Mi clasificador es mejor que el tuyo

Breve introducción a la comparación estadística de clasificadores

Por: Mario Juez-Gil

UNIVERSIDAD DE BURGOS

Junta de Castilla y León
Consejería de Educación

UNIÓN EUROPEA
Fondo Social Europeo
El FSE invierte en tu futuro

Qué hago aquí

Qué hago aquí

- Investigo sobre *Machine Learning* y quiero saber cómo comparar clasificadores.

Qué hago aquí

- Investigo sobre *Machine Learning* y quiero saber cómo comparar clasificadores.
- Quiero aprender a interpretar los análisis comparativos que he visto en algunos *papers*.

Qué hago aquí

- Investigo sobre *Machine Learning* y quiero saber cómo comparar clasificadores.
- Quiero aprender a interpretar los análisis comparativos que he visto en algunos *papers*.
- Josema y José Antonio me han engañado para que venga.

Hoja de ruta

- Contexto: Machine learning, entrenamiento, validación cruzada, métricas de rendimiento…
- Comparación estadística basada en contraste de hipótesis
- Comparación bayesiana de clasificadores
- Líneas de trabajo futuras en este ámbito

Machine Learning

Machine Learning

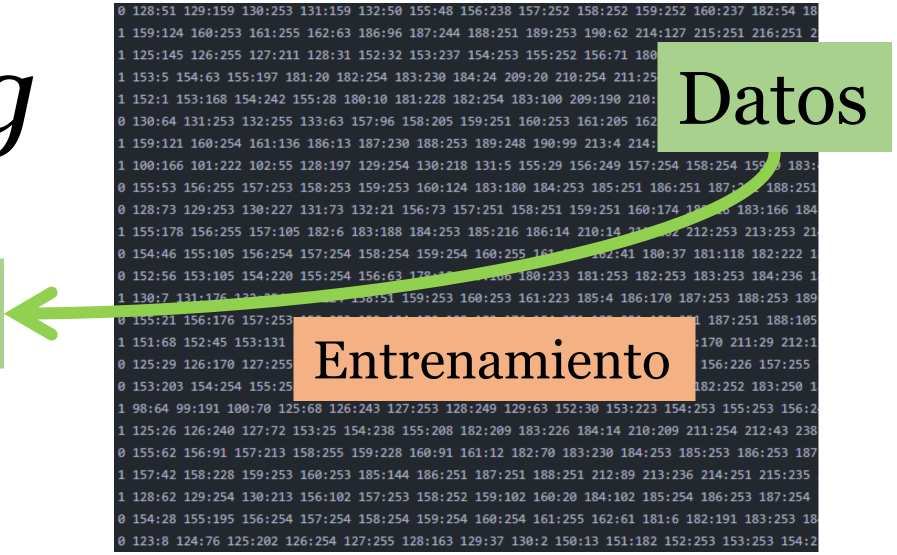
```
/**
  * Trains a Rotation Forest classification model.
  *
  * @param dataset Training dataset.
  * @return Rotation Forest classification model.
  */
@Since("2.4.4")
protected def train(dataset: Dataset[_]): RotationForestClassificationModel = {
  val categoricalFeatures =
    MetadataUtils.getCategoricalFeatures(dataset.schema($(featuresCol)))
  require(categoricalFeatures.size == 0, "Rotation Forest cannot work with " +
    "categorical features. Consider transforming the categorical features " +
    "using OneHotEncoderEstimator for example.")

  val numClasses = MetadataUtils.getNumClasses(dataset.schema($(labelCol))) match {
    case Some(n: Int) => n
    case None => throw new IllegalArgumentException("The dataset must " +
    "contain a label column with spark ML metadata. See spark ml " +
    "StringIndexer transformer for more information.")
  }

  val rnd = new Random($(seed))

  val numFeatures = countFeatures(dataset)
  val classes = 0d to (numClasses - 1) by 1d toList

  val (scaledDataset, scaler) = scaleDatasetAndGetScalerModel(dataset)
  val cachedDataset = scaledDataset.cache()
  val ensemble = buildEnsemble(cachedDataset, numFeatures, classes, rnd)
  cachedDataset.unpersist()

  new RotationForestClassificationModel(ensemble, scaler, numClasses)
}
```

Algoritmo

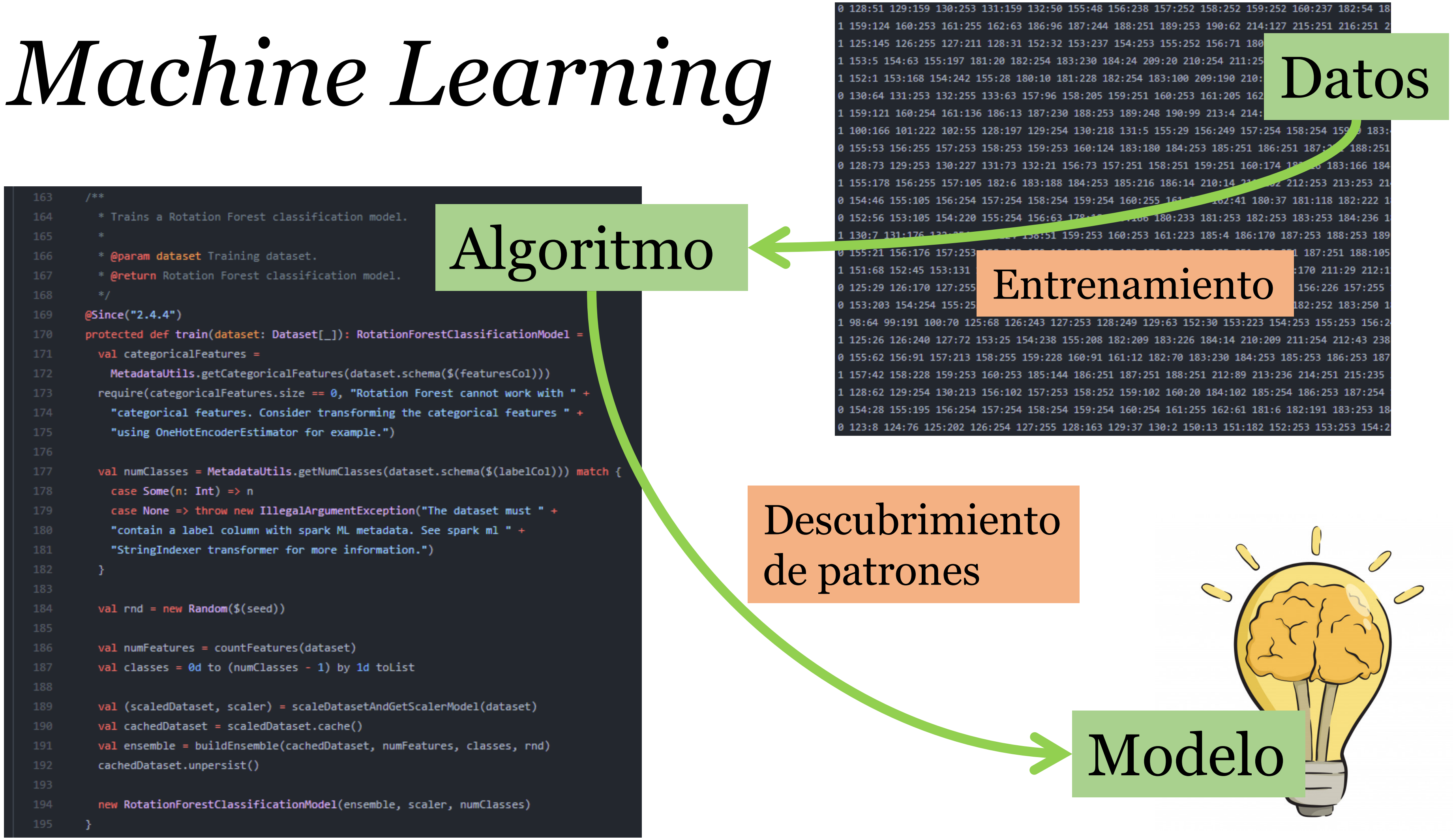
Machine Learning

```scala
/**
  * Trains a Rotation Forest classification model.
  *
  * @param dataset Training dataset.
  * @return Rotation Forest classification model.
  */
@Since("2.4.4")
protected def train(dataset: Dataset[_]): RotationForestClassificationModel = {
  val categoricalFeatures =
    MetadataUtils.getCategoricalFeatures(dataset.schema($(featuresCol)))
  require(categoricalFeatures.size == 0, "Rotation Forest cannot work with " +
    "categorical features. Consider transforming the categorical features " +
    "using OneHotEncoderEstimator for example.")

  val numClasses = MetadataUtils.getNumClasses(dataset.schema($(labelCol))) match {
    case Some(n: Int) => n
    case None => throw new IllegalArgumentException("The dataset must " +
    "contain a label column with spark ML metadata. See spark ml " +
    "StringIndexer transformer for more information.")
  }

  val rnd = new Random($(seed))

  val numFeatures = countFeatures(dataset)
  val classes = 0d to (numClasses - 1) by 1d toList

  val (scaledDataset, scaler) = scaleDatasetAndGetScalerModel(dataset)
  val cachedDataset = scaledDataset.cache()
  val ensemble = buildEnsemble(cachedDataset, numFeatures, classes, rnd)
  cachedDataset.unpersist()

  new RotationForestClassificationModel(ensemble, scaler, numClasses)
}
```

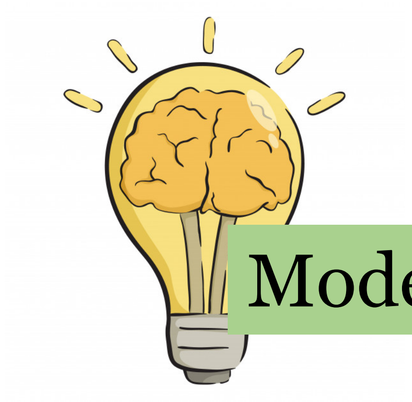
Algoritmo



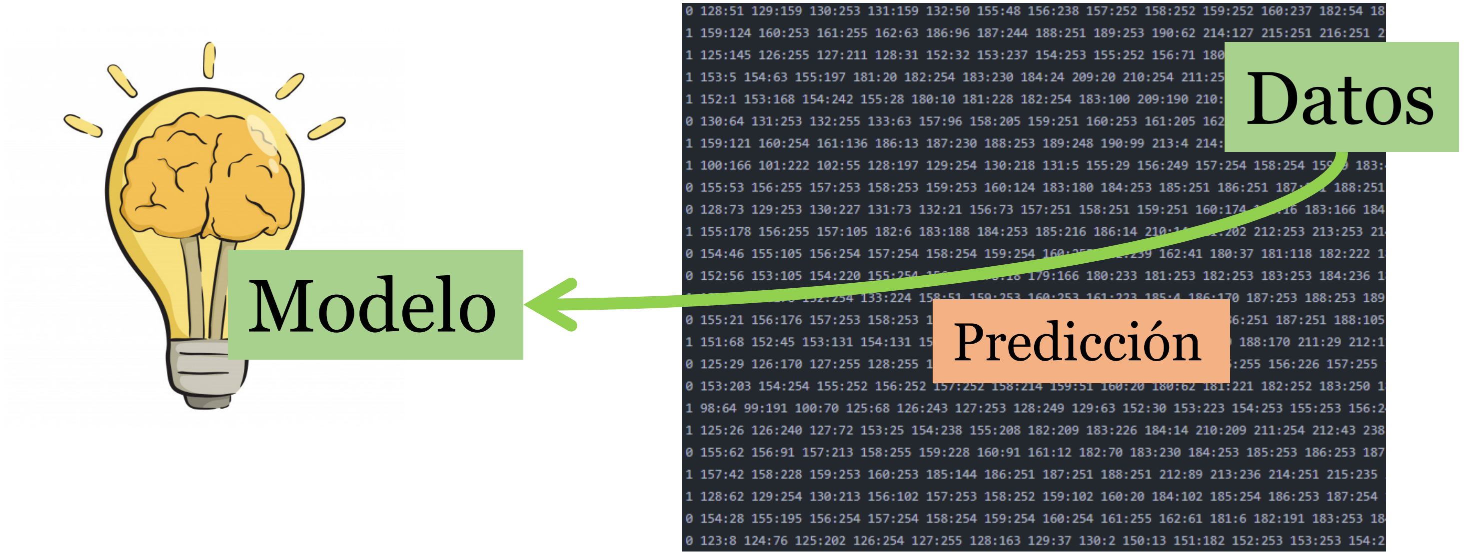
Machine Learning

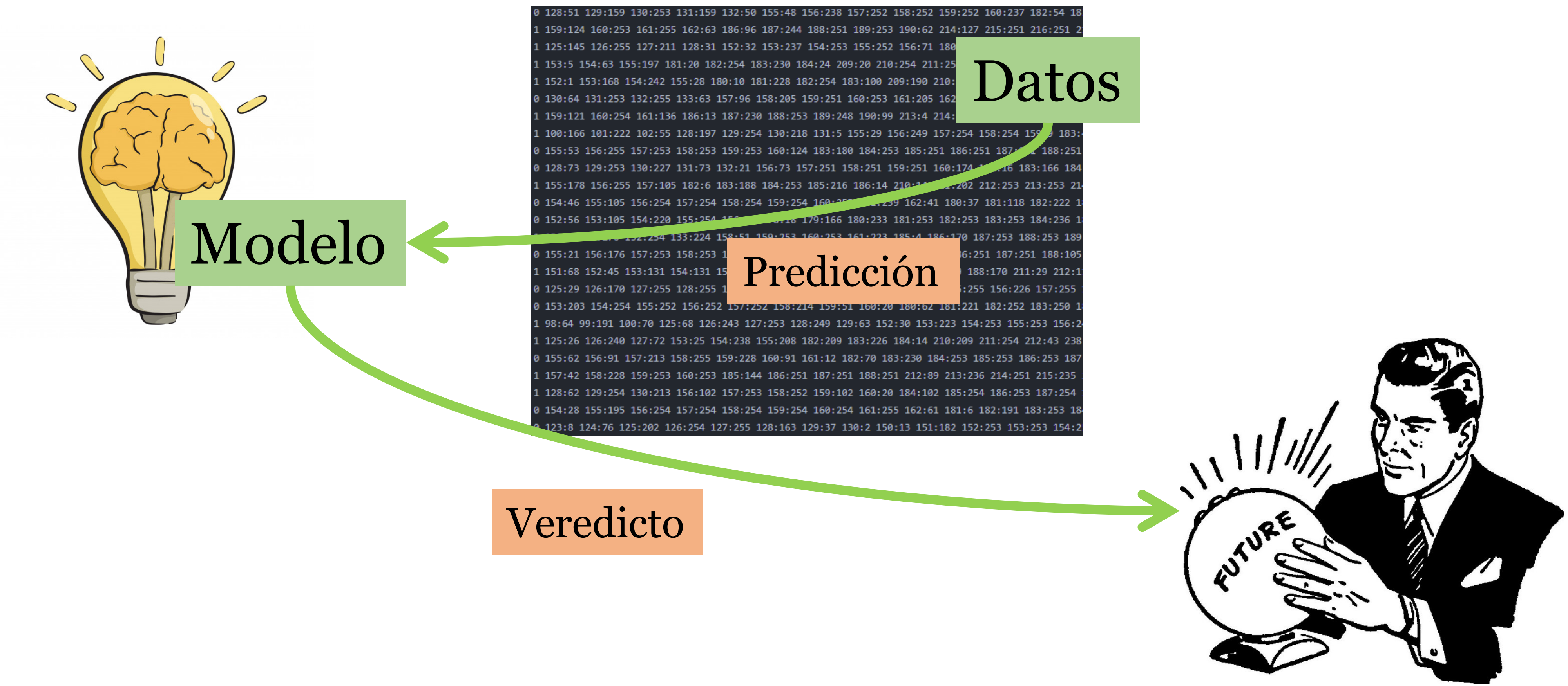


UNIVERSIDAD DE BURGOS
Junta de Castilla y León
Consejería de Educación
UNIÓN EUROPEA
Fondo Social Europeo
El FSE invierte en tu futuro

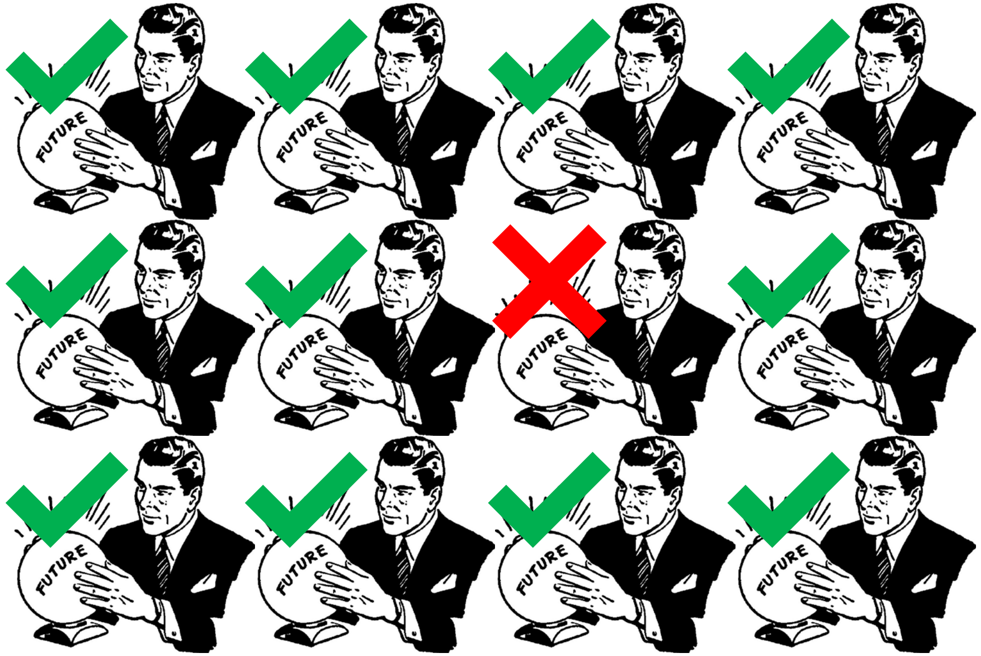


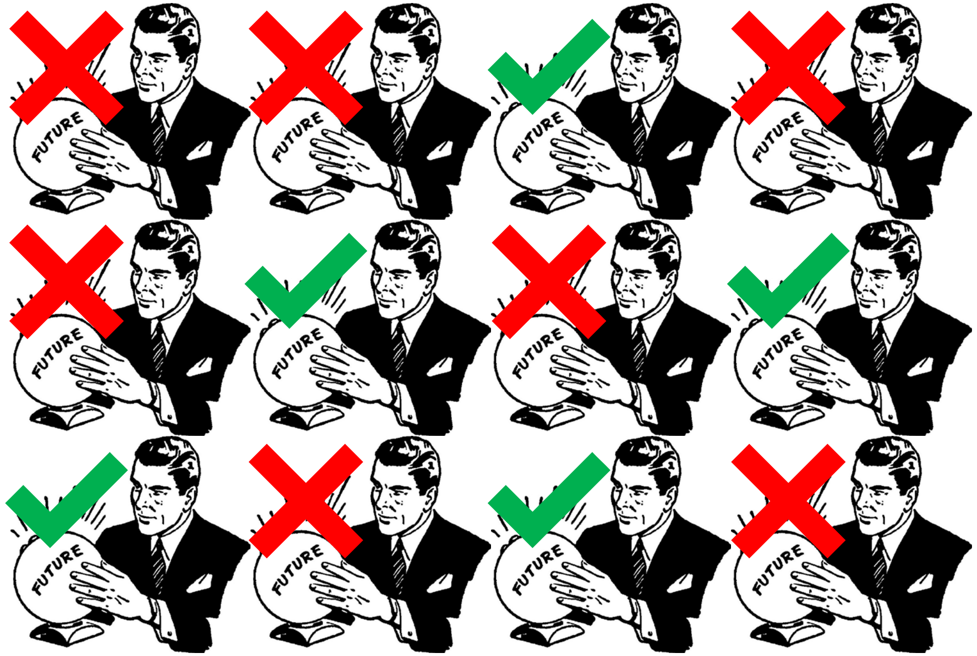
Modelo


Datos
Modelo
Predicción


Datos
Modelo
Predicción
Veredicto
FUTURE

Mi clasificador 😎 Tu clasificador 🤕



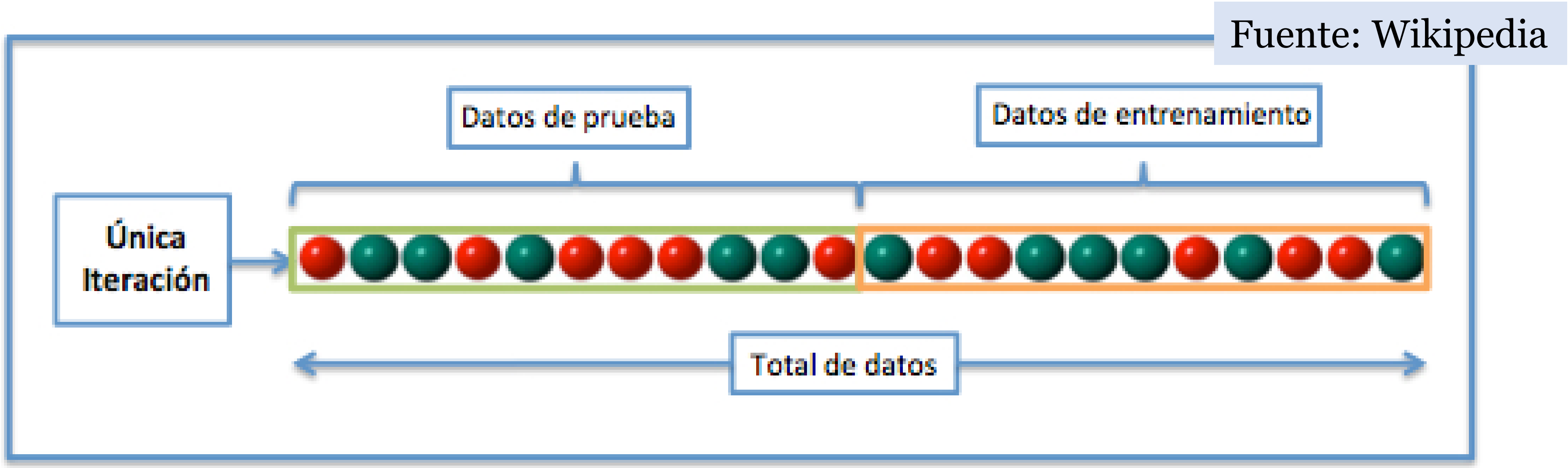


Mi clasificador 🤔

Tu clasificador 🤔

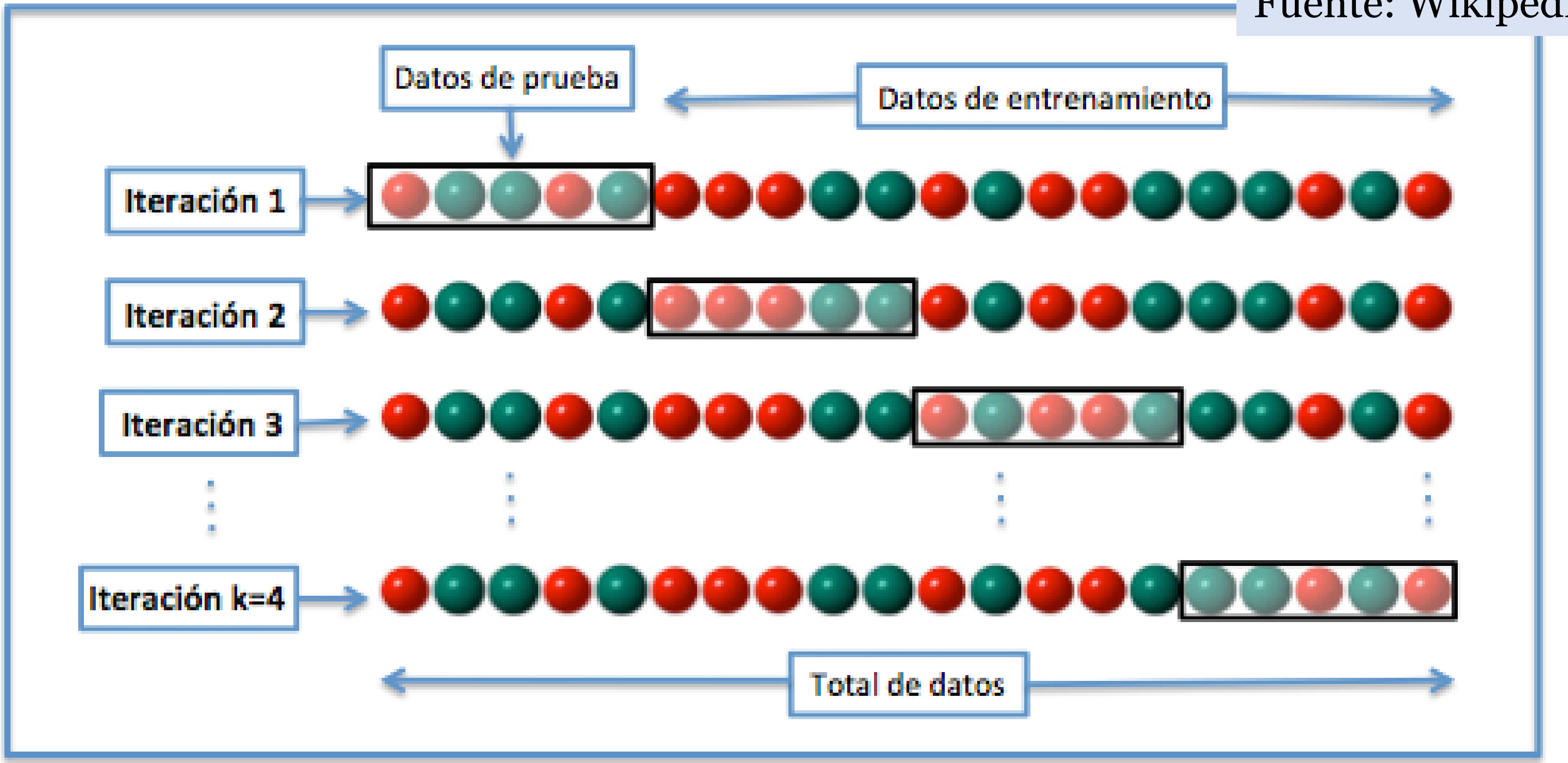
UNIVERSIDAD DE BURGOS
Junta de Castilla y León
Consejería de Educación
UNIÓN EUROPEA
Fondo Social Europeo
El FSE invierte en tu futuro

División de datos de entrenamiento



Validación cruzada

Fuente: Wikipedia



Métricas de rendimiento

Accuracy

Métricas de rendimiento

ROC AUC

MAE

F-Score

Accuracy

MCC

RMSE

G-Mean

Métricas de rendimiento

- ROC AUC
- MAE
- F-Score
- Accuracy
- MCC
- RMSE
- G-Mean

UNIVERSIDAD DE BURGOS
Junta de Castilla y León
Consejería de Educación
UNIÓN EUROPEA
Fondo Social Europeo
El FSE invierte en tu futuro

Métricas de rendimiento

ROC AUC

MAE

F-Score

Accuracy

MCC

RMSE

G-Mean

Métricas de rendimiento

ROC AUC

F-Score

MAE

Accuracy

MCC

RMSE

G-Mean

Métricas de rendimiento

ROC AUC

MAE

F-Score

Accuracy

RMSE

G-Mean

MCC

Comparación de clasificadores

Mi clasificador

FOLD	ACCURACY
1	0,68
2	0,72
3	0,88
4	0,61
5	0,75
media	**0,73 ± 0,1**

Tu clasificador

FOLD	ACCURACY
1	0,63
2	0,75
3	0,76
4	0,64
5	0,69
media	**0,69 ± 0,06**

Comparación de clasificadores

Mi clasificador 😎

FOLD	ACCURACY
1	0,68
2	0,72
3	0,88
4	0,61
5	0,75
media	**0,73 ± 0,1**

Tu clasificador 🤕

FOLD	ACCURACY
1	0,63
2	0,75
3	0,76
4	0,64
5	0,69
media	**0,69 ± 0,06**

0,73 > 0,69

Comparación de clasificadores

Mi clasificador 😎

FOLD	ACCURACY
1	0,68
2	0,72
3	0,88
4	0,61
5	0,75
media	**0,73 ± 0,1**

Tu clasificador 😟

FOLD	ACCURACY
1	0,63
2	0,75
3	0,76
4	0,64
5	0,69
media	**0,69 ± 0,06**

0,73 > 0,69

- ¿Un conjunto de datos es suficiente?

Comparación de clasificadores

Mi clasificador 😎 Tu clasificador 😟

FOLD	ACCURACY
1	0,68
2	0,72
3	0,88
4	0,61
5	0,75
media	**0,73 ± 0,1**

FOLD	ACCURACY
1	0,63
2	0,75
3	0,76
4	0,64
5	0,69
media	**0,69 ± 0,06**

0,73 > 0,69

- ¿Un conjunto de datos es suficiente?
- ¿Diferencia <u>significativa</u>?

Comparación de clasificadores

Mi clasificador 😎 Tu clasificador 😟

FOLD	ACCURACY
1	0,68
2	0,72
3	0,88
4	0,61
5	0,75
media	0,73 ± 0,1

FOLD	ACCURACY
1	0,63
2	0,75
3	0,76
4	0,64
5	0,69
media	0,69 ± 0,06

0,73 > 0,69

- ¿Un conjunto de datos es suficiente?
- ¿Diferencia significativa?

Comparación estadística de clasificadores

2006

Journal of Machine Learning Research 7 (2006) 1–30 Submitted 8/04; Revised 4/05; Published 1/06

Statistical Comparisons of Classifiers over Multiple Data Sets

Janez Demšar JANEZ.DEMSAR@FRI.UNI-LJ.SI
Faculty of Computer and Information Science
Tržaška 25
Ljubljana, Slovenia

Editor: Dale Schuurmans

Abstract

While methods for comparing two learning algorithms on a single data set have been scrutinized for quite some time already, the issue of statistical tests for comparisons of more algorithms on multiple data sets, which is even more essential to typical machine learning studies, has been all but ignored. This article reviews the current practice and then theoretically and empirically examines several suitable tests. Based on that, we recommend a set of simple, yet safe and robust non-parametric tests for statistical comparisons of classifiers: the Wilcoxon signed ranks test for comparison of two classifiers and the Friedman test with the corresponding post-hoc tests for comparison of more classifiers over multiple data sets. Results of the latter can also be neatly presented with the newly introduced CD (critical difference) diagrams.

Keywords: comparative studies, statistical methods, Wilcoxon signed ranks test, Friedman test, multiple comparisons tests

2008

On the Appropriateness of Statistical Tests in Machine Learning

Janez Demšar JANEZ.DEMSAR@FRI.UNI-LJ.SI
University of Ljubljana, Faculty of Computer Science, Tržaška 25, Ljubljana, Slovenia

Abstract

One of the greatest machine learning problems of today is an intractable number of new algorithms being presented at our conferences, workshops and journals. A similar rush of ideas and results also plagues most other scientific fields and some have already questioned the usefulness of statistical tests for telling the true relations from the false. Statistical tests have been criticized as conceptually wrong almost from their inception. They do not work well in situations when numerous groups conduct similar research. Not measuring what we are really interested in, they can promote the randomly successful ideas instead of the good but unlucky ones. We unfortunately see no established alternatives in other fields of science which could be transplanted to our field. We however speculate on a possible spontaneously appearing solution in a form of a worldwide peer review.

Machine learning is being suffocated by the ease with which we can *generate* new algorithms or, in most cases, slight variations of the existing ones by using flexible and extendible frameworks such as Weka (Wit-

cessful in beating the competition, although usually only marginally.

This perpetual enhancement of our methods – often without the new methods actually getting any wide attention and use after being published – should be suspicious and alarming by itself. Instead, we are getting used to read and hear about new and *statistically significantly* better methods... and do not pay any attention to them.

This pessimistic paper will first discuss why null-hypothesis significance testing is problematic in principle, next section will show why it is becoming inapplicable in most modern science including ours, followed and concluded by a section presenting several non-viable alternatives. The basic message of the paper is, however, that any evaluations and comparisons – statistical or non-statistical – of new methods should be taken with a grain of salt (as well as this paper itself).

1. Objections to Significance Tests

Null-hypothesis significance testing (NHST) is aimed at distinguishing between random and non-random differences or, in general, relations found in experimental results. Yet its appropriateness has been dis-

2017

Journal of Machine Learning Research 18 (2017) 1-36 Submitted 6/16; Revised 5/17; Published 8/17

Time for a Change: a Tutorial for Comparing Multiple Classifiers Through Bayesian Analysis

Alessio Benavoli[†] ALESSIO@IDSIA.CH
Giorgio Corani[†] GIORGIO@IDSIA.CH
Janez Demšar[♮] JANEZ.DEMSAR@FRI.UNI-LJ.SI
Marco Zaffalon[†] ZAFFALON@IDSIA.CH
[♮]*Faculty of Computer and Information Science, University of Ljubljana, Vecna pot 113, SI-1000 Ljubljana, Slovenia*
[†]*Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA) Galleria 2, 6928 Manno, Switzerland*

Editor: David Barber

Abstract

The machine learning community adopted the use of null hypothesis significance testing (NHST) in order to ensure the statistical validity of results. Many scientific fields however realized the shortcomings of frequentist reasoning and in the most radical cases even banned its use in publications. We should do the same: just as we have embraced the Bayesian paradigm in the development of new machine learning methods, so we should also use it in the analysis of our own results. We argue for abandonment of NHST by exposing its fallacies and, more importantly, offer better—more sound and useful—alternatives for it.

Keywords: comparing classifiers, null hypothesis significance testing, pitfalls of p-values, Bayesian hypothesis tests, Bayesian correlated t-test, Bayesian hierarchical correlated t-test, Bayesian signed-rank test

Comparación estadística de clasificadores

Antes de 2006:

- t-test
  - Test paramétrico
  - Muestras deben seguir una distribución normal o tamaño de la muestra: +30
  - Sensible a outliers (datos anómalos bajan fiabilidad del test)

Comparación estadística de clasificadores

Antes de 2006:

- t-test
  - Test paramétrico
  - Muestras d[...] una dist[...] t[...] +30
  - Se[...]ers (datos an[...]os bajan fiabilidad del test)

NO SON ADECUADOS

A partir de 2006:

- Tests no paramétricos
  - Las muestras no deben ajustarse a distribuciones
  - Basados en rankings
  - Ej: Wilcoxon Signed-Rank

Comparación estadística de clasificadores

Contraste de hipótesis:

- H_0 (hipótesis nula): Rendimiento equivalente
- H_1 (hipótesis alternativa): El rendimiento no es equivalente

Rechazo de hipótesis → mi clasificador es mejor que el tuyo

UNIVERSIDAD DE BURGOS

Junta de Castilla y León
Consejería de Educación

UNIÓN EUROPEA
Fondo Social Europeo
El FSE invierte en tu futuro

Wilcoxon Signed-Rank Test

Ranking de diferencias

Dataset	Tu Clasificador	Mi Clasificador	Diferencia
adult (sample)	0.763	0.768	+0.005
breast cancer	0.599	0.591	−0.008
bc wisconsin	0.954	0.971	+0.017
cmc	0.628	0.661	+0.033
ionosphere	0.882	0.888	+0.006
iris	0.936	0.931	−0.005
liver disorders	0.661	0.668	+0.007
lung cancer	0.583	0.583	0.000
lymphography	0.775	0.838	+0.063
mushroom	1.000	1.000	0.000
primary tumor	0.940	0.962	+0.022
rheum	0.619	0.666	+0.047
voting	0.972	0.981	+0.009
wine	0.957	0.978	+0.021

Wilcoxon Signed-Rank Test

Ranking de diferencias

Dataset	Tu Clasificador	Mi Clasificador	Diferencia	Rank
adult (sample)	0.763	0.768	+0.005	3.5
breast cancer	0.599	0.591	−0.008	7
bc wisconsin	0.954	0.971	+0.017	9
cmc	0.628	0.661	+0.033	12
ionosphere	0.882	0.888	+0.006	5
iris	0.936	0.931	−0.005	3.5
liver disorders	0.661	0.668	+0.007	6
lung cancer	0.583	0.583	0.000	1.5
lymphography	0.775	0.838	+0.063	14
mushroom	1.000	1.000	0.000	1.5
primary tumor	0.940	0.962	+0.022	11
rheum	0.619	0.666	+0.047	13
voting	0.972	0.981	+0.009	8
wine	0.957	0.978	+0.021	10

Wilcoxon Signed-Rank Test

Ranking de diferencias

Dataset	Tu Clasificador	Mi Clasificador	Diferencia	Rank
adult (sample)	0.763	0.768	+0.005	3.5
breast cancer	0.599	0.591	−0.008	7
bc wisconsin	0.954	0.971	+0.017	9
cmc	0.628	0.661	+0.033	12
ionosphere	0.882	0.888	+0.006	5
iris	0.936	0.931	−0.005	3.5
liver disorders	0.661	0.668	+0.007	6
lung cancer	0.583	0.583	0.000	1.5
lymphography	0.775	0.838	+0.063	14
mushroom	1.000	1.000	0.000	1.5
primary tumor	0.940	0.962	+0.022	11
rheum	0.619	0.666	+0.047	13
voting	0.972	0.981	+0.009	8
wine	0.957	0.978	+0.021	10

R+ = 93

R- = 12

Wilcoxon Signed-Rank Test

R+ = 93, R- = 12

T = min(R+, R-) = 12

N = 14

Tabla de valores críticos

	alpha values						
n	0.001	0.005	0.01	0.025	0.05	0.10	0.20
5	--	--	--	--	--	0	2
6	--	--	--	--	0	2	3
7	--	--	--	0	2	3	5
8	--	--	0	2	3	5	8
9	--	0	1	3	5	8	10
10	--	1	3	5	8	10	14
11	0	3	5	8	10	13	17
12	1	5	7	10	13	17	21
13	2	7	9	13	17	21	26
14	4	9	12	17	21	25	31
15	6	12	15	20	25	30	36
16	8	15	19	25	29	35	42
17	11	19	23	29	34	41	48
18	14	23	27	34	40	47	55
19	18	27	32	39	46	53	62
20	21	32	37	45	52	60	69
21	25	37	42	51	58	67	77
22	30	42	48	57	65	75	86
23	35	48	54	64	73	83	94
24	40	54	61	72	81	91	104
25	45	60	68	79	89	100	113
26	51	67	75	87	98	110	124
27	57	74	83	96	107	119	134

Wilcoxon Signed-Rank Test

$R+ = 93$, $R- = 12$
$T = \min(R+, R-) = 12$
$N = 14$

$12 < 21$,
se rechaza H_o,
Mi clasificador es mejor que el tuyo

Tabla de valores críticos

	alpha values						
n	0.001	0.005	0.01	0.025	0.05	0.10	0.20
5	--	--	--	--	--	0	2
6	--	--	--	--	0	2	3
7	--	--	--	0	2	3	5
8	--	--	0	2	3	5	8
9	--	0	1	3	5	8	10
10	--	1	3	5	8	10	14
11	0	3	5	8	10	13	17
12	1	5	7	10	13	17	21
13	2	7	9	13	17	21	26
14	4	9	12	17	21	25	31
15	6	12	15	20	25	30	36
16	8	15	19	25	29	35	42
17	11	19	23	29	34	41	48
18	14	23	27	34	40	47	55
19	18	27	32	39	46	53	62
20	21	32	37	45	52	60	69
21	25	37	42	51	58	67	77
22	30	42	48	57	65	75	86
23	35	48	54	64	73	83	94
24	40	54	61	72	81	91	104
25	45	60	68	79	89	100	113
26	51	67	75	87	98	110	124
27	57	74	83	96	107	119	134

Problemas del contraste de hipótesis

$\alpha = 0.05$

Mi clasificador es mejor que el tuyo con un 95% de probabilidad

Problemas del contraste de hipótesis

$\alpha = 0.05$

~~Mi clasificador es mejor que el tuyo con un 95% de probabilidad~~

Problemas del contraste de hipótesis

$\alpha = 0.05$

Hay un 5% de probabilidad de que dos clasificadores iguales presenten las mismas (o mayores) diferencias vistas en los experimentos.

Problemas del contraste de hipótesis

Pero yo quiero saber:

¿Con qué probabilidad mi clasificador es mejor que el tuyo?

Teorema de Bayes

Teorema de Bayes

Tu clasificador y mi clasificador son iguales

Resultados experimentales

Mi clasificador es mejor que el tuyo
(con un 80% de probabilidad)

Teorema de Bayes

$$P(A|E) = \frac{P(E|A)P(A)}{P(E)}$$

Probabilidad a priori

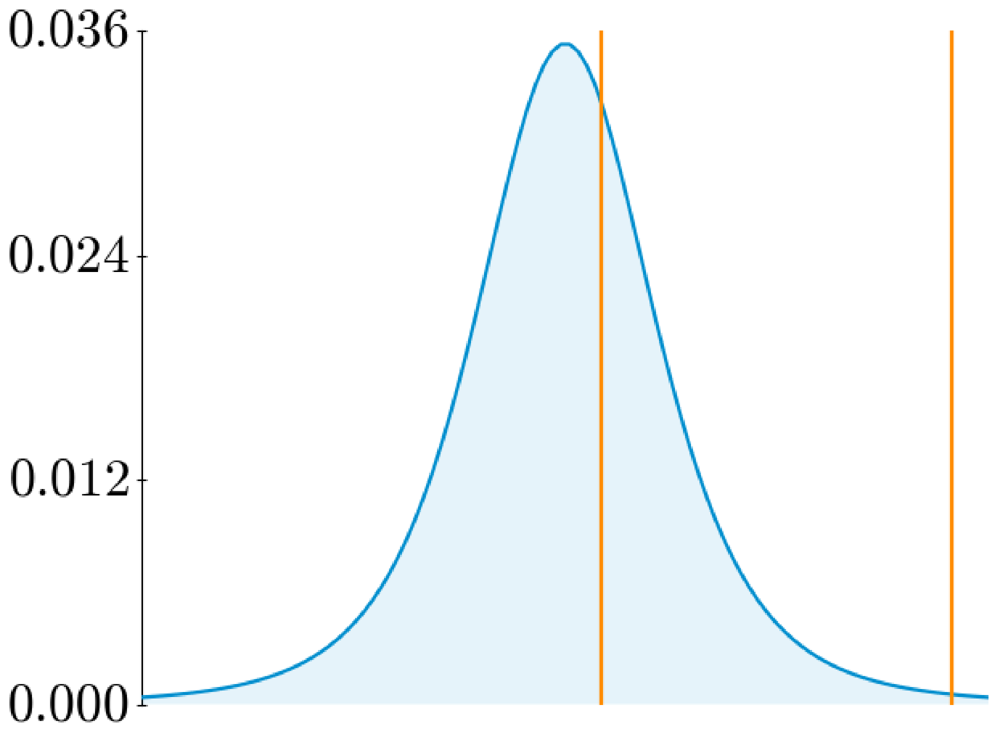
Probabilidad a posteriori

Evidencias

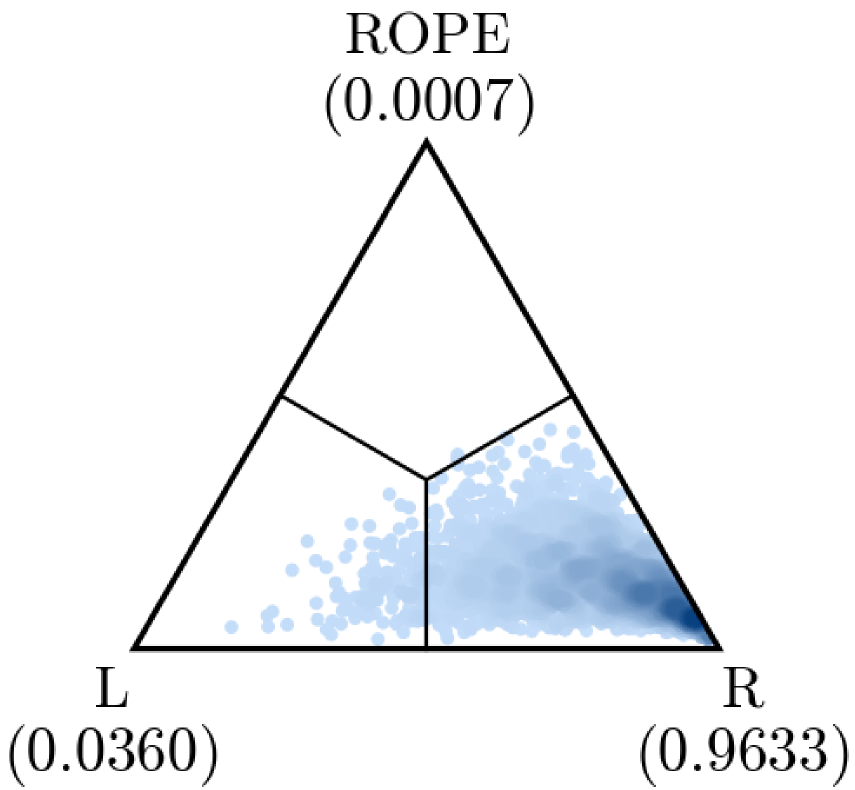
Comparación bayesiana de clasificadores

- Probabilidad de que mi clasificador sea mejor que el tuyo (L)
- Probabilidad de que tu clasificador sea mejor que el mío (R)
- Probabilidad de que ambos clasificadores rindan igual (ROPE)
  - Region Of Practical Equivalence

Comparación bayesiana de clasificadores



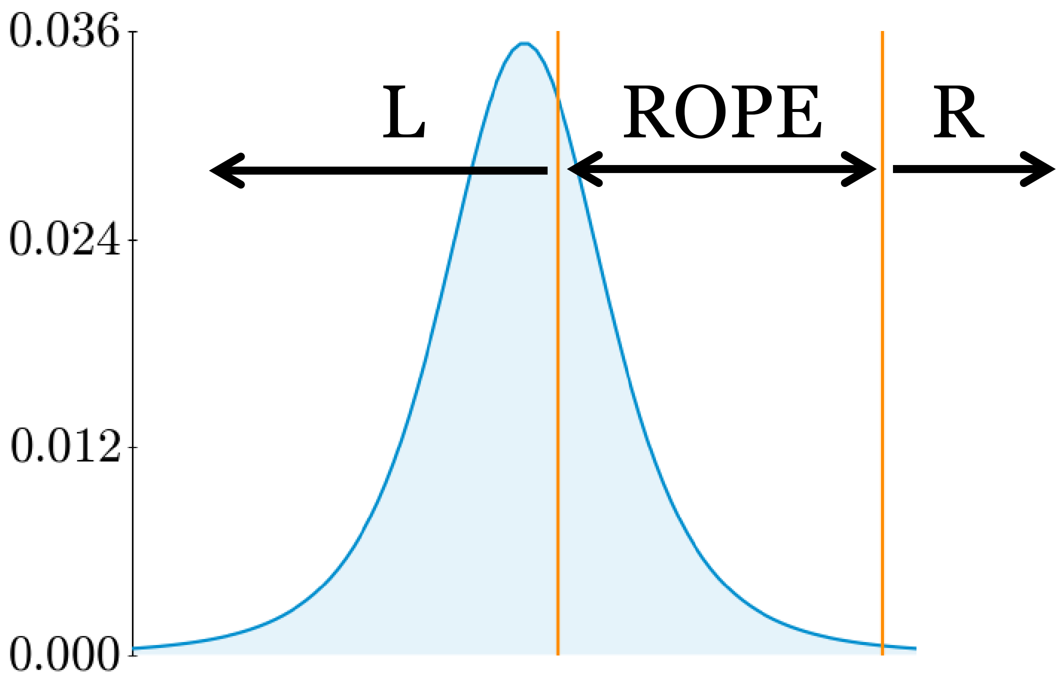
Un conjunto de datos



Varios conjuntos de datos

Un conjunto de datos

- Correlated t-test
- Necesario utilizar validación cruzada
- Distribución se calcula analíticamente



UNIVERSIDAD DE BURGOS
Junta de Castilla y León
Consejería de Educación
UNIÓN EUROPEA
Fondo Social Europeo
El FSE invierte en tu futuro

Varios conjuntos de datos

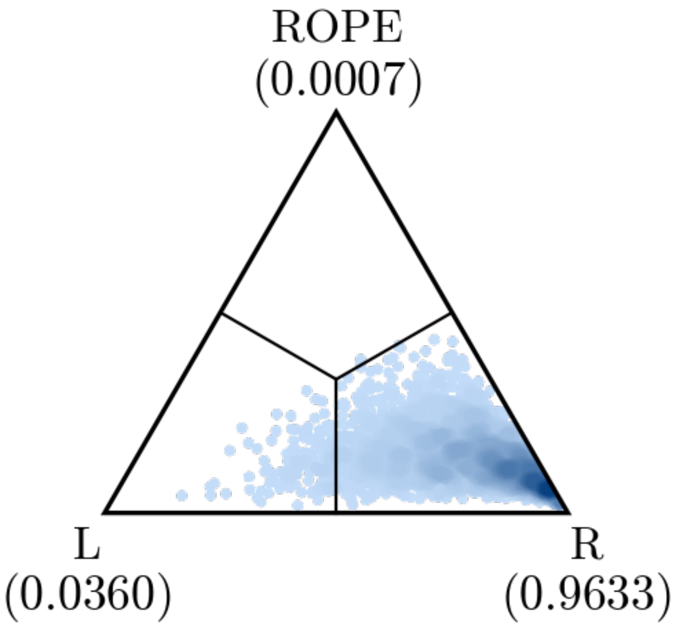
- Sign Test
- Signed-Rank Test

Una medida por conjunto de datos
Versión bayesiana del test de Wilcoxon

- Hierarchical Test → Todas las medidas de todos los folds

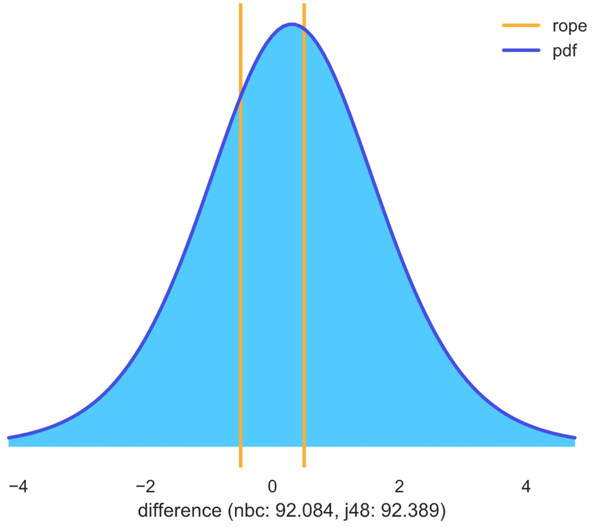
Varios conjuntos de datos

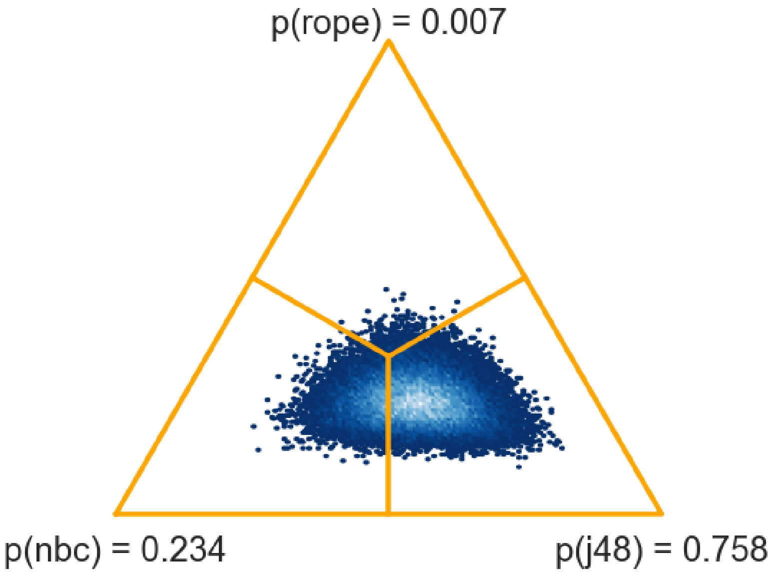
- Coste del cálculo analítico de la distribución muy elevado
- Se aproxima la distribución utilizando Monte Carlo



¿Cómo hago estas comparaciones?

- Baycomp: Librería en Python
  - https://baycomp.readthedocs.io

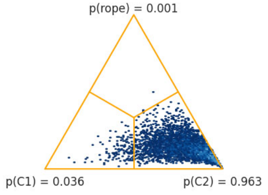




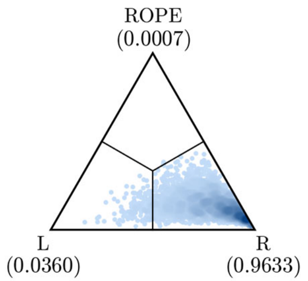
¿Cómo hago estas comparaciones?

- Baycomp_plotting: Visualización alternativa
  - [https://github.com/**mjuez**/baycomp_plotting](https://github.com/mjuez/baycomp_plotting)

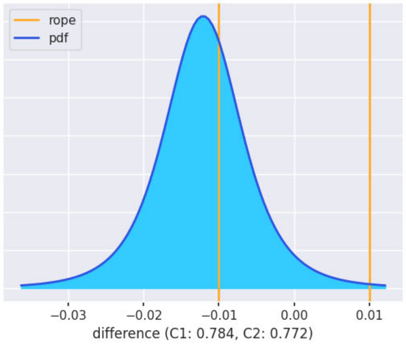
baycomp default plotting:



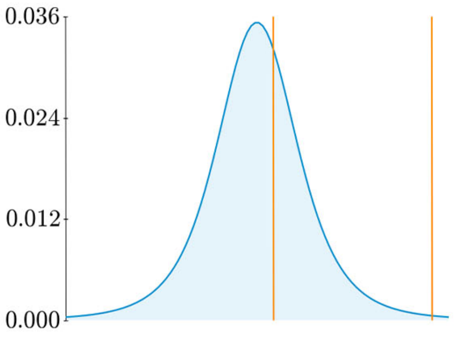
baycomp_plotting:



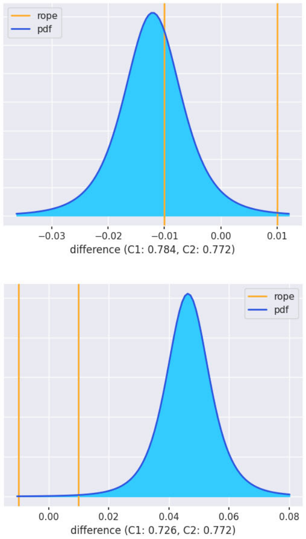
baycomp default plotting:



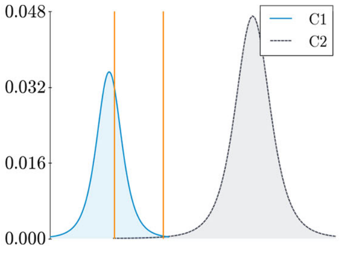
baycomp_plotting:



baycomp default plotting:



baycomp_plotting:



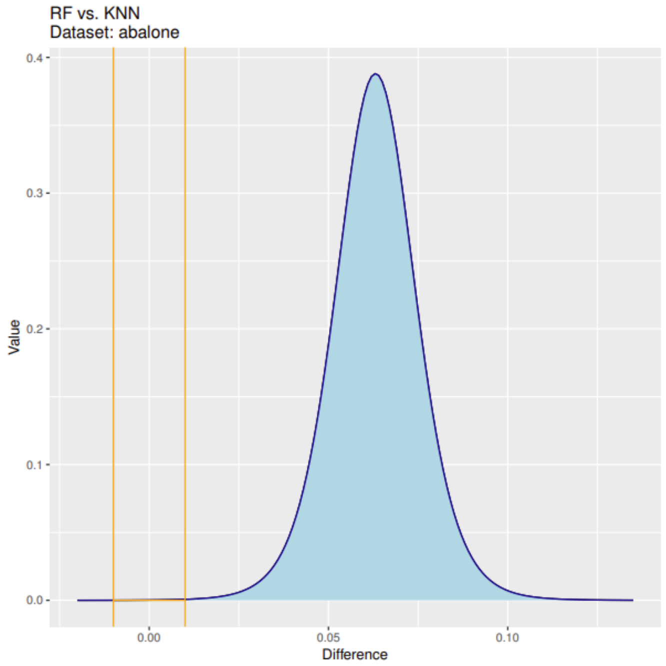
UNIVERSIDAD DE BURGOS

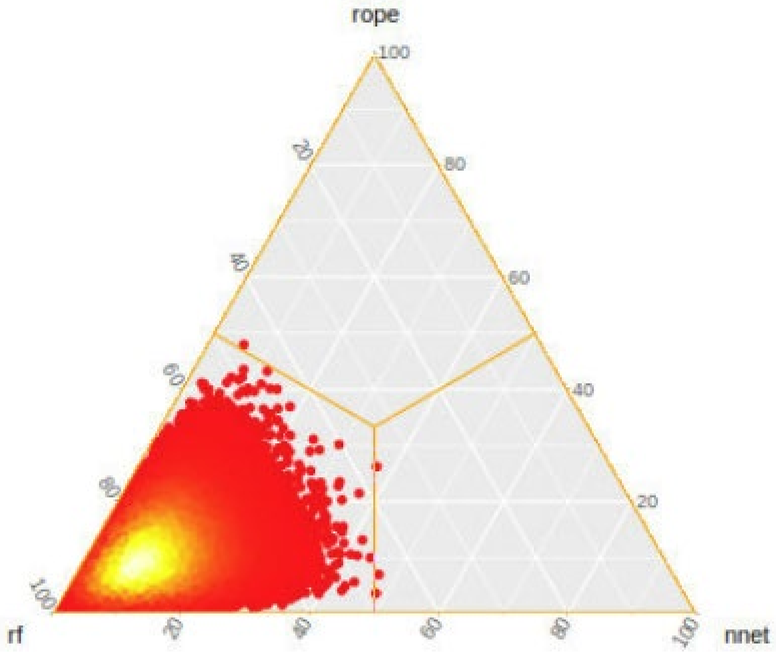
Junta de Castilla y León
Consejería de Educación

UNIÓN EUROPEA
Fondo Social Europeo
El FSE invierte en tu futuro

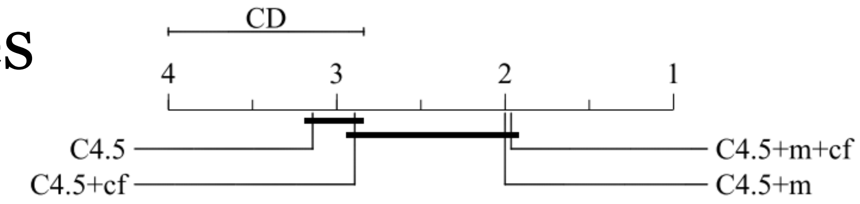
¿Cómo hago estas comparaciones?

- rNPBST: Librería en R
  - https://jacintocc.github.io/rNPBST





¿Hacia dónde vamos?



- Comparaciones de más de 2 clasificadores
  - Contraste de hipótesis: ANOVA, Friedmann
  - No existe alternativa bayesiana (aún)
- Resultados condicionados por los conjuntos de datos
  - Teorema: No Free Lunch
  - Mi clasificador es mejor que el tuyo… frente a estos datos concretos
- Alternativas al uso de métricas de rendimiento
  - Métodos gráficos

Referencias

- Demšar, J. (2006). **Statistical comparisons of classifiers over multiple data sets.** *The Journal of Machine Learning Research*, 7, 1-30.
- Demšar, J. (2008). **On the appropriateness of statistical tests in machine learning.** In *Workshop on Evaluation Methods for Machine Learning in conjunction with ICML* (p. 65).
- Benavoli, A., Corani, G., Demšar, J., & Zaffalon, M. (2017). **Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis**. *The Journal of Machine Learning Research*, *18*(1), 2653-2688.
- Carrasco, J., Garcıa, S., & Herrera, F. **shinytests: Una herramienta gráfica para la comparacion estadıstica en mineria de datos.**
- Stapor, K., Ksieniewicz, P., García, S., & Woźniak, M. (2021). **How to design the fair experimental classifier evaluation.** *Applied Soft Computing*, *104*, 107219.