Bayesian parametric and hierarchical approaches for comparing classifiers

Giorgio Corani
giorgio@idsia.ch

Contents

- Comparing cross-validated classifiers on a single data set
  - Frequentist correlated (a.k.a. corrected) t-test
  - Bayesian correlated t-test
- Comparing cross-validated classifiers on multiple data sets
  - Signed-rank test
  - Bayesian hierarchical t-test

Aim of hypothesis testing when comparing classifiers

- Detecting equivalent classifiers
- Declaring two classifiers as significantly different when the difference has a *practical* impact.
- We will argue that both objectives are better met by adopting a Bayesian estimation approach than by NHST.

Two cross-validated classifiers on the same data set

- You assess accuracy by k-folds x-validation with paired folds.
- You compute the difference d on each fold.

	Cross-validation fold				*Mean*	*Var*
	1	2	...	10	$\bar{d}$	S_d^2
Naive Bayes	.70	.78	...	.73		
Decision tree	.68	.77	...	.71		
Difference (d)	.02	.01	...	.02	**1.5**	**0.3**

- The d_i's are samples from a population with mean δ.
- You can run p times cross-validation, so you have $m = p \cdot k$ results (in most cases m=100).

Frequentist inference (NHST) about δ

- The two-sided test is:

$$H_0: \quad \delta = 0$$
$$H_1: \quad \delta \neq 0$$

- Two perfectly equivalent classifiers do not exist: at least they differ by a tiny difference.
- The null hypothesis is surely wrong!
- This is a known drawback of point null hypothesis.

Correlated t-test (Nadeau & Bengio, 2003)

- It adjusts the t-test accounting for correlation ρ, due to the overlapping training sets built during cross-validation.
- It is impossible to estimate ρ from data.
- The test heuristically assumes $\rho = \frac{n_{te}}{n}$ where n_{te} and n are the size of the test set and of the whole data set.
- The correlated t-test is better calibrated than the standard t-test when analyzing the cross-validation results.

Correlated t-test (Nadeau & Bengio, 2003)

The statistic is:

$$t = \frac{\overline{d}}{\sqrt{S_d^2(\frac{1}{m} + \frac{\rho}{1-\rho})}}$$

- It follows a Student distribution with $m - 1$ degrees of freedom.
- The correlation correction is applied to the standard error (denominator of the statistics).

Declaring significance

- The test claims significance when the absolute value of the statistic exceeds the (1-$\alpha/2$) quantile of the Student distribution.
- The p-value is the probability of the statistic assuming under H_0 a more extreme value than the observed t.
- Equivalently, the test rejects the null hypothesis when the p-value is lower than α.
- In this case the test claims the two classifiers to be significantly different.

Non-rejection of the null

- If the p-value is larger than α, the test *does not reject* H_0.
- In this case the test draws a non-committal conclusion: H_0 might be true or no; we do not have enough evidence for rejecting it.
- This outcome provides no evidence of H_0 being true.

Bayesian approaches for comparing classifiers

- Different works in recent years:
  - Lacoste et al., AISTATS 2012
  - Brodersen et al., JMLR 2012
  - Also the literature of neuroscience has moved towards Bayesian hypothesis testing: Melinscak et al., J. Neuroscience, 2016.
- None of those approaches models the correlation which exist between cross-validation results.

The Bayesian correlated t-test (Corani and Benavoli, 2015)

- It yields the posterior distribution of δ modeling the correlation of the cross-validation results d_i.

- Idea:
  - Consider a multivariate normal (MVN) with $m = p \cdot k$ components (p runs of k-fold cv).
  - Each component has mean δ and variance σ^2
  - All components are cross-correlated with correlation $\rho = \frac{n_{te}}{n}$.
  - The d_i's are jointly sampled from such MVN.

Generative model - I

$$p(\delta, \tau) \sim \text{NormalGamma}(\mu_0, \text{k}_0, \text{a}, \text{b})$$

$$\mathbf{d}|\delta, \tau \sim \text{MVN}(\mathbf{1}\delta, \mathbf{\Sigma_{m\times m}})$$

$$\mathbf{\Sigma}_{m\times m} = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \rho\sigma^2 \dots & \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 \dots & \\ \rho\sigma^2 & \rho\sigma^2 & \sigma^2 \dots & \\ \dots & \dots & \dots & \dots \end{bmatrix}$$

- τ is the precision (inverse of the variance)
- $\mathbf{d}$ is the vector of the cross-validation results.
- $\mathbf{1}$ is a vector of ones.

Generative model - II

$$p(\delta, \tau) \sim \text{NormalGamma}(\mu_0, \text{k}_0, \text{a}, \text{b})$$

$$\mathbf{d}|\delta, \tau \sim \text{MVN}(\mathbf{1}\delta, \mathbf{\Sigma_{m \times m}})$$

$$\mathbf{\Sigma}_{m \times m} = \begin{bmatrix} \sigma^2 & \rho\sigma^2 & \rho\sigma^2 & \dots \\ \rho\sigma^2 & \sigma^2 & \rho\sigma^2 & \dots \\ \rho\sigma^2 & \rho\sigma^2 & \sigma^2 & \dots \\ \dots & \dots & \dots & \dots \end{bmatrix}$$

- The diagonal elements of $\mathbf{\Sigma}$ are equal.
- The variables have equal correlation $\rho = \frac{n_{te}}{n}$ with each other.

Posterior distribution of the difference of accuracy

- Prior and likelihood of the previous model are conjugate.
- The posterior joint distribution of δ, τ is Normal-Gamma.
- Marginalizing the precision out, the posterior distribution of δ is a Student.
- Adopting non-informative prior over the parameters:

$$p(\delta|\boldsymbol{d}) = St\left(\delta; n-1, \bar{d}, \sqrt{\left(\frac{1}{n} + \frac{\rho}{1-\rho}\right)\hat{\sigma}^2}\right)$$

Once standardized, it corresponds to the sampling distribution of the NHST correlated test.

Matching decisions

- The NHST and the Bayesian test take matching decision, if we aim at making inference about whether δ is positive or negative.

- This is done but running the one-sided NHST test
  $H_0 : \delta \leq 0$
  $H_1 : \delta > 0$
- and comparing its conclusion with the posterior probabilities:
  $P(\delta \leq 0)$
  $P(\delta > 0)$

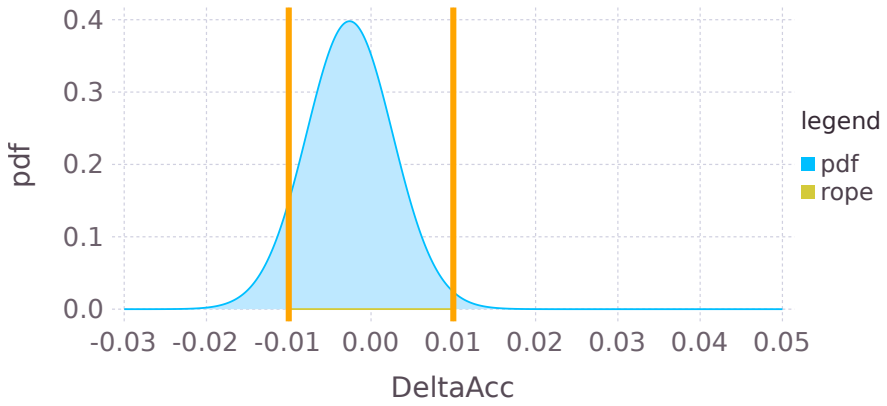
Further considerations

- The posterior depends only on prior and data, not on the sampling intentions.
- You get the full posterior distribution of δ, that you can analyze as you prefer.
- Checking the posterior probability of δ being positive or negative might be too simplistic.

The region of practical equivalence (ROPE)

- We want the probability of one classifier *practically equivalent* or *practically better* than the other.
- Rule of thumb:
  - two classifiers are practically equivalent if $-0.01 < \delta < 0.01$.
  - two classifiers are practically different if $|\delta| > 0.01$.
- The interval (-0.01, 0.01) is our region of practical equivalence (rope) (Kruschke and Liddell, 2015).

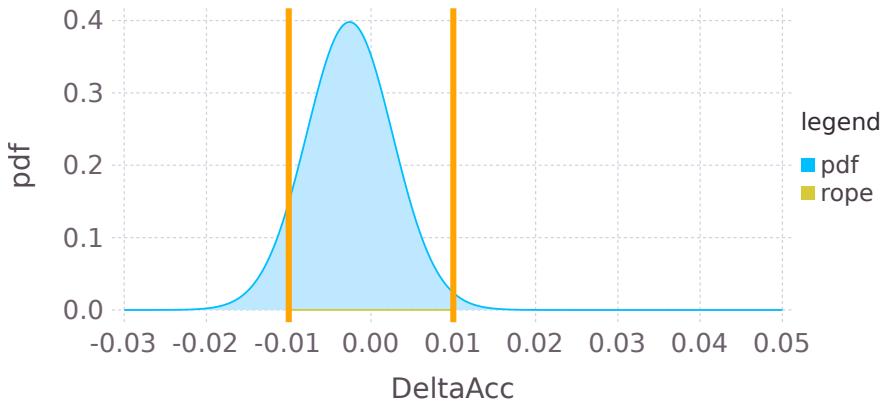
Computing with the ROPE - I



Posterior $P(\delta|\mathbf{d})$

- The probability mass within the ROPE is the probability of the two classifiers being practically equivalent.

Computing with the ROPE - II



Posterior $P(\delta|\mathbf{d})$

- The probability mass in the outer regions is the probability of one classifier being more accurate than the other, the difference being of practical interest.

Experimental comparison

- We consider five classifiers: naive Bayes (nbc), AODE, hidden naive Bayes (nnb), j48, j48 grafted.
- We run them in WEKA on 54 data sets, performing 10 runs of 10-folds cross-validation on each data set.
- Then we perform all the $54 \cdot 10 = 540$ pairwise comparisons and we compare the conclusions of the NHST and of the Bayesian test.

When NHST does not reject the null hypothesis

pair	Data sets	**Bayesian decision**	
	(out of 54)	P(rope) >.95	No decision
nbc–aode	35	6	29
nbc–hnb	30	0	30
nbc–j48	27	2	25
nbc–j48gr	27	2	25
...	...	...	...
j48–j48gr	50	40	10
total	341	74	267
rates		22%	78%

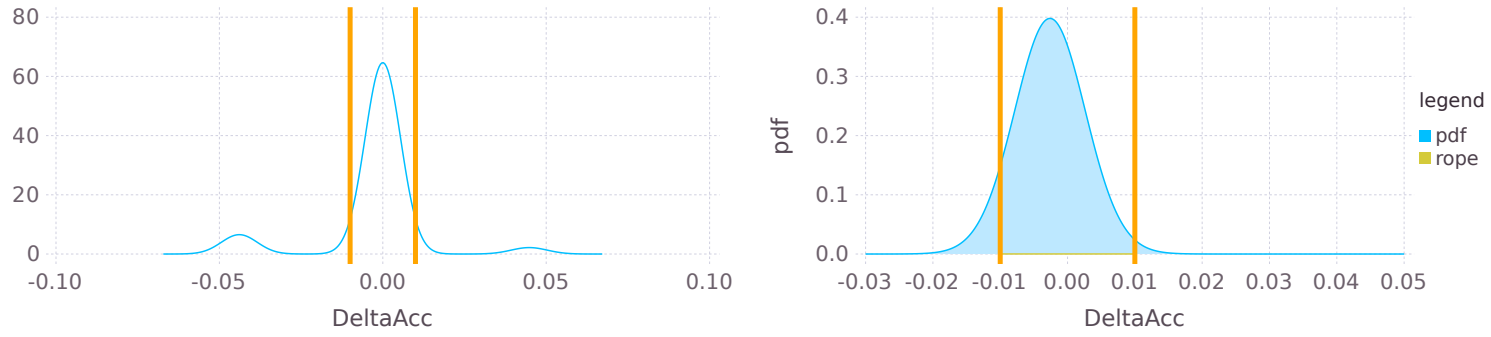
When NHST does not reject the null

	# non-rejections	**Bayesian decision**	
		P(rope) >.95	No decision
	341	74	267
rates		22%	78%

- in about 20% of such cases Bayesian analysis detects the two classifiers as practically equivalent with probability >95%.
- in the remaining cases it nevertheless provides you with informative posterior distributions.

Example: AODE vs naive Bayes (*audiology*)

- NHST: p-value = 0.6 (non informative)



Density plot Posterior $P(\delta|\mathbf{d})$

- Bayesian conclusion: δ belongs with probability 90% to the rope: more informative.

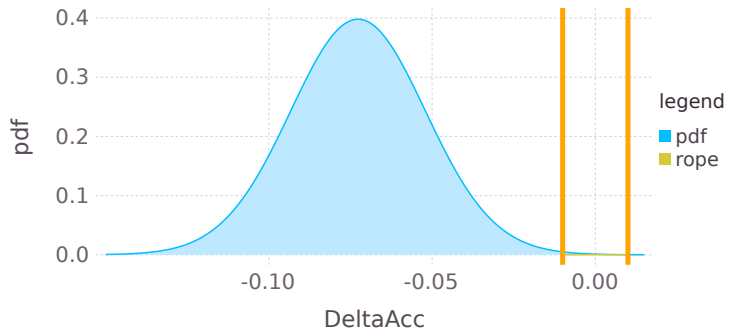
When NHST does reject the null

pair	Data sets (out of 54)	Bayesian decision (95%) rope	difference	no decision
nbc–aode	19	1	14	4
...	...	...	...	...
j48–j48gr	4	2	1	1
total	199	6	142	51
rates		3%	71%	26%

The Bayesian estimation procedure confirms the significance of only 70% of the significances claimed by NHST.

Example: AODE vs naive Bayes (*iris*)

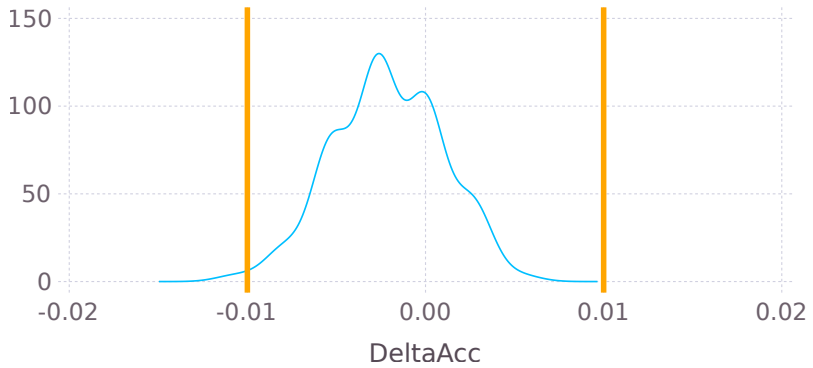
- NHST: p-value = 0.000



Posterior $P(\delta|\mathbf{d})$

- Bayesian conclusion: δ belongs with probability >95% to the *left* of the rope.
- Both test conclude AODE to be significantly more accurate than naive Bayes on iris.

Disagreeing conclusions - I

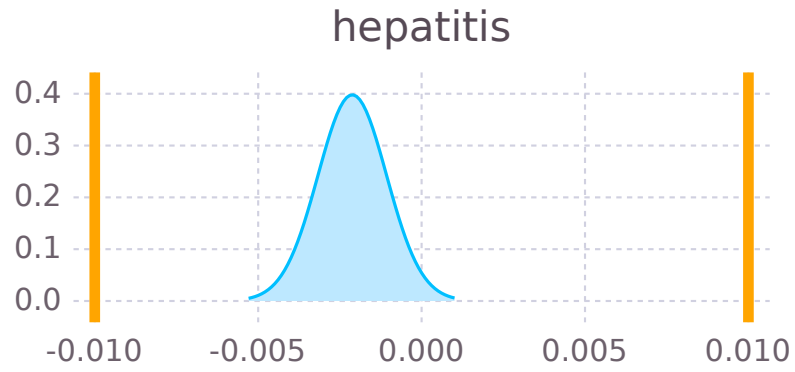


naive Bayes vs AODE on hepatitis

Density plot of the cross-validation differences.

- The d_i's are mostly negative and thus in favor of aode.
- Moreover, variance is low. Thus NHST claims significance (p<0.05).
- Yet the differences are small and lie within the rope.

Disagreeing conclusions - II



Posterior $P(\delta|\mathbf{d})$

- The posterior mean of δ is negative, but its entire distribution lies within the ROPE.
- Bayesian estimation declares the two classifiers to be practically equivalent.
- We argue that this conclusion is sensible.

Statistical significances which no practical impact

Statistical Science
2006, Vol. 21, No. 1, 1–15
DOI: 10.1214/088342306000000060

Classifier Technology and the Illusion of Progress

David J. Hand

Abstract. A great many tools have been developed for supervised classification, ranging from early methods such as linear discriminant analysis through to modern developments such as neural networks and support vector machines. A large number of comparative studies have been conducted in attempts to establish the relative superiority of these methods. This paper argues that these comparisons often fail to take into account important aspects of real problems, so that the apparent superiority of more sophisticated methods may be something of an illusion. In particular, simple methods typically yield performance almost as good as more sophisticated methods, to the extent that the difference in performance may be swamped by other sources of uncertainty that generally are not considered in the classical supervised classification paradigm.

Summing up

- A principle framework for inference on δ, overcoming the issues of NHST.
- Equipping the Bayesian test with rope allows to detect classifiers which are practically equivalent.
- Moreover, it allows claiming significances which have a practical impact.

A more general case

Comparing classifiers cross-validated on **multiple** data sets

Classifiers on multiple data sets (NHST)

	Cross-validation fold				*Mean*	*Var*
	1	2	...	10	$\bar{d}$	S_d^2
			Nursery			
Naive Bayes	.98	.98	...	.93		
Decision tree	.96	.97	...	.91		
Difference	.02	.01	...	.02	**1.5**	**0.3**
			Spambase			
Naive Bayes	.67	.67	...	.68		
Decision tree	.65	.69	...	.70		
Difference	.02	−.02	...	−.02	**-.02**	**.03**
			Further data sets ...			

- You have a collection of k data sets.

Goal

- You want to make inference about δ_0, the median difference of accuracy between the two classifiers in the population of data sets.

Classifiers on multiple data sets (NHST)

You **should** account for all the cross-validation results.

Instead you have to proceed as follows:

- compute the *average difference* of accuracy $\bar{d}$ on each dataset $(\bar{d}_1, \bar{d}_2, \ldots, \bar{d}_k)$
- treat the $\bar{d}_i$'s as i.i.d and run a NHST test, typically the signed rank
    - Unrealistic: Each data set has different complexity and sample size. Each estimate $\bar{d}_i$'s has different uncertainty!

The test outcome depends on k

- The signed-rank test makes non-parametric inference about δ_0.
- Its null hypothesis is $\delta_0 = 0$. As already pointed out, any point hypothesis is surely wrong.
- You can easily reject the null of the signed-rank, if you test on a large enough collection of data sets.

Bayesian approach

- We are unaware of previous Bayesian approaches able to compare classifiers on multiple data sets considering the variability and the correlation of the cross-validation results on each data set

- Thus we refer to our hierarchical model (Corani et al., 2016)

Hierarchical test: borrowed assumptions

It borrows the following assumptions from the Bayesian correlated t-test:

- The cross-validation differences on a given data are MVN-distributed.
- Each component of the MVN has mean δ, dev std. σ.
- The components are equally cross-correlated ($\rho = \frac{n_{te}}{n}$).

On the i-th data set we have:

- actual parameters δ_i and σ_i
- MLE estimates $\bar{d}_i$ and S_i

Supplementary assumptions

Both the δ_i's and the σ_i's are sampled from high-level distributions:

- The $\delta_i's$ are drawn from a a high-level Student distribution with mean δ_0.
- The Student distribution is more flexible than the Gaussian (it has three parameters). Moreover it is robust to outliers (used in robust Bayesian analysis).
- The σ_i's are drawn from a high level distribution $U(0, \bar{\sigma})$ as in literature (Gelman 2008).

Core of the model

$$\delta_1 ... \delta_k \sim t(\delta_0, \sigma_0, \nu)$$
$$\sigma_1 ... \sigma_k \sim \text{unif}(0, \bar{\sigma})$$
$$\mathbf{x}_i \sim MVN(\mathbf{1}\delta_i, \Sigma_i) \quad i = 1, 2, \ldots, k$$

- It explicitly allows each data set to have its own σ_i, getting rid of the unrealistic i.i.d. assumption.

Prior over the Student parameters

$$\delta_0 \sim \text{unif}(-1, 1)$$
$$\sigma_0 \sim \text{unif}(0, \bar{\sigma_0})$$
$$\nu \sim Ga(\alpha, \beta)$$

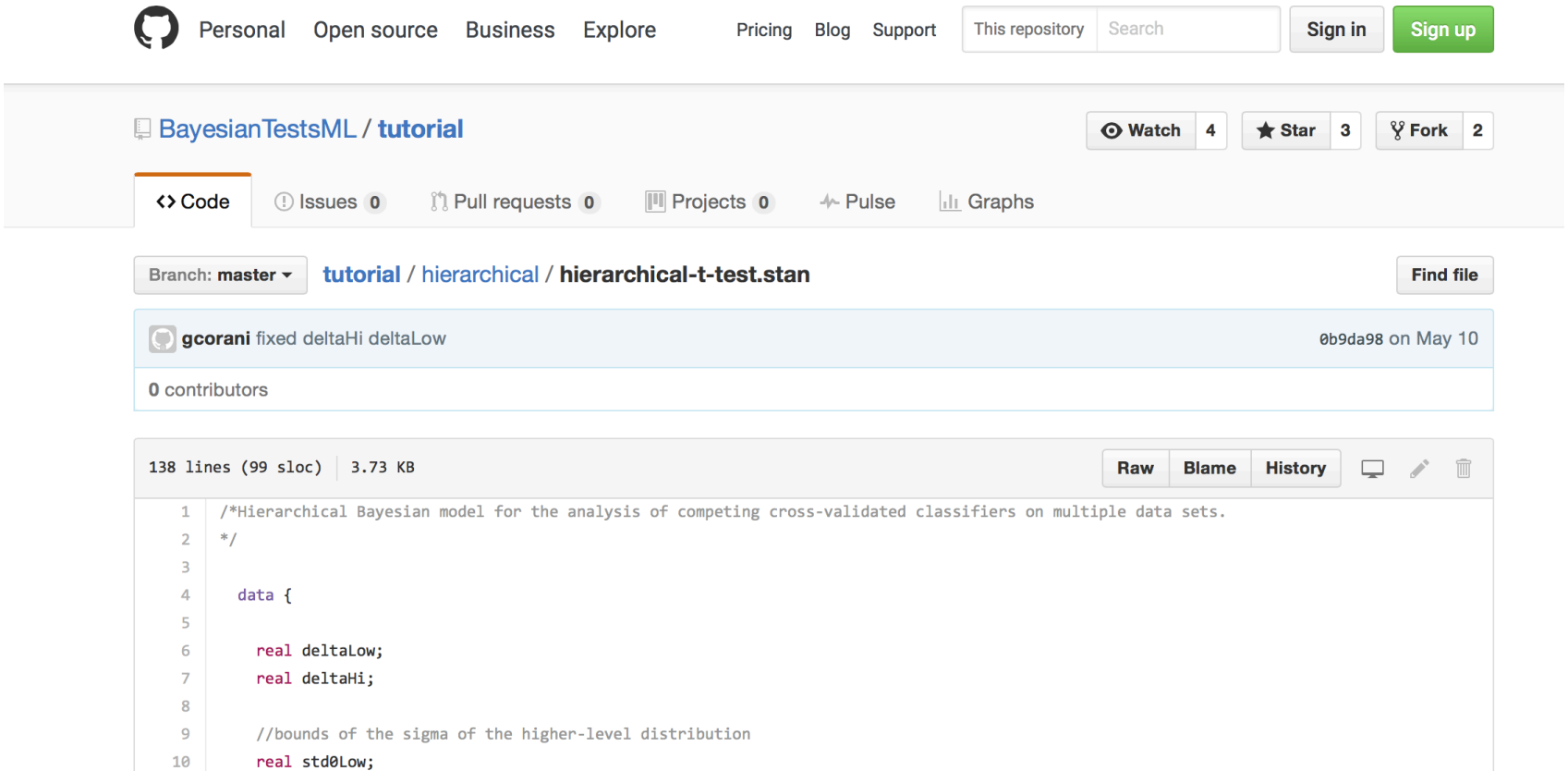
- The prior on δ_0 and σ_0 are non-informative and weakly informative respectively.
- The prior on the degrees of freedom is taken from (Kruschke 2015).

Shrinkage estimator

- The hierarchical model jointly estimates the δ_i's, while previous methods estimate independently each δ_i as $\bar{d}_i$.
- The hierarchical model applies *shrinkage* to the $\bar{d}_i$'s, which are brought closer to each other.
- Estimates on data sets characterized by more uncertainty (because more noisy or smaller sample size) are shrunken more (borrowing strenght).
- We prove both analytically and via simulation that such shrunken estimates are more accurate than the MLE estimates ($\bar{d}_i$).
- This is a general property of the shrinkage estimator, which is known to dominate the MLE one.

Inference of the model

- We infer the model parameters numerically using Stan (`http://mc-stan.org/`).
- All the code of this tutorial is available online.



Inference of the model

- Each data set yields 100 (10 runs of 10-folds cross-validation).
- Dealing with 50 data sets we thus analyze a matrix of 5000 numbers (rather than only 50 means as in the traditional case)
- This requires about 3 mins. on standard laptop.
- We then analyze the posterior samples of δ_0

Simulation framework - I

Let us simulate a population of data sets whose actual mean difference of accuracy is δ_0 .

Repeat 500 times for each value of $k = \{10, 20, 30, 40, 50\}$:

- For each of the k data sets draw:
  $\delta_i \sim N(\delta_0, 0.03) \;\; i = (1, 2, \ldots, k)$
  - differences of accuracy larger than 0.1 are rare
  - results are similar if we sample from a Student or even from a mixture.

For each sampled δ_i:

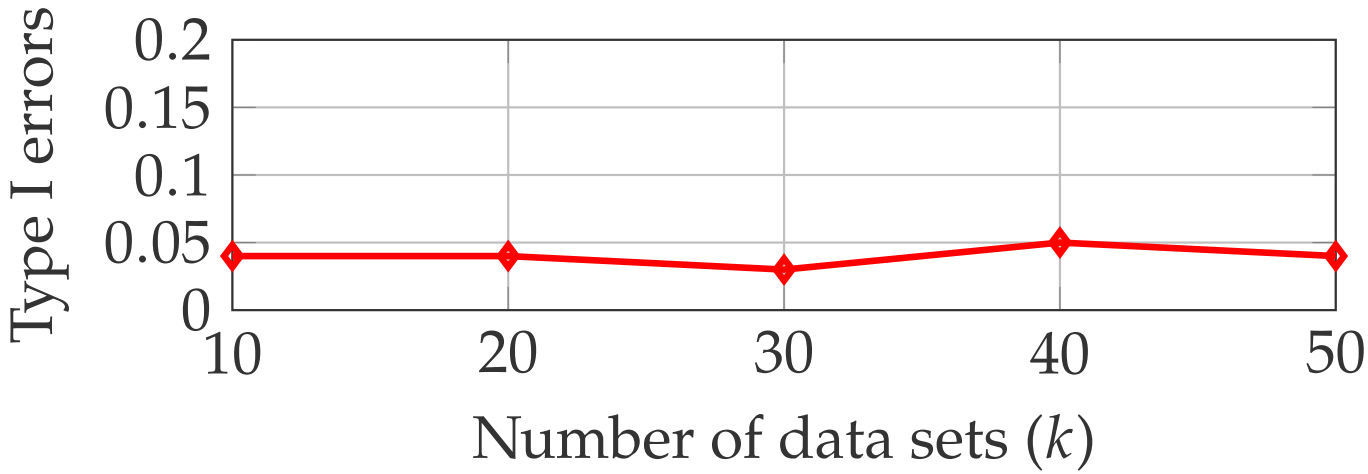
- implement a mechanism for sampling the training set and learning two classifiers so that their difference of accuracy is δ_i when learned on infinite samples.
- sample the training set and run cross-validation of the two classifiers, obtaining a vector of measures which fluctuate around δ_i.
- the variance of the measures mostly depends on the sample size of the training set.
- details on how to do this are given in the paper.

Taking decisions

Eventually we analyze the results obtained on the k data sets:

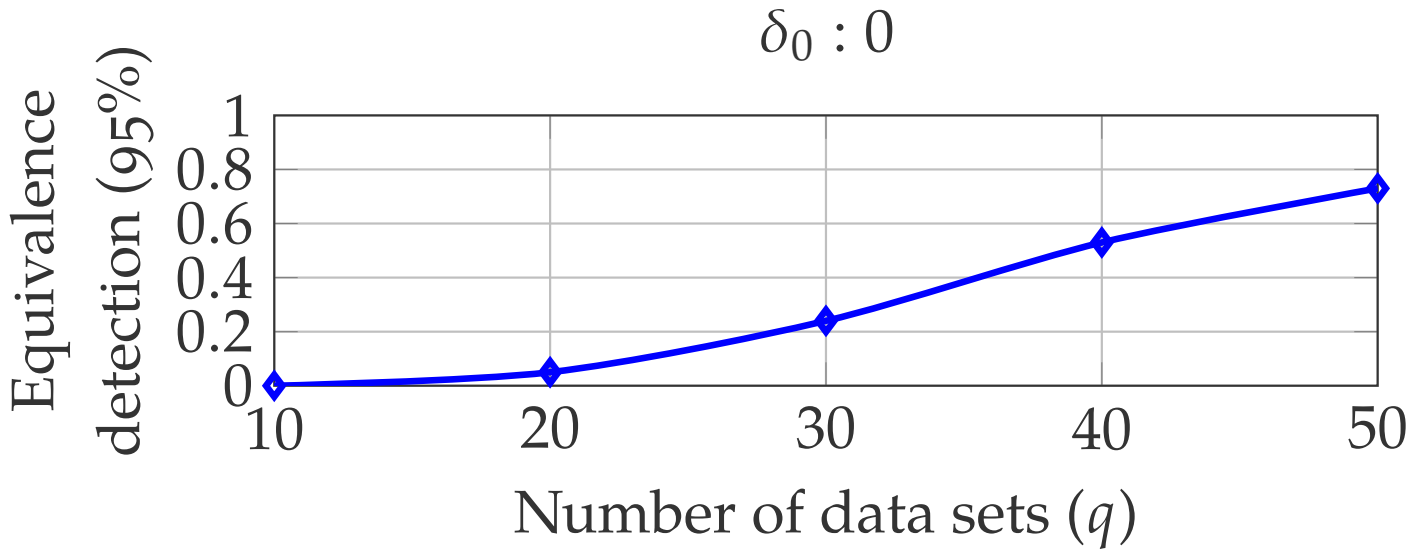
- with the signed-rank test (α=0.05)
- with the Bayesian hierarchical test
  - declaring them *practically equivalent* when we have more than 95% probability in the ROPE.
  - declaring them *significantly different* when we have more than than 95% probability in one of the regions surrounding the ROPE.

Two formally equivalent classifiers ($\delta_0 = 0$)



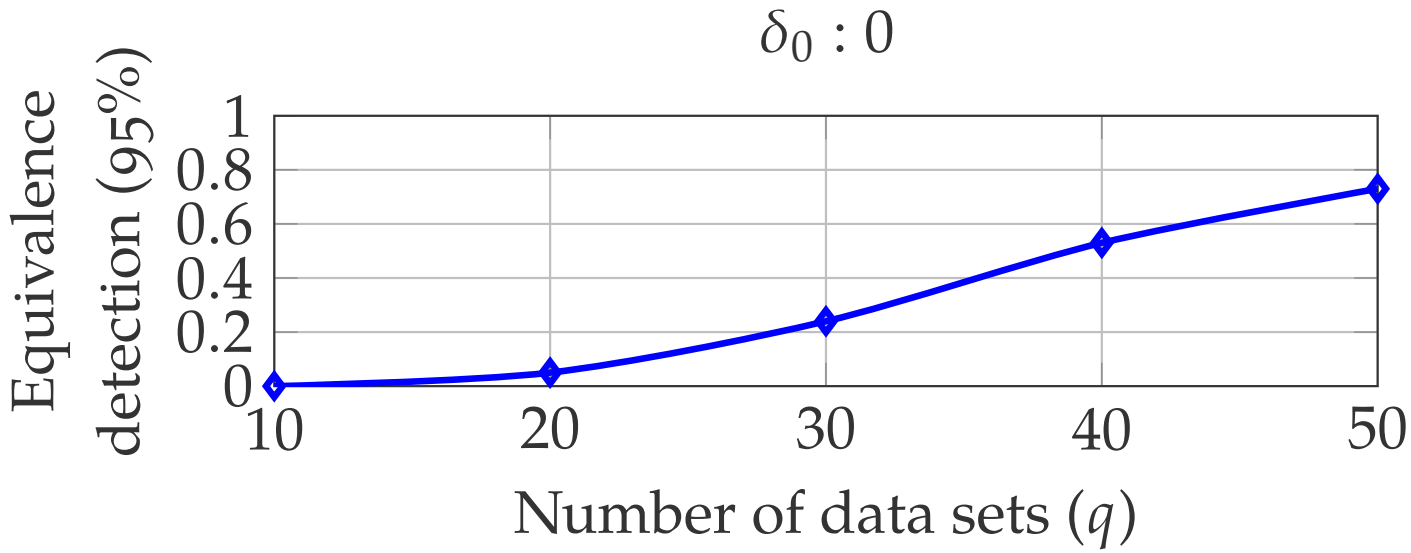
- The signed-rank is correctly calibrated, regardless the value of k.
- This might look great but:
  - It draws no conclusion in 95% of the cases.
  - It draws a wrong conclusion (Type I error) in the remaining 5%.

Two formally equivalent classifiers (hierarchical test)



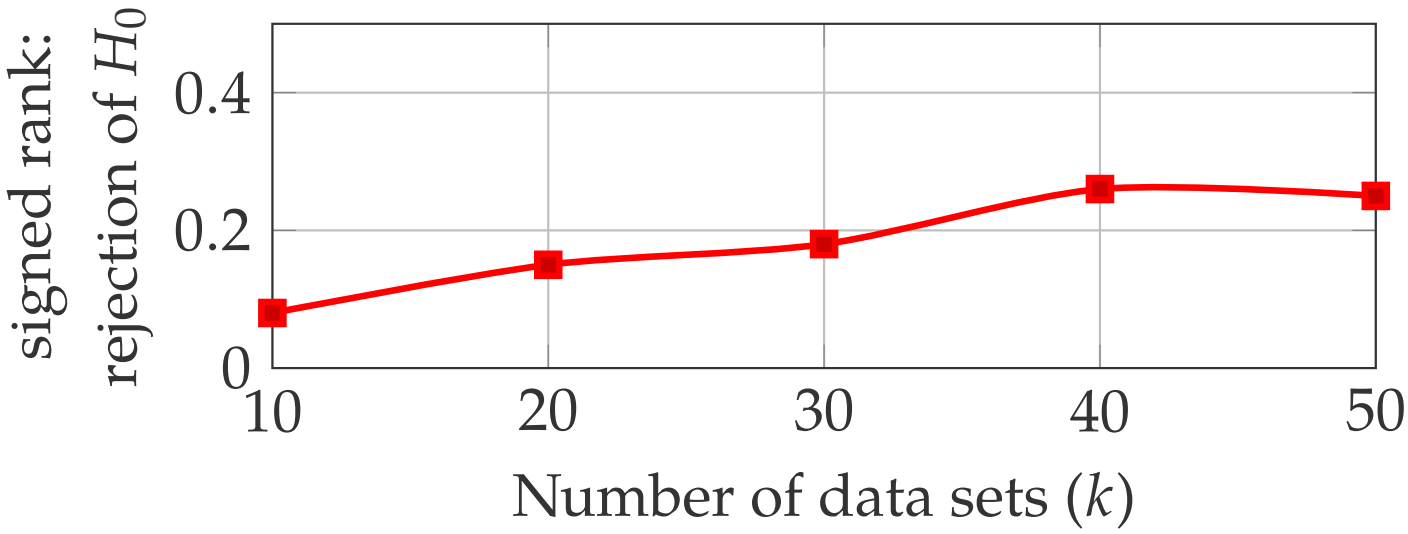
- First advantage: no Type I error.
- Second advantage: ability of recognizing the two classifiers as equivalent (we do so when the ROPE contains more than 95% of the posterior).

Two formally equivalent classifiers (hierarchical test)



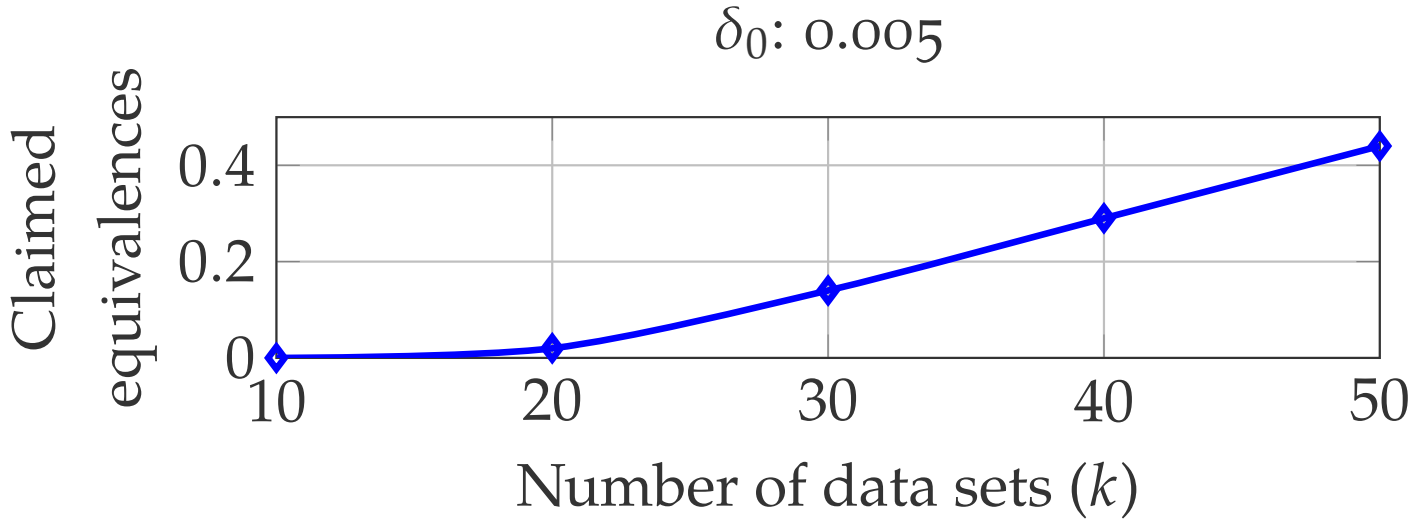
- Equivalence detection increases with k, as evidence accumulates.
- The model learns from the data!
- Instead the NHST does 5% Type I errors for any value of k.

Two practically equivalent classifiers (δ_0= 0.005)



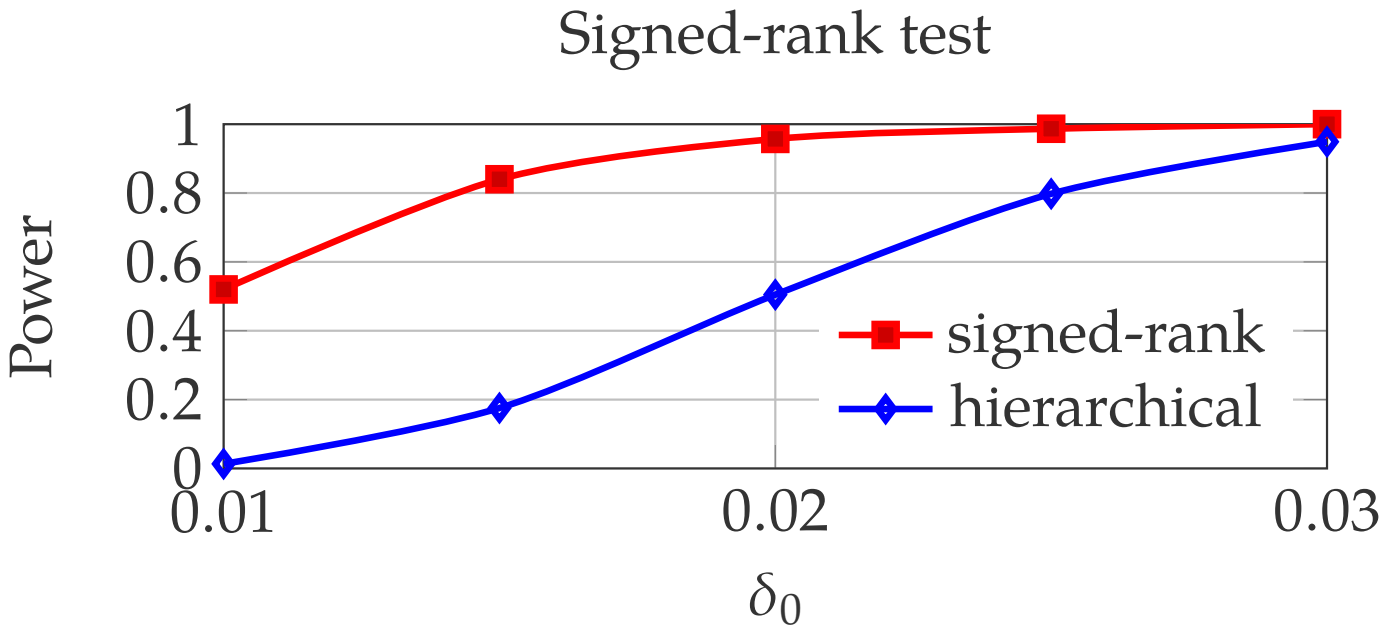
- The signed-rank rejects more H_0 as the number of data sets increases, despite the difference δ_0 being trivial.
- Beware statistical significance without practical effect!

Analyzing the same simulations with the hierarchical test



- The hierarchical test **never** declares the two classifiers as significantly different.
- Instead it is able to declare them as practically equivalent; its frequency in doing so increases with k.

Two significantly different classifiers (k = 50)



- We compute the frequency with which the signed-rank and the hierarchical test declare the two classifiers significantly different, varying δ_0.
- The signed-rank test is more powerful.

Analysis of real classifiers on 54 UCI data sets

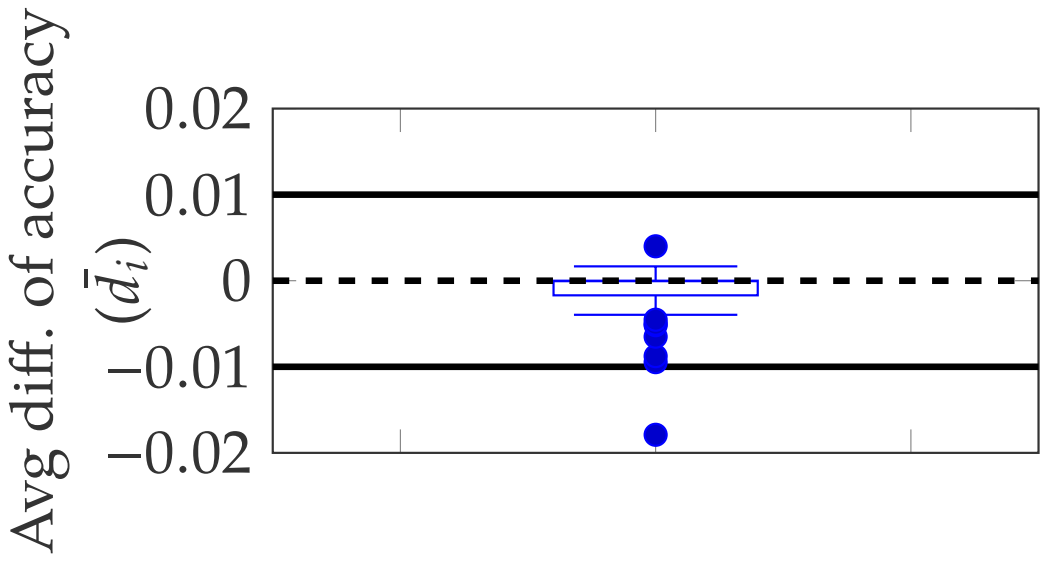
- We perform 10 runs of 10-folds cross-validation on each data set.
- All experiments have been performed using WEKA.

Analysis on real data sets

		Signed rank	Hierarchical test			
left	*right*	*p value*	*p(left)*	*p(rope)*	*p(right)*	
nbc	hnb	0.00	0.00	0.00	1.00	agree
nbc	j48	0.46	0.20	0.01	0.79	
hnb	j48gr	0.08	0.92	0.05	0.03	
j48	j48gr	0.00	0.00	1.00	0.00	disagree

- In the two middle cases no test take a decision.
- Yet the Bayesian outcome is more informative.

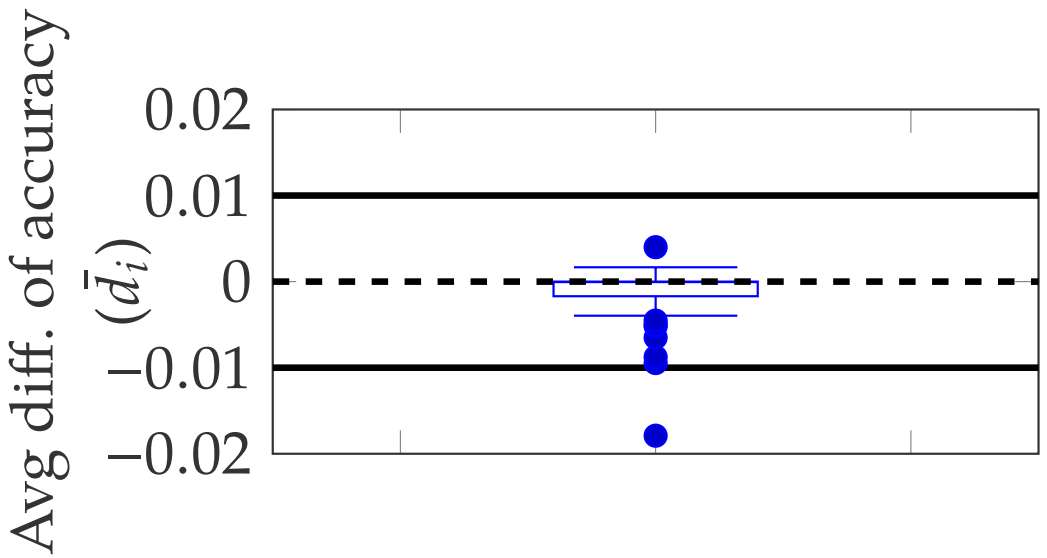
Analysis of the case J48 vs J48gr



Boxplots of the differences of accuracy $\bar{d}_i$'s on 54 data sets.

- The solid line shows the rope.
- The dashed line shows the 0.

Analysis of the case J48 vs J48gr



- Most difference are negative, thus in favor of J48gr. Variance is low. Thus NHST claims significance.
- But differences are small-sized. They all lie within the rope.
- We argue that the two classifiers should be regarded as practically equivalent.

Conclusions

- The Bayesian correlated and hierarchical t-test yield a more informative output than their NHST counterpart.
- They are more conservative in claiming significances and they are able to detect practically equivalent classifiers.
- In case of disagreement between the two test, the conclusion drawn on the basis of the Bayesian approach looks (at leat to us!) more sensible.
- Running:
  - the correlated t-test uses the same distribution of the NHST
  - the Stan code of the hierarchical test is available. It can be called from R, Python, Matlab, etc.