next δ_i. Moreover, the mean of the δ_i's is δ_0=0.02 (in favor of lda).

Assessing the conclusions of the tests

We run 200 times the following procedure:

- random selection of 12 out of 18 settings for each Friedman function, thus selecting 36 settings;
- in each setting:
  - generate a data set according to the specific of the setting;
  - run 10 runs of 10-folds cross-validation of lda and cart using paired folds;
- analyze the cross-validation results on the q=36 data sets using the signed rank and the hierarchical test.

We start checking the *power* of the tests, defined as the proportion of simulations in which the null hypothesis is rejected (signed-rank) or the posterior probability p(right) exceeds 95% (hierarchical test).

The two tests have roughly the same power: 28% for the signed-rank and 27.5% for the hierarchical test. In the remaining simulations the signed-rank does not reject H_0; in those cases it conveys no information since the p-values cannot be interpreted.

We can instead interpret the posterior odds yielded by the hierarchical model, obtaining the following results:

- in 11% of the simulations both $o(right, rope)$ and $o(right, left)$ are larger than 20, providing strong evidence in favor of lda even though p(right) does not exceed 95%;
- in a further 33% of the simulations both $o(right, rope)$ and $o(right, left)$ are larger than 3, providing at least positive evidence in favor of lda.

We have moreover to point out a 2% of simulations in which the posterior odds provide erroneously positive evidence for rope over both right and left. In no case there is positive evidence for left over either rope or right.

Thus the interpretation of posterior odds allows drawing meaningful conclusions even when the 95% threshold is not exceeded. The probabilities are sensibly estimated, even if $p(\delta_i)$ is unavoidably misspecified.

As a further check we compare MSE_{MLE} and MSE_{Shr}. Also in this case MSE_{MLE} is much lower than MSE_{Shr} (Fig 5), with an average reduction of about 60%. This further confirms the properties of the shrinkage estimator.

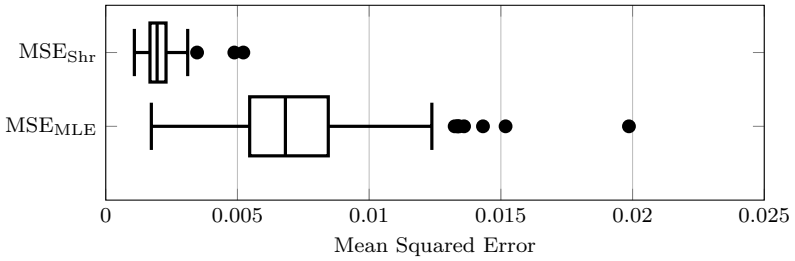

Fig. 5 Boxplots of MSE_{MLE} and MSE_{Shr} over 200 repetitions of our experiment with the Friedman functions.

4.8 Sensitivity analysis on real-world data sets

We now consider real data sets. In this case we *cannot* know the actual δ_i's: we could repeat a few hundred times cross-validation but the resulting estimates would have large uncertainty as already discussed.

We exploit this setting to perform sensitivity analysis and to further compare the conclusions drawn by the hierarchical model and of the signed-rank test.

We consider 54 data sets taken from the webpage[1] of WEKA data sets. We consider four classifiers: naive Bayes (nbc), hidden naive Bayes (hnb), decision tree (j48), grafted decision tree (j48gr). Witten et al. (2011) provides a summary description of all such classifiers with pointers to the relevant papers. We perform 10 runs of 10-folds cross-validation for each classifier on each data set. We run all experiments using the WEKA[2] software.

A fundamental step of Bayesian analysis is to check how the posterior conclusions depend on the chosen prior and how the model fits the data. The hierarchical model shows some sensitivity on the choice of $p(\delta_i)$, being instead robust to the other assumptions (see later for further discussion). The Student distribution is more flexible than the Gaussian and we have found that it consistently provides better fit to the data. Yet, the model conclusions are sometimes sensitive on the prior on the degrees of freedom $p(\nu)$ of the Student.

In Table 5 we compare the posterior inferences of the model, using the prior $p(\nu) = Gamma(2, 0.1)$ (proposed by Juárez & Steel (2010)) or using the more flexible model described in Sec. 3, where the parameters of the Gamma are described as random variables with their own prior distributions. Such two variants are referred to as *Gamma(2,0.1)* and *hierarchical* in Table 5.

	Hierarchical			**Gamma(2,0.1)**		
pair	left	rope	right	left	rope	right
nbc-hnb	1.00	0.00	0.00	1.00	0.00	0.00
nbc-j48	0.80	0.02	0.18	0.80	0.01	0.20
nbc-j48gr	0.84	0.02	0.14	0.84	0.01	0.15
hnb-j48	0.03	0.10	0.87	0.03	0.02	0.95
hnb-j48gr	0.03	0.07	0.90	0.03	0.02	0.95
j48-j48gr	0.00	1.00	0.00	0.00	1.00	0.00

Table 5 Posterior probabilities computed by two variants of the hierarchical model.

In some cases the estimates of the two models differ by some points (Tab. 5). This means that the actual high-level distribution from which the δ_i's are sampled is not a Student (or a Gaussian), otherwise the estimate of the two models would converge.

Which model better fits the data? We respond to this question by adopting a visual approach. We start considering that the shrinkage estimates of the δ_i's are identical between the two models. We then compute the density plot of the shrinkage estimates (our best estimate of the δ_i's). We take such density as the ground truth (this is actually our best approximation to the ground truth) and we plot it in thick black (Fig. 6). Then we sample 8000 δ_i's from both variants of the model, obtaining two further densities. We then plot the three densities for each pair of classifiers (Fig. 6). We produce all the density plots using the default kernel density estimation provided in R. In general the hierarchical model, being more flexible, fits better the data than the model equipped with a simple Gamma prior.

4.8.1 Sensitivity on the prior on σ_0 and σ_i

The model conclusions are moreover robust with respect to the specification of the priors $p(\sigma_i)$ and $p(\sigma_0)$. Recall that σ_i is the standard deviation on the i-th data set while σ_0 is the standard deviation of the high-level distribution.

[1] `http://www.cs.waikato.ac.nz/ml/weka/datasets.html`

[2] `http://www.cs.waikato.ac.nz/ml/weka/`

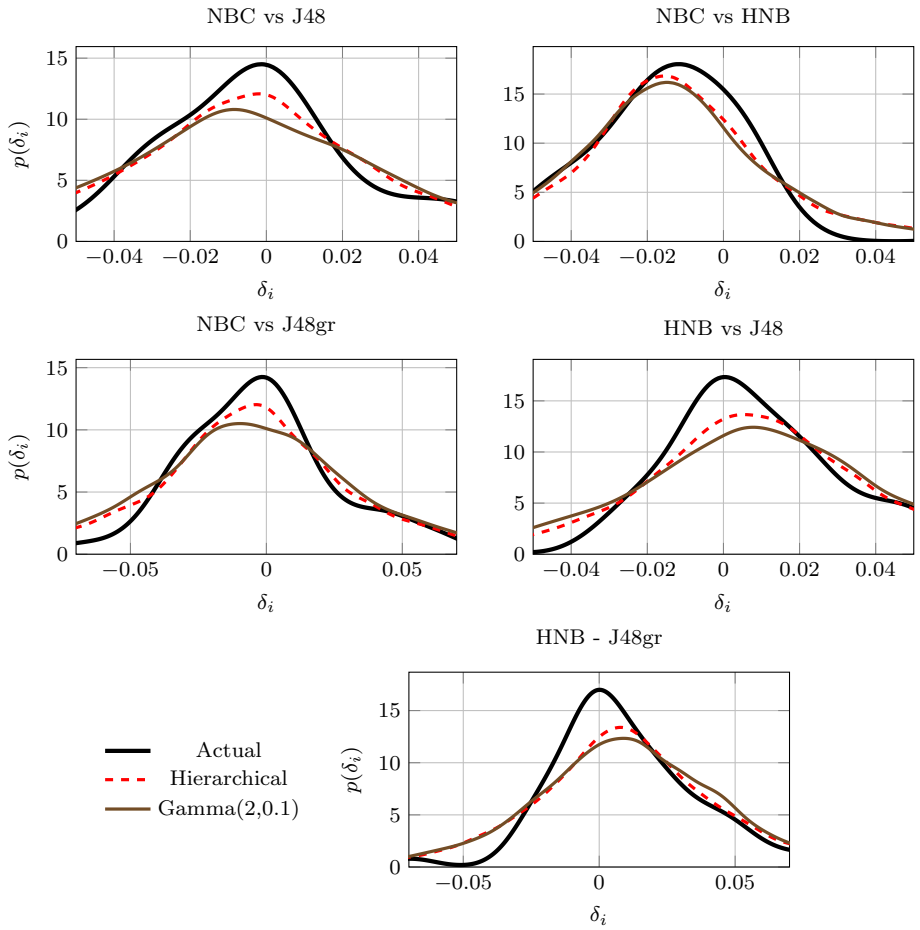

Fig. 6 Comparison of the densities estimated by $p(\delta_i)$ of two variants of the hierarchical model in selected cases.

Our model assumes $\sigma_i \sim \text{unif}(0, \bar{\sigma})$ where $\bar{\sigma} = 1000\bar{s}$ where $\bar{s}$ is the average of the sample standard deviations of the different data sets. The posterior distribution of σ_i is however substantially unchanged if we adopt instead $\bar{\sigma} = 100\bar{s}$.

The same consideration applies to σ_0, whose prior is $p(\sigma_0) = unif(0, \bar{s_0})$. We obtain the same posterior distribution for σ_0 using as upper bound $\bar{s_0} = 1000 s_{\bar{x}}$ or $\bar{s_0} = 100 s_{\bar{x}}$, where $s_{\bar{x}}$ is the standard deviation of the $\bar{x}_i$'s.

4.9 Comparing the signed-rank and the hierarchical test

We compare the conclusions of the hierarchical model and of the signed-rank test on the same cases of the previous section. The results are given in Tab. 6.

Both the signed-rank and the hierarchical test claim with 95% confidence hnb to be significantly more accurate than nbc.

In the following comparisons apart from the last one, the two tests do not draw any conclusion with 95% confidence. The signed-rank does not reject the null hypothesis, while the hierarchical test does not achieve probability larger than 95%.

	Hierarchical			Signed-rank
pair	left	rope	right	p-value
nbc-hnb	1.00	0.00	0.00	0.00
nbc-j48	0.80	0.02	0.18	0.46
nbc-j48gr	0.84	0.02	0.14	0.39
hnb-j48	0.03	0.10	0.87	0.07
hnb-j48gr	0.03	0.07	0.90	0.08
j48-j48gr	0.00	1.00	0.00	0.00

Table 6 Posterior probabilities of the hierarchical model and p-values of the signed-rank.

When the signed-rank test does not reject the null hypothesis, it draws a non-informative conclusion. We can instead always interpret the posterior odds yielded by the hierarchical model. When comparing nbc and j48, there is a positive evidence for right (j48 being more accurate than nbc) over left and strong evidence for right over rope. We thus conclude that there is positive evidence of j48 being practically more accurate than nbc. Similarly, we conclude that there is positive evidence of j48gr being practically more accurate than nbc

When comparing hnb and j48, there is strong evidence for right (hnb being more accurate than j48) over both left and rope. We conclude that there is strong evidence of j48 being practically more accurate than hnb. We draw the same conclusion when comparing hnb and j48gr.

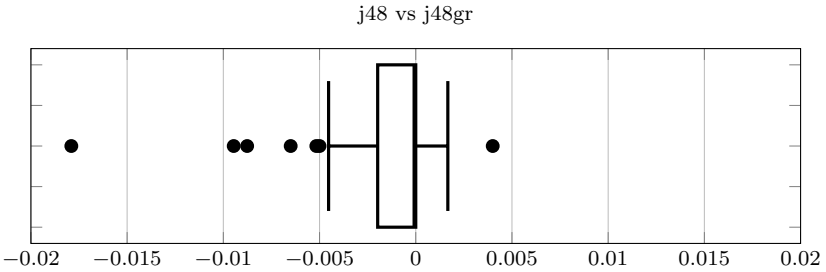

Fig. 7 Boxplots of the differences of accuracy $\bar{x}_i$'s between j48 and j48gr on 54 data sets.

The two test draw opposite conclusions when comparing j48 and j48gr. The signed-rank declares j48gr to be significantly more accurate than j48 (p-value 0.00) while the hierarchical model declares them to be practically equivalent, with p(rope)=1. The reason why the two tests achieved opposite conclusions is that the differences have a consistent sign but are small-sized. Most data sets yield a positive difference in favor of j48gr; this leads the signed rank test to claim significance. Yet the differences lies mostly within the rope (Fig. 7). The hierarchical model shrinks them further towards the overall mean and eventually claims the two classifiers to be practically equivalent. The posterior probabilities remain unchanged even adopting the half-sized rope (-0.005, 0.005). It thus seems fair to conclude that, even if most signs are in favor of j48gr, the accuracies of j48 and j48gr are practically equivalent.

5 Conclusions

The proposed approach is a realistic model of the data generated by cross-validation across multiple data sets. Through the rope it also defines a sensible null hypothesis which can be verified, allowing the test to detect classifiers that are practically equivalent. The interpretation of the posterior odds allows drawing meaningful conclusions even when the posterior probabilities do not exceed

95%. Thanks to shrinkage, the hierarchical model estimates the δ_i's more accurately than the usual approach of averaging (independently on each data set) the cross-validation differences. An interesting research direction is thus the adoption of a non-parametric approach for the high-level distribution $p(\delta_i)$. This is a non-trivial task which we leave for future research.

Acknowledgments

The research in this paper has been partially supported by the Swiss NSF grants ns. IZKSZ2_162188 and n. 200021_146606.

6 Appendix

6.1 Implementing two classifiers with known difference of accuracy

On the i-th data set we need to simulate two classifier whose actual difference of accuracy is δ_i. We start by sampling the instances from a naive Bayes model with two features. Let us denote by C the class variables with states $\{c_0, c_1\}$ and by F and G the two features with states $\{f_0, f_1\}$ and $\{g_0, g_1\}$. The naive Bayes model is thus $G \leftarrow C \rightarrow F$. The parameters of the conditional probability tables are: $P(c_0)$=0.5; $P(f_0|c_0) = \theta_f$; $P(f_0|c_1) = 1 - \theta_f$; $P(g_0|c_0) = \theta_g$; $P(g_0|c_1) = 1 - \theta_g$ with $\theta_f > 0.5$. The remaining elements of the conditional probability tables are the complement to 1 of the above elements. We set θ_f=0.9 and $\theta_g = \theta_f + \delta_i$. We sample the data set from this naive Bayes model.

During cross-validation we train and test the two competing classifiers $C \rightarrow F$ and $C \rightarrow G$. Their expected accuracies are θ_f and θ_g respectively, and thus their expected difference of accuracy is $\delta_i = \theta_f - \theta_g$. To explain this statement, let us first consider classifier $C \rightarrow F$. Assume that the marginal probabilities of the class have been correctly estimated. The classification thus depends only on the conditional probability of the feature given the class. If F=f_0 the most probable class is c_0 as long as $\hat{P}(c_0|f_0) > 0.5$, where $\hat{P}$ denotes the conditional probability estimated from data. The accuracy of this prediction is θ_f. It F=f_1, the most probable class is c_1 as long as $\hat{P}(c_1|f_1) = \theta_f > 0.5$. Also the accuracy of this prediction is θ_f. If the bias of conditional probability ($\hat{P}(c_0|f_0)$ >0.5 and $\hat{P}(c_1|f_1)$ >0.5) is correctly estimated the accuracy of classifier $C \rightarrow F$ on a large test set is θ_f. Analogously, the accuracy of classifier $C \rightarrow G$ in the same conditions is θ_g, so that their difference is δ_i. Since the sampled data set have finite size the mean difference of accuracy $\bar{x}_i$ measured by cross-validation will fluctuate with some variance around δ_i.

6.2 Proofs

Proof of Proposition 1 Consider the hierarchical model:

$$\begin{aligned} & P(\bar{\boldsymbol{x}}, \boldsymbol{\delta}, \delta_0, \sigma_0^2) \\ & = \prod_{i=1}^{q} N(\mathbf{x}_i; \mathbf{1}\delta_i, \boldsymbol{\Sigma}) N(\delta_i; \delta_o, \sigma_o^2) p(\delta_o, \sigma_o^2), \end{aligned} \tag{11}$$

We aim at computing the derivative of the $\log(P(\bar{\boldsymbol{x}}, \boldsymbol{\delta}, \delta_0, \sigma_0^2))$ w.r.t. the parameter $\delta_i, \delta_0, \sigma_o^2$. Consider the quadratic term from the first and second Gaussian:

$$\frac{1}{2}(\mathbf{x}_i - \mathbf{1}\delta_i)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \mathbf{1}\delta_i) + \frac{1}{2\sigma_o^2}(\delta_i - \delta_o)^2;$$

its derivatives w.r.t. δ_i is $\mathbf{1}^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \mathbf{1}\delta_i) + \frac{1}{\sigma_o^2}(\delta_i - \delta_o)$. Exploiting the fact that

$$\begin{aligned}&\mathbf{1}^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \mathbf{1}\delta_i) = \mathbf{1}^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \mathbf{1}\bar{x}_i + \mathbf{1}\bar{x}_i - \mathbf{1}\delta_i)\\ &= \mathbf{1}^T\boldsymbol{\Sigma}^{-1}(\mathbf{1}\bar{x}_i - \mathbf{1}\delta_i),\end{aligned}$$

it follows that

$$\frac{d}{\delta_i}\ln(P(\cdot)) \propto \frac{1}{\sigma_n^2}(\bar{x}_i - \delta_i) + \frac{1}{2\sigma_o^2}(\delta_i - \delta_o)^2,$$

where $\sigma_n^2 = \frac{1}{\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}} = \frac{1}{n^2}\mathbf{1}^T\boldsymbol{\Sigma}\mathbf{1}$. The latter equality can be derived by Corani & Benavoli (2015)[Appendix], i.e.,

$$\frac{1}{\mathbf{1}^T\boldsymbol{\Sigma}^{-1}\mathbf{1}} = \frac{n}{1+(n-1)\rho} = \frac{1}{n^2}\mathbf{1}^T\boldsymbol{\Sigma}\mathbf{1}.$$

The other derivatives can be computed easily.

Proof of Proposition 2 Let us consider the likelihood:

$$\begin{aligned}&p(\mathbf{x}_i|\delta_i, \boldsymbol{\Sigma}) = N(\mathbf{x}_i; \mathbf{1}\delta_i, \boldsymbol{\Sigma})\\ &= \frac{\exp(-\frac{1}{2}(\mathbf{x}_i - \mathbf{1}\delta_i)^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \mathbf{1}\delta_i))}{(2\pi)^{n/2}\sqrt{|\boldsymbol{\Sigma}|}}.\end{aligned} \tag{12}$$

Let us define $\bar{x}_i = \sum_{j=1}^n \mathbf{x}_{ij}/n$. The MSE of the maximum likelihood estimator is:

$$\mathrm{MSE}_{\mathrm{MLE}} = \iint (\delta_i - \bar{x}_i)^2 N(\mathbf{x}_i; \mathbf{1}\delta_i, \boldsymbol{\Sigma}) p(\delta_i) d\mathbf{x}_i d\delta_i.$$

Consider that $(\delta_i - \bar{x}_i)^2 = \left(\delta_i - \frac{1}{n}\mathbf{1}^T\mathbf{x}_i\right)^2$ where $\frac{1}{n}\mathbf{1}^T$ is a linear transformation of the variable $\mathbf{x}_i$. From the properties of the Normal distribution, it follows that

$$\int \left(\delta_i - \frac{1}{n}\mathbf{1}^T\mathbf{x}_i\right)^2 N(\mathbf{x}_i; \mathbf{1}\delta_i, \boldsymbol{\Sigma}) d\mathbf{x}_i = \frac{1}{n^2}\mathbf{1}^T\boldsymbol{\Sigma}\mathbf{1}$$

and since

$$\int (\frac{1}{n^2}\mathbf{1}^T\boldsymbol{\Sigma}\mathbf{1}) p(\delta_i) d\mathbf{x}_i d\delta_i = \frac{1}{n^2}\mathbf{1}^T\boldsymbol{\Sigma}\mathbf{1},$$

we derive the first result.

Proof of Proposition 3 The MSE of the shrunken estimator can be obtained in a similar way. First observe that

$$\begin{aligned}&(\delta_i - w\bar{x}_i - (1-w)\delta_0)^2\\ &= w^2(\delta_i - \bar{x}_i)^2 + (1-w)^2(\delta_i - \delta_0)^2\\ &+ 2w(1-w)(\delta_i - \bar{x}_i)(\delta_i - \delta_0)\end{aligned}$$

and its expected value w.r.t. $N(\mathbf{x}_i; \delta_i, \sigma_n^2)p(\delta_i)$ is:

$$\begin{aligned}&\int \left[w^2\sigma_n^2 + (1-w)^2(\delta_i - \delta_0)^2\right] p(\delta_i) d\delta_i\\ &= w^2\sigma_n^2 + (1-w)^2\sigma_0^2,\end{aligned} \tag{13}$$

where we have denoted $\sigma_n^2 = \frac{1}{n^2}\mathbf{1}^T\boldsymbol{\Sigma}\mathbf{1}$.

References

Benavoli, Alessio, Corani, Giorgio, Mangili, Francesca, Zaffalon, Marco, and Ruggeri, Fabrizio. A Bayesian Wilcoxon signed-rank test based on the Dirichlet process. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1026–1034, 2014.

Benavoli, Alessio, Corani, Giorgio, Demsar, Janez, and Zaffalon, Marco. Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis. *Journal of Machine Learning Research*, under review.

Carpenter, Bob, Gelman, Andrew, Hoffman, Matthew, Lee, Daniel, Goodrich, Ben, Betancourt, Michael, Brubaker, Marcus, Guo, Jiqiang, Li, Peter, and Riddell, Allen. Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1):1–32, 2017.

Corani, Giorgio and Benavoli, Alessio. A Bayesian approach for comparing cross-validated algorithms on multiple data sets. *Machine Learning*, 100(2):285–304, 2015.

Demšar, Janez. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

Friedman, Jerome H. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67, 1991.

Gelman, Andrew. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian analysis*, 1(3):515–534, 2006.

Hand, David J et al. Classifier technology and the illusion of progress. *Statistical science*, 21(1): 1–14, 2006.

Juárez, Miguel A and Steel, Mark FJ. Model-based clustering of non-Gaussian panel data based on skew-t distributions. *Journal of Business & Economic Statistics*, 28(1):52–66, 2010.

Krueger, Tammo, Panknin, Danny, and Braun, Mikio. Fast cross-validation via sequential testing. *Journal of Machine Learning Research*, 16:1103–1155, 2015.

Kruschke, John. *Doing Bayesian data analysis: A tutorial with R, Jags and Stan*. Academic Press, 2015.

Kruschke, John K. Bayesian estimation supersedes the t–test. *Journal of Experimental Psychology: General*, 142(2):573, 2013.

Lacoste, Alexandre, Laviolette, François, and Marchand, Mario. Bayesian comparison of machine learning algorithms on single and multiple datasets. In *Proc.of the Fifteenth International Conference on Artificial Intelligence and Statistics (AISTATS-12)*, pp. 665–675, 2012.

Murphy, Kevin P. *Machine learning: a probabilistic perspective*. MIT press, 2012.

Nadeau, Claude and Bengio, Yoshua. Inference for the generalization error. *Machine Learning*, 52 (3):239–281, 2003.

Raftery, Adrian E. Bayesian model selection in social research. *Sociological methodology*, 25:111–164, 1995.

Wasserstein, Ronald L and Lazar, Nicole A. The ASA's statement on p-values: context, process, and purpose. *The American Statistician*, 2016.

Witten, Ian H, Frank, Eibe, and Hall, Mark. *Data Mining: Practical machine learning tools and techniques (third edition)*. Morgan Kaufmann, 2011.