

AN INTRODUCTION TO CONFORMAL PREDICTION

Henrik Linusson

June 13, 2017

Dept. of Information Technology, University of Borås, Sweden
henrik.linusson@hb.se

```
%> whoami
henrik_linusson

%> pwd
/sweden/universities/borås

%> groups
msc phdstudent csl@bs copa

%> git clone http://github.com/donlnz/nonconformist
Cloning into 'nonconformist'

%> mail -s "COPA tutorial confusion" henrik.linusson@hb.se
Thanks for the conformal prediction tutorial at COPA!
I have some questions...
```

A motivating example

Conformal prediction at a glance

Conformal classification

Conformal regression

Validity and efficiency

Considerations and modifications

Other scenarios, topics, suggested reading

References

A MOTIVATING EXAMPLE

How good is your prediction?

You want to estimate the risk of cancer recurrence in patient x_{k+1}

To your disposal, you have:

1. A set of historical observations $(x_1, y_1), \dots, (x_k, y_k)$
 - x_i describes a patient by age, tumor size, etc
 - y_i is a measurement of cancer recurrence in patient x_i
2. Some machine learning (classification or regression) algorithm

```
import pandas as pd

breast_cancer = pd.read_csv('./data/breast-cancer.csv')

# (x_1, y_1), ..., (x_k, y_k)
x_train = breast_cancer.values[:-1, :-1]
y_train = breast_cancer.values[:-1, -1]

# (x_{k+1}, y_{k+1})
x_test = breast_cancer.values[-1, :-1]
y_test = breast_cancer.values[-1, -1]
```

```
from sklearn.neighbors import KNeighborsClassifier

knn = KNeighborsClassifier(n_neighbors=5)
knn.fit(x_train, y_train)

print(knn.predict(x_test))
print(knn.predict_proba(x_test))
```

```
['no-recurrence-events']
[[ 0.8  0.2 ]]
```

How good is your prediction, really?

- Your classifier the patient will have no recurrence events.
Is it right?
- Your probability estimator says it's 80% likely that the patient won't have a recurrence event.
How good is the estimate?
- Your regression model says the patient should have 0.4 recurrence events in the future.
How close is that to the true value?

Will you trust your model?

The simple answer:

We expect past performance to indicate future performance.

The simple answer:

We expect past performance to indicate future performance.

- The model is 71% accurate on the test data,
so we assume it's accurate for 71% of production data.
- The model has an AUC of 0.65 on the test data,
so we assume it has an AUC of 0.65 on production data.
- The model has an RMSE of 0.8 on the test data,
so we assume it has an RMSE of 0.8 on production data.

The simple answer:

We expect past performance to indicate future performance.

- The model is 71% accurate on the test data, so we assume it's accurate for 71% of production data.
- The model has an AUC of 0.65 on the test data, so we assume it has an AUC of 0.65 on production data.
- The model has an RMSE of 0.8 on the test data, so we assume it has an RMSE of 0.8 on production data.

But...

How good are these estimates? Do we have any guarantees? Specifically, what about patient x_{k+1} ? What performance should we expect from the model for this particular instance?

Conformal Prediction

¹V. Vovk, A. Gammerman, and G. Shafer, Algorithmic learning in a random world. Springer, 2005

Conformal Prediction

- Provides error bounds on a per-instance basis (unlike PAC theory).
 - Probabilities are well-calibrated (80% means 80%).

¹V. Vovk, A. Gammerman, and G. Shafer, Algorithmic learning in a random world. Springer, 2005

Conformal Prediction

- Provides error bounds on a per-instance basis (unlike PAC theory).
 - Probabilities are well-calibrated (80% means 80%).
- No need to know prior probabilities (unlike Bayesian learning).
 - Only requires that data is exchangeable (i.i.d. \rightarrow exchangeability).

¹V. Vovk, A. Gammerman, and G. Shafer, Algorithmic learning in a random world. Springer, 2005

Conformal Prediction

- Provides error bounds on a per-instance basis (unlike PAC theory).
 - Probabilities are well-calibrated (80% means 80%).
- No need to know prior probabilities (unlike Bayesian learning).
 - Only requires that data is exchangeable (i.i.d. \rightarrow exchangeability).
- Can be used with any machine learning algorithm.

¹V. Vovk, A. Gammerman, and G. Shafer, Algorithmic learning in a random world. Springer, 2005

Conformal Prediction

- Provides error bounds on a per-instance basis (unlike PAC theory).
 - Probabilities are well-calibrated (80% means 80%).
- No need to know prior probabilities (unlike Bayesian learning).
 - Only requires that data is exchangeable (i.i.d. \rightarrow exchangeability).
- Can be used with any machine learning algorithm.
- Can be applied online, offline or semi-offline.

¹V. Vovk, A. Gammerman, and G. Shafer, Algorithmic learning in a random world. Springer, 2005

Conformal Prediction

- Provides error bounds on a per-instance basis (unlike PAC theory).
 - Probabilities are well-calibrated (80% means 80%).
- No need to know prior probabilities (unlike Bayesian learning).
 - Only requires that data is exchangeable (i.i.d. \rightarrow exchangeability).
- Can be used with any machine learning algorithm.
- Can be applied online, offline or semi-offline.
- The framework is rigorously proven, and simple to implement.
 - Developed by Vladimir Vovk, Alex Gammerman & Glenn Shafer.¹

¹V. Vovk, A. Gammerman, and G. Shafer, Algorithmic learning in a random world. Springer, 2005

CONFORMAL PREDICTION AT A GLANCE

Some intuition

Assume we have

- Some distribution $\mathbf{Z} : \mathbf{X} \times \mathbf{Y}$ generating examples
- Some function $f(z) \rightarrow \mathbb{R}$

Some intuition

- Apply $f(z)$ to some, say 4, examples from Z
- Call the resulting scores $\alpha_1, \alpha_2, \alpha_3, \alpha_4$.
 - For simplicity, $\alpha_1 \leq \alpha_2 \leq \alpha_3 \leq \alpha_4$

α_1 α_2 α_3 α_4

Some intuition

If we draw new examples from Z , and apply $f(z)$ to them

- Given that all examples are exchangeable,
- we can estimate distribution of scores, relative to $\alpha_1, \dots, \alpha_4$

Some intuition

If we draw new examples from Z , and apply $f(z)$ to them

- Given that all examples are exchangeable,
- we can estimate distribution of scores, relative to $\alpha_1, \dots, \alpha_4$

20% 20% 20% 20% 20%

α_1 α_2 α_3 α_4

$$P[f(z) \leq \alpha_3] = 0.6$$

$$P[f(z) \leq \alpha_4] = 0.8$$

For any distribution Q

Let $X_1, \dots, X_k \sim Q$, where $X_i \leq X_{i+1}$

Let $W \sim Q$

$$P[W \leq X_i] = \frac{i}{k+1}$$

Equal-depth binning made simple!

Some intuition

Let $f(z_i) = |y_i - h(x_i)|$

where h is a regression model trained on the domain of \mathbf{Z} .

Some intuition

Let $f(z_i) = |y_i - h(x_i)|$

where h is a regression model trained on the domain of Z .

20%

α_1

20%

α_2

20%

α_3

20%

α_4

20%

$$P[|y_i - h(x_i)| \leq \alpha_3] = 0.6$$

$$P[|y_i - h(x_i)| \leq \alpha_4] = 0.8$$

Some intuition

We know (x_i, y_i) for all examples that generated $\alpha_1, \dots, \alpha_4$,
i.e., we can obtain values for $\alpha_1, \dots, \alpha_4$.

20%	20%	20%	20%	20%
0.03	0.07	0.11	0.13	

$$P[|y_i - h(x_i)| \leq 0.11] = 0.6$$

$$P[|y_i - h(x_i)| \leq 0.13] = 0.8$$

Some intuition

For a novel example, where we know x_i but not y_i , we still know that

$$P[|y_i - h(x_i)| \leq 0.11] = 0.6$$

$$P[|y_i - h(x_i)| \leq 0.13] = 0.8$$

and can obtain $h(x_i)$ from our regression model, e.g. $h(x_i) = 0.3$.

20%

0.03

20%

0.07

20%

0.11

20%

0.13

20%

Some intuition

For a novel example, where we know x_i but not y_i , we still know that

$$P[|y_i - h(x_i)| \leq 0.11] = 0.6$$

$$P[|y_i - h(x_i)| \leq 0.13] = 0.8$$

and can obtain $h(x_i)$ from our regression model, e.g. $h(x_i) = 0.3$.

20%

20%

20%

20%

20%

0.03

0.07

0.11

0.13

$$P[|y_i - 0.3| \leq 0.11] = 0.6$$

$$P[|y_i - 0.3| \leq 0.13] = 0.8$$

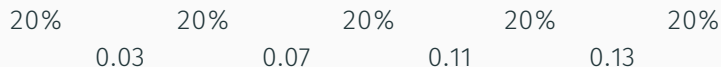
Some intuition

For a novel example, where we know x_i but not y_i , we still know that

$$P[|y_i - h(x_i)| \leq 0.11] = 0.6$$

$$P[|y_i - h(x_i)| \leq 0.13] = 0.8$$

and can obtain $h(x_i)$ from our regression model, e.g. $h(x_i) = 0.3$.



$$P[|y_i - 0.3| \leq 0.11] = 0.6$$

$$P[|y_i - 0.3| \leq 0.13] = 0.8$$

$$P[y_i \in 0.3 \pm 0.11] = 0.6$$

$$P[y_i \in 0.3 \pm 0.13] = 0.8$$

When does conformal prediction work?

We already noted a few things:

- Training data and test data belong to the same distribution (they are identically distributed)
- Choice of $f(z)$ is irrelevant (w.r.t. validity), as long as it is symmetric (training patterns and test patterns are treated equally)

What else might we need?

- statistical independence
- exchangeability (order of observations is irrelevant)

Identically, independently and exchangeably distributed sampling (*iid*)

- Draw random numbers (with replacement) according to $\mathbb{Z} \sim U[0, 3]$
- $P\{1, 2, 3\} = P\{2, 1, 3\} = P\{1, 1, 1\}$

Identically, non-independently and exchangeably distributed sampling

- Draw random numbers (without replacement) according to $\mathbb{Z} \sim U[0, 3]$
- $P\{1, 2, 3\} = P\{2, 1, 3\} \neq P\{1, 1, 1\}$

Identically, non-independently and non-exchangeably distributed sampling

- Draw random numbers (without replacement) according to $\mathbb{Z} \sim U[0, 3]$, but skip any number smaller than its predecessor
- $P\{1, 2, 3\} \neq P\{2, 1, 3\}$

Conformal predictors output multi-valued **prediction regions**

- Sets of labels or real-valued intervals

Given

- a test pattern x_i , and
- a significance level ϵ

A conformal predictor outputs

- A prediction region Γ_i^ϵ that contains y_i with probability $1 - \epsilon$

$$Y_c = \{\text{iris_setosa}, \text{iris_versicolor}, \text{iris_virginica}\}$$

$$Y_r = \mathbb{R}$$

Point predictions

$$h_c(x_{k+1}) = \text{iris_setosa}$$

$$h_c(x_{k+2}) = \text{iris_versicolor}$$

$$h_c(x_{k+3}) = \text{iris_virginica}$$

$$h_r(x_{k+1}) = 0.3$$

$$h_r(x_{k+2}) = 0.2$$

$$h_r(x_{k+3}) = 0.6$$

Point predictions

$$h_c(x_{k+1}) = \text{iris_setosa}$$

$$h_c(x_{k+2}) = \text{iris_versicolor}$$

$$h_c(x_{k+3}) = \text{iris_virginica}$$

$$h_r(x_{k+1}) = 0.3$$

$$h_r(x_{k+2}) = 0.2$$

$$h_r(x_{k+3}) = 0.6$$

$$P[y_i = h_c(x_i)] = ?$$

$$\Delta[y_i, h_r(x_i)] = ?$$

Prediction regions

$$h_c(x_{k+1}) = \{\text{iris_setosa}\}$$

$$h_c(x_{k+2}) = \{\text{iris_setosa}, \text{iris_versicolor}\}$$

$$h_c(x_{k+3}) = \{\text{iris_setosa}, \text{iris_versicolor}, \text{iris_virginica}\}$$

$$h_r(x_{k+1}) = [0.2, 0.4]$$

$$h_r(x_{k+2}) = [0, 0.5]$$

$$h_r(x_{k+3}) = [0.5, 0.7]$$

Prediction regions

$$h_c(x_{k+1}) = \{\text{iris_setosa}\}$$

$$h_c(x_{k+2}) = \{\text{iris_setosa}, \text{iris_versicolor}\}$$

$$h_c(x_{k+3}) = \{\text{iris_setosa}, \text{iris_versicolor}, \text{iris_virginica}\}$$

$$h_r(x_{k+1}) = [0.2, 0.4]$$

$$h_r(x_{k+2}) = [0, 0.5]$$

$$h_r(x_{k+3}) = [0.5, 0.7]$$

$$P[y_i \in h_c(x_i)] = 1 - \epsilon$$

$$P[y_i \in h_r(x_i)] = 1 - \epsilon$$

To perform conformal prediction, we need

- A function $f(z) \rightarrow \mathbb{R}$
- A set of training examples, $Z^k \subset Z : X^n \times Y$
- A statistical test

Overall rationale

1. Apply $f(z)$ to training examples in Z^k , estimate distribution of $f(z) \sim Q$
2. For every possible output $\tilde{y} \in Y$, apply $f(z)$ to (x_{k+1}, \tilde{y})
3. Reject \tilde{y} if it appears unlikely that $f[(x_{k+1}, \tilde{y})] \sim Q$

The function $f(z)$

We call this the **nonconformity function**

- A function that measures the “strangeness” of a pattern (x_i, y_i)
- Any function $f(z) \rightarrow \mathbb{R}$ works (produces valid predictions)

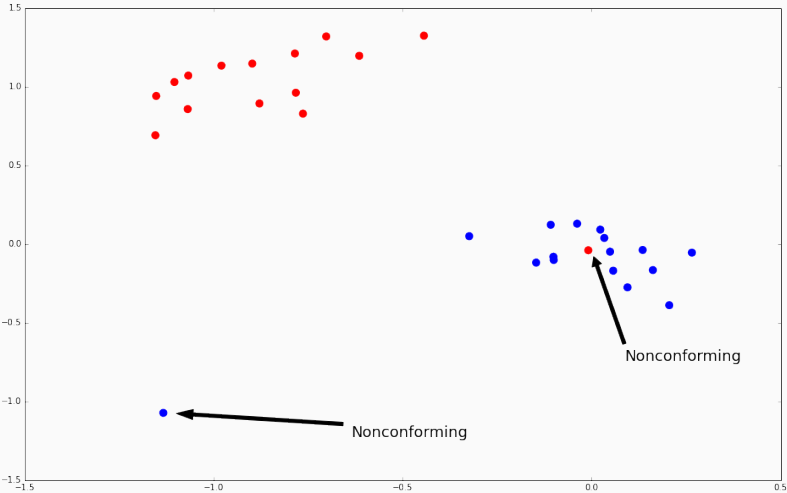
Properties of a good nonconformity function (that produces small prediction sets)

- Give low scores to patterns (x_i, y_i)
- Give large scores to patterns $(x_i, \neg y_i)$

Common choice: $f(z) = \Delta[h(x_i), y_i]$

- h is called the **underlying model**
- “Our random forest misclassified this example, it must be weird!”

CONFORMAL PREDICTION AT A GLANCE



Probability estimate for correct class

If the probability estimate for an example's correct class is low, the example is strange.

Margin of a probability estimating model

If an example's true class is not clearly separable from other classes, it is strange.

Distance to neighbors with same class (or distance to neighbors with different classes)

If an example is not surrounded by examples that share its label, it is strange.

Absolute error of a regression model

If the prediction is far from the true value, the example is strange.

`rand(0, 1)`

Even if it's not useful, it's still valid.

Conformal prediction process

1. Define a nonconformity function.
2. Measure the nonconformity of labeled examples $(x_1, y_1), \dots, (x_k, y_k)$.
3. For a new pattern x_i , test all possible outputs $\tilde{y} \in Y$:
 - 3.1 Measure the nonconformity of (x_i, \tilde{y}) .
 - 3.2 Is (x_i, \tilde{y}) particularly nonconforming compared to the training examples? Then \tilde{y} is probably an incorrect classification. Otherwise, include it in the prediction region.

To determine whether an example is “too nonconforming”, we use a statistical test.

To determine whether an example is “too nonconforming”, we use a statistical test.

$$p_i^{\tilde{y}} = \frac{|\{z_j \in Z : \alpha_j > \alpha_i^{\tilde{y}}\}|}{k+1} + \theta \frac{|\{z_j \in Z : \alpha_j = \alpha_i^{\tilde{y}}\}| + 1}{k+1}, \theta \sim U[0, 1]$$

(Portion of examples at least as nonconforming as the tentatively labeled test example)

To determine whether an example is “too nonconforming”, we use a statistical test.

$$p_i^{\tilde{y}} = \frac{|\{z_j \in Z : \alpha_j > \alpha_i^{\tilde{y}}\}|}{k+1} + \theta \frac{|\{z_j \in Z : \alpha_j = \alpha_i^{\tilde{y}}\}| + 1}{k+1}, \theta \sim U[0, 1]$$

(Portion of examples at least as nonconforming as the tentatively labeled test example)

Prediction region

$$\Gamma_i^\epsilon = \{\tilde{y} \in Y : p_i^{\tilde{y}} > \epsilon\}$$

To determine whether an example is “too nonconforming”, we use a statistical test.

$$p_i^{\tilde{y}} = \frac{|\{z_j \in Z : \alpha_j > \alpha_i^{\tilde{y}}\}|}{k+1} + \theta \frac{|\{z_j \in Z : \alpha_j = \alpha_i^{\tilde{y}}\}| + 1}{k+1}, \theta \sim U[0, 1]$$

(Portion of examples at least as nonconforming as the tentatively labeled test example)

Prediction region

$$\Gamma_i^\epsilon = \{\tilde{y} \in Y : p_i^{\tilde{y}} > \epsilon\}$$

- Classification — known $\alpha_i^{\tilde{y}}$, find $p_i^{\tilde{y}}$
- Regression — known $p_i^{\tilde{y}}$, find $\alpha_i^{\tilde{y}}$

Transductive conformal prediction (TCP) — $f(z, Z)$

Original conformal prediction approach

- Requires retraining model for each new test example
- For regression problems, only certain models (e.g. kNN) can be used as of yet

Inductive conformal prediction (ICP) — $f(z)$

Revised approach

- Requires model to be trained only once
- Requires that some data is set aside for calibration
 - To avoid violating exchangeability assumption

CONFORMAL CLASSIFICATION

Divide the training set Z into two disjoint subsets

A proper training set Z_t

A calibration set Z_c where $|Z_c| = q$

Divide the training set Z into two disjoint subsets

A proper training set Z_t

A calibration set Z_c where $|Z_c| = q$

Fit a model h using Z_t

This is the underlying model

Divide the training set Z into two disjoint subsets

A **proper training set** Z_t

A **calibration set** Z_c where $|Z_c| = q$

Fit a model h using Z_t

This is the **underlying model**

Choose an $f(z)$, e.g. $f(z_i) = 1 - \hat{P}_h(y_i | x_i)$

This is the **nonconformity function**

Divide the training set Z into two disjoint subsets

A **proper training set** Z_t

A **calibration set** Z_c where $|Z_c| = q$

Fit a model h using Z_t

This is the **underlying model**

Choose an $f(z)$, e.g. $f(z_i) = 1 - \hat{P}_h(y_i | x_i)$

This is the **nonconformity function**

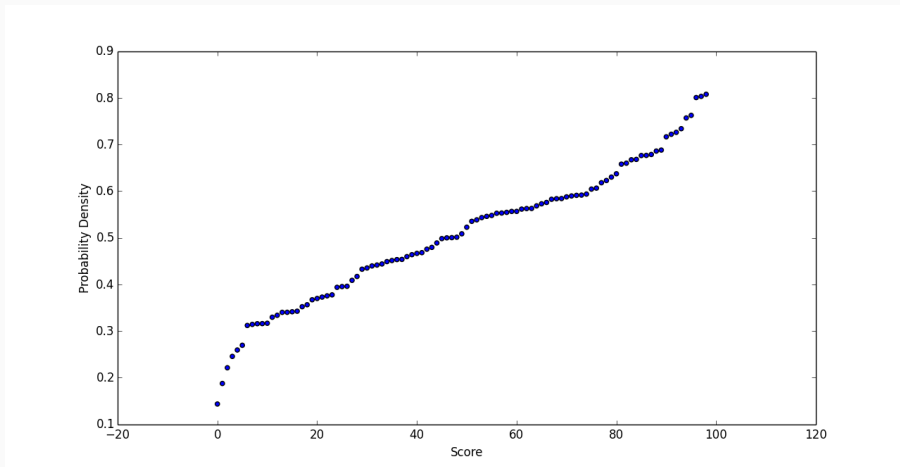
Apply $f(Z)$ to $\forall z_i \in Z_c$

Save these **calibration scores**

We denote these $\alpha_1, \dots, \alpha_q$

INDUCTIVE CONFORMAL CLASSIFICATION

Apply $f(z)$ to Z_c , and obtain a set of calibration scores $\alpha_1, \dots, \alpha_q$



For each $\tilde{y} \in Y$

Let $\alpha_i^{\tilde{y}} = f[(x_i, \tilde{y})]$

Calculate

$$p_i^{\tilde{y}} = \frac{|\{z_j \in Z_c : \alpha_j > \alpha_i^{\tilde{y}}\}|}{q+1} + \theta \frac{|\{z_j \in Z_c : \alpha_j = \alpha_i^{\tilde{y}}\}| + 1}{q+1}, \theta \sim U[0, 1]$$

For each $\tilde{y} \in Y$

Let $\alpha_i^{\tilde{y}} = f[(x_i, \tilde{y})]$

Calculate

$$p_i^{\tilde{y}} = \frac{|\{z_j \in Z_c : \alpha_j > \alpha_i^{\tilde{y}}\}|}{q+1} + \theta \frac{|\{z_j \in Z_c : \alpha_j = \alpha_i^{\tilde{y}}\}| + 1}{q+1}, \theta \sim U[0, 1]$$

Fix a significance level $\epsilon \in (0, 1)$

For each $\tilde{y} \in Y$

Let $\alpha_i^{\tilde{y}} = f[(x_i, \tilde{y})]$

Calculate

$$p_i^{\tilde{y}} = \frac{|\{z_j \in Z_c : \alpha_j > \alpha_i^{\tilde{y}}\}|}{q+1} + \theta \frac{|\{z_j \in Z_c : \alpha_j = \alpha_i^{\tilde{y}}\}| + 1}{q+1}, \theta \sim U[0, 1]$$

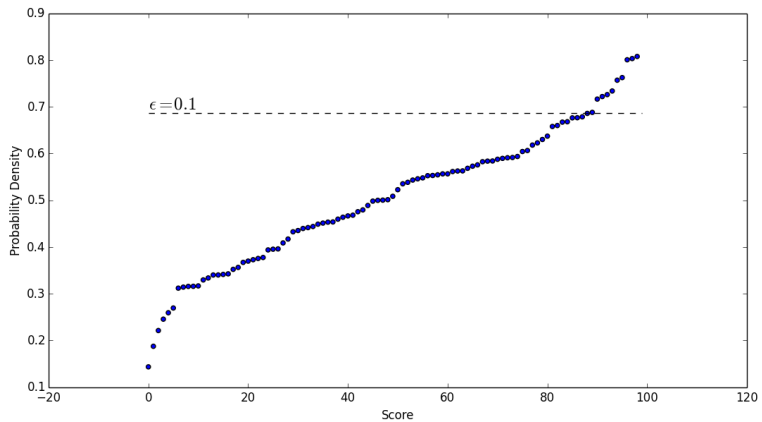
Fix a significance level $\epsilon \in (0, 1)$

Prediction region

$$\Gamma_i^\epsilon = \{\tilde{y} \in Y : p_i^{\tilde{y}} > \epsilon\}$$

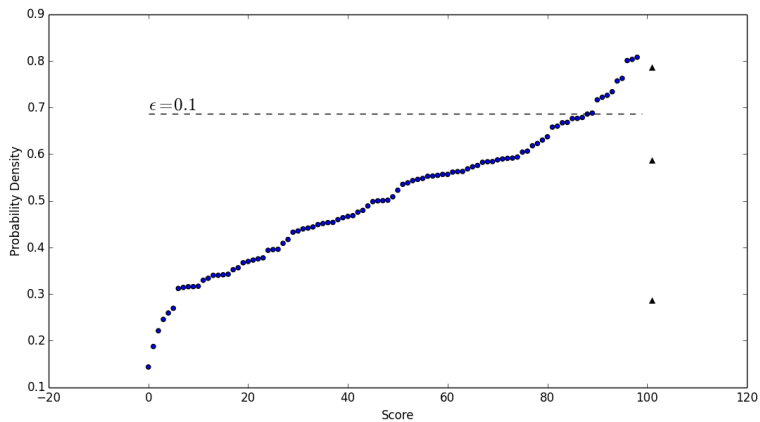
INDUCTIVE CONFORMAL CLASSIFICATION

Choose a significance level ϵ



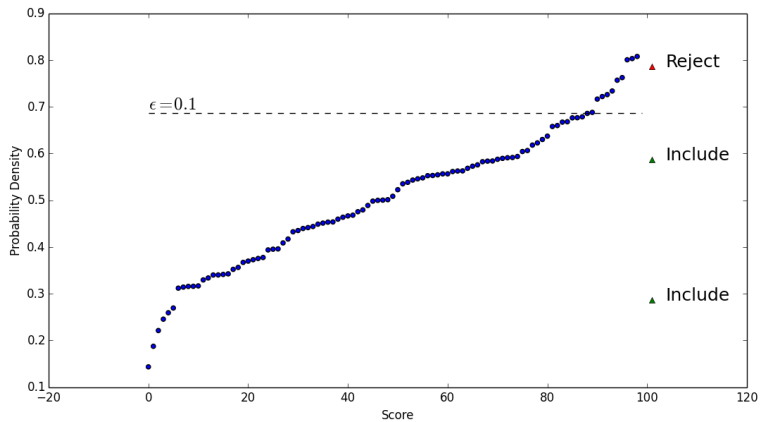
INDUCTIVE CONFORMAL CLASSIFICATION

Obtain α_i using $f(z)$ for each possible class $(x_i, \tilde{y}_1), (x_i, \tilde{y}_2), (x_i, \tilde{y}_3), \dots$, resulting in $\alpha_i^{\tilde{y}_1}, \alpha_i^{\tilde{y}_2}, \alpha_i^{\tilde{y}_3}, \dots$



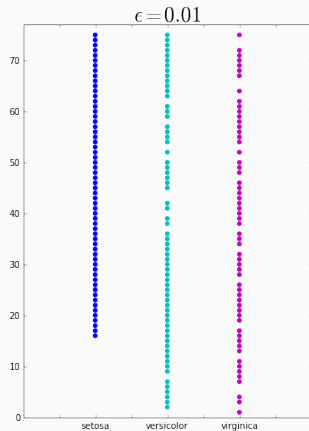
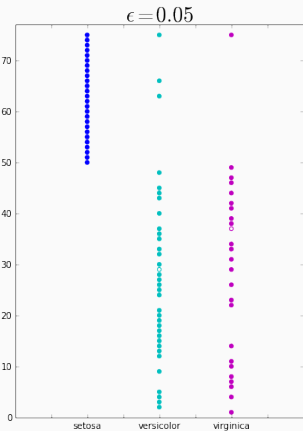
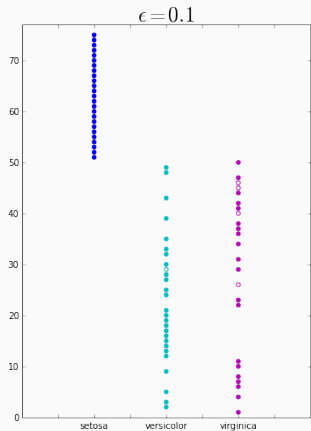
INDUCTIVE CONFORMAL CLASSIFICATION

Reject/include based on the p-value statistic, and the chosen ϵ



INDUCTIVE CONFORMAL CLASSIFICATION

Iris, Random Forest



CONFORMAL REGRESSION

Divide the training set Z into two disjoint subsets

A proper training set Z_t

A calibration set Z_c where $|Z_c| = q$

Divide the training set Z into two disjoint subsets

A proper training set Z_t

A calibration set Z_c where $|Z_c| = q$

Fit a model h using Z_t

This is the underlying model

Divide the training set Z into two disjoint subsets

A **proper training set** Z_t

A **calibration set** Z_c where $|Z_c| = q$

Fit a model h using Z_t

This is the **underlying model**

Let $f(z_i) = |y_i - h(x_i)|$

This is the **nonconformity function**

Divide the training set Z into two disjoint subsets

A **proper training set** Z_t

A **calibration set** Z_c where $|Z_c| = q$

Fit a model h using Z_t

This is the **underlying model**

Let $f(z_i) = |y_i - h(x_i)|$

This is the **nonconformity function**

Apply $f(z)$ to $\forall z_i \in Z_c$

Save these **calibration scores**, sorted in descending order

We denote these $\alpha_1, \dots, \alpha_q$

Fix a significance level $\epsilon \in (0, 1)$

Let $s = \lfloor \epsilon(q + 1) \rfloor$.

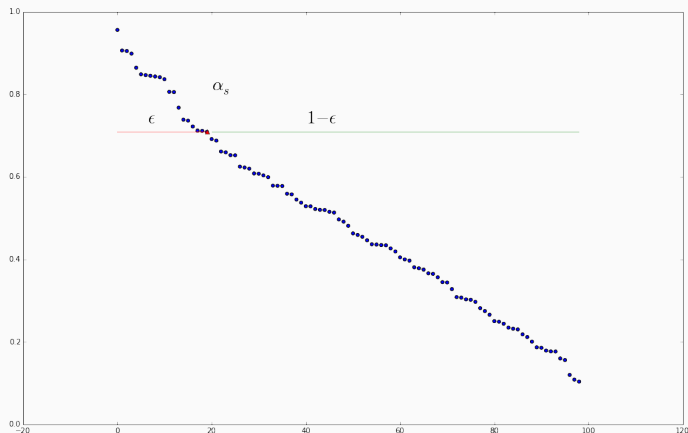
This is the index of the $(1 - \epsilon)$ -percentile nonconformity score, α_s .

INDUCTIVE CONFORMAL REGRESSION

Fix a significance level $\epsilon \in (0, 1)$

Let $s = \lfloor \epsilon(q + 1) \rfloor$.

This is the index of the $(1 - \epsilon)$ -percentile nonconformity score, α_s .



The prediction for a new example is $\Gamma_i^\epsilon = h(x_i) \pm \alpha_s$

The interval contains y_i with probability $1 - \epsilon$

The prediction for a new example is $\Gamma_i^\epsilon = h(x_i) \pm \alpha_s$

The interval contains y_i with probability $1 - \epsilon$

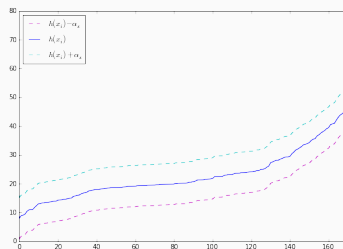
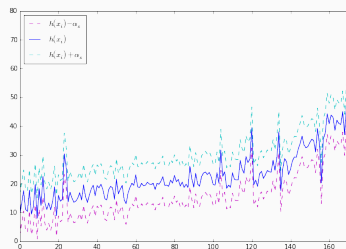
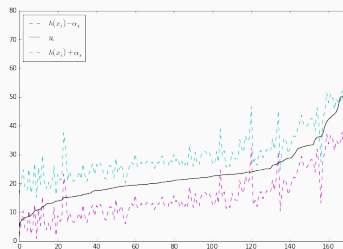
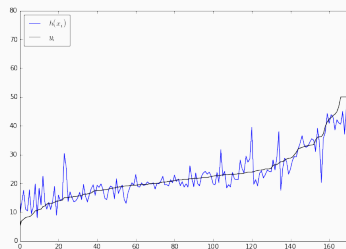
Note

For regression, we can't enumerate each $\tilde{y} \in Y$, instead we work backwards, i.e., fix the p-value and then find an appropriate $\alpha_{\tilde{y}}$.

- Hence, our nonconformity function must be (partially) invertible for quick calculation of intervals

INDUCTIVE CONFORMAL REGRESSION

Boston Housing, Random Forest, $\epsilon = 0.1$



Static prediction interval size

Using $f(z_i) = |y_i - h(x_i)|$ and $\Gamma_i^\epsilon = h(x_i) \pm \alpha_S$
means each prediction interval has the same size (α_S).

But we want individual bounds for each x_i ...

Static prediction interval size

Using $f(z_i) = |y_i - h(x_i)|$ and $\Gamma_i^\epsilon = h(x_i) \pm \alpha_s$
means each prediction interval has the same size (α_s).

But we want individual bounds for each x_i ...

Normalized nonconformity functions

Normalized nonconformity functions utilize an additional term σ_i .

$$f(z_i) = \frac{|y_i - h(x_i)|}{\sigma_i}$$

σ_i is an estimate of the difficulty of predicting y_i

A common practice is to let σ be predicted by a model, e.g., $\sigma_i = \hat{\Delta}[y_i, h(x_i)]$

Static prediction interval size

Using $f(z_i) = |y_i - h(x_i)|$ and $\Gamma_i^\epsilon = h(x_i) \pm \alpha_s$
means each prediction interval has the same size (α_s).

But we want individual bounds for each x_i ...

Normalized nonconformity functions

Normalized nonconformity functions utilize an additional term σ_i .

$$f(z_i) = \frac{|y_i - h(x_i)|}{\sigma_i}$$

σ_i is an estimate of the difficulty of predicting y_i

A common practice is to let σ be predicted by a model, e.g., $\sigma_i = \hat{\Delta}[y_i, h(x_i)]$

The normalized prediction for a new example is $\Gamma_i^\epsilon = h(x_i) \pm \alpha_s \sigma_i$

Divide the training set Z into two disjoint subsets

A **proper training set** Z_t

A **calibration set** Z_c

Fit a model h using Z_t

In addition

- Let E_t be the residual errors of h (i.e. the errors that h makes on Z_t)
- Fit a model g using $X_t \times E_t$

$$f(z_i) = \frac{|y_i - h(x_i)|}{g(x_i) + \beta}$$

β is a sensitivity parameter that determines the impact of normalization

Apply $f(z)$ to $\forall z_i \in Z_c$

Save these **calibration scores**, sorted in descending order

Fix a significance level $\epsilon \in (0, 1)$

Let $s = \lfloor \epsilon(q + 1) \rfloor$

This is the index of the $(1 - \epsilon)$ -percentile nonconformity score, α_s .

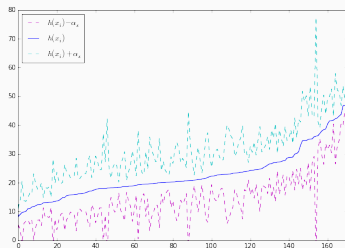
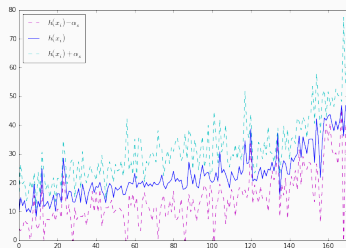
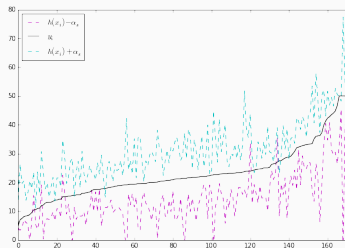
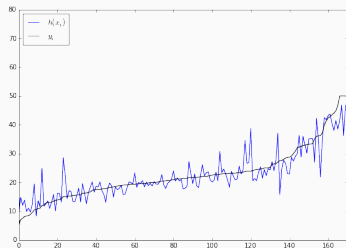
Prediction region

The prediction for a new example is $\Gamma_i^\epsilon = h(x_i) \pm \alpha_s(g(x_i) + \beta)$

Interval contains y_i with probability $1 - \epsilon$

INDUCTIVE CONFORMAL REGRESSION

Boston Housing, Random Forest, normalized nonconformity function, $\epsilon = 0.1$



How good is your prediction?

You want to estimate the risk of cancer recurrence in patient x_{k+1}

To your disposal, you have:

1. A set of historical observations $(x_1, y_1), \dots, (x_k, y_k)$
 - x_i describes a patient by age, tumor size, etc
 - y_i is a measurement of cancer recurrence in patient x_i
2. Some machine learning (classification or regression) algorithm
3. Conformal prediction

```
import pandas as pd

breast_cancer = pd.read_csv('./data/breast-cancer.csv')

# proper training set
x_train = breast_cancer.values[:-100, :-1]
y_train = breast_cancer.values[:-100, -1]

# calibration set
x_cal = breast_cancer.values[-100:-1, :-1]
y_cal = breast_cancer.values[-100:-1, -1]

# (x_{k+1}, y_{k+1})
x_test = breast_cancer.values[-1, :-1]
y_test = breast_cancer.values[-1, -1]

# Omitted: convert y_train, y_cal, y_test to numeric
```

```
import numpy as np
from sklearn.neighbors import KNeighborsClassifier
from nonconformist.icp import IcpClassifier
from nonconformist.nc import NcFactory

knn = KNeighborsClassifier(n_neighbors=5)
nc = NcFactory.create_nc(knn)
icp = IcpClassifier(nc)

icp.fit(x_train, y_train)
icp.calibrate(x_cal, y_cal)

print(icp.predict(np.array([x_test]), significance=0.05))
```

```
[[ True  False ]]
```

Installation options:

- `git clone http://github.com/donlnz/nonconformist`
- `pip install nonconformist`

Nonconformist supports:

- Conformal classification (inductive)
- Conformal regression (inductive)
- Mondrian (e.g., class-conditional) models
- Normalization
- Aggregated conformal predictors (\approx icp ensembles)
- Out-of-bag calibration
- Plug-and-play using sklearn
- User extensions

VALIDITY AND EFFICIENCY

Conformal predictors are subject to two desiderata

Validity — coherence between ϵ and error rate

Efficiency — size of prediction regions (i.e. informativeness)

Conformal predictors are automatically valid

Efficiency depends on the nonconformity function (and thus the underlying model)

Conformal predictors are subject to two desiderata

Validity — coherence between ϵ and error rate

Efficiency — size of prediction regions (i.e. informativeness)

Conformal predictors are automatically valid

Efficiency depends on the nonconformity function (and thus the underlying model)

Confidence-efficiency trade-off

The more confidence we require in a prediction, the larger it will (likely) be

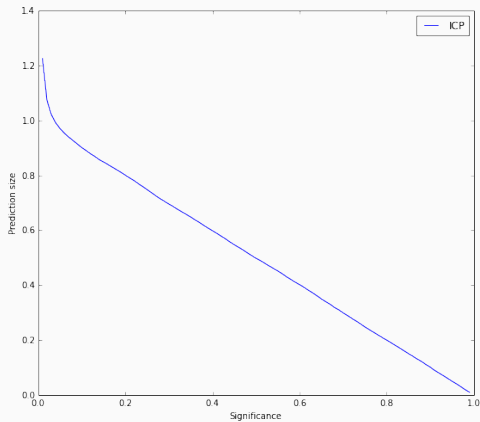
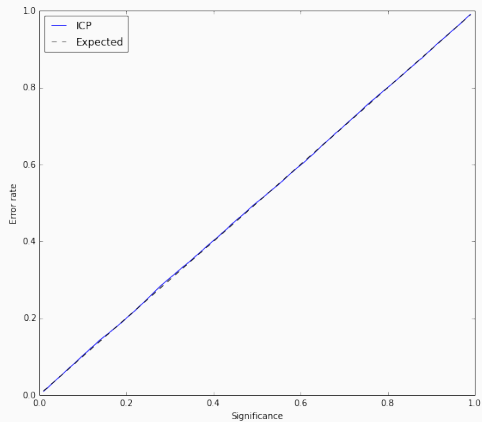
ϵ	errors	size
0.01	0.006	38.31
0.05	0.040	16.90
0.10	0.089	11.46
0.20	0.191	7.562

Table: Boston 10x10 RF CV

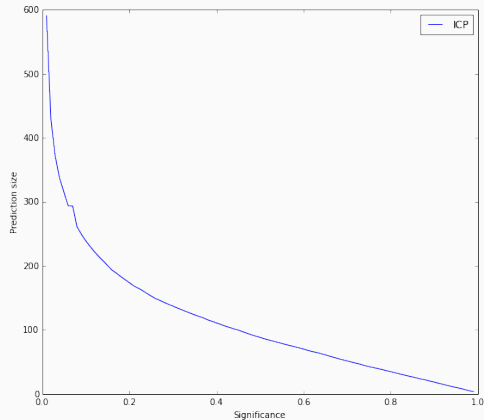
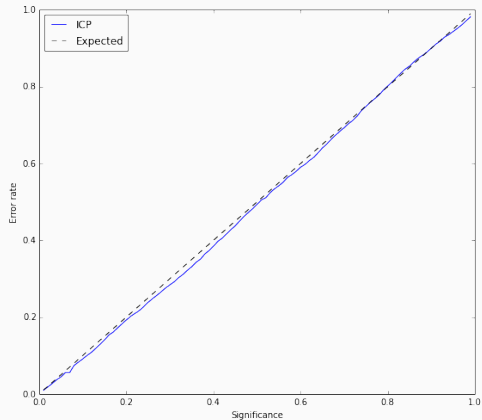
ϵ	errors	size
0.01	0.011	2.347
0.05	0.055	1.052
0.10	0.100	0.930
0.20	0.202	0.804

Table: Iris 10x10 RF CV

Digits, Random Forest, 10x10 CV



Diabetes, Random Forest, 10x10 CV



Empirical validity is measured by observing the error rate of a conformal predictor.

Efficiency can be measured in many different ways.²

Examples — regression

- Average size of prediction interval

Examples — classification

- Average number of classes per prediction (AvgC)
- Rate of predictions containing a single class (OneC)
- Average p-value

²V. Vovk, V. Fedorova, I. Nouretdinov, and A. Gammerman, “Criteria of efficiency for conformal prediction,” 2014

CONSIDERATIONS AND MODIFICATIONS

Conformal predictors are, by default, unconditional

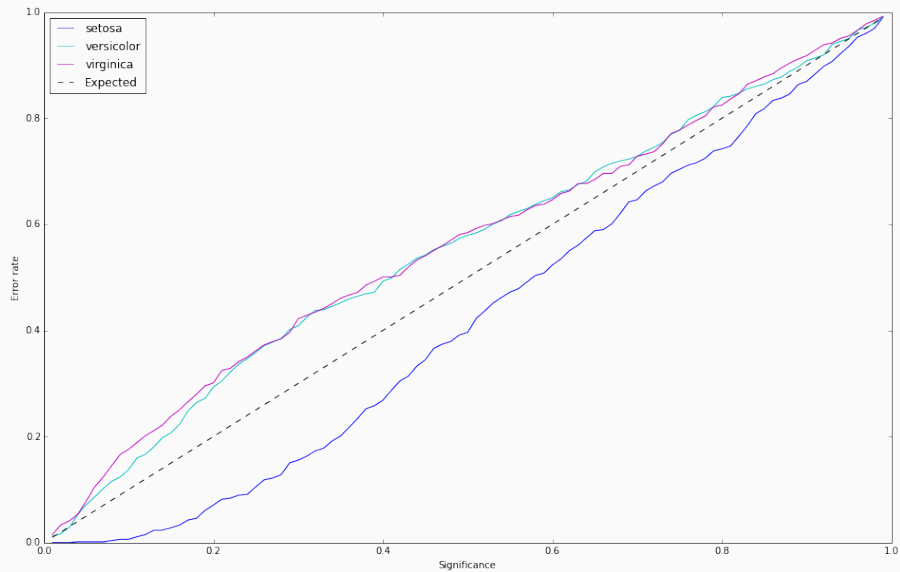
Their guaranteed error rate applies to the entire test set, on average.

- Difficult patterns (e.g. minority class) may see a greater error rate than expected
- Easy patterns (e.g. majority class) may see a smaller error rate than expected

Example — Iris data set

- One linearly separable class (easy)
- Two linearly non-separable classes (difficult)

CONDITIONAL CONFORMAL PREDICTION



Conditional conformal predictors³ help solve this by

Dividing the problem space into several disjoint subspaces

- e.g. let each class represent a subspace, or
- define subspace based on some input variable(s) (age, gender, etc.)

Guaranteeing an error rate at most ϵ for each subspace

³V. Vovk, “Conditional validity of inductive conformal predictors,” *Journal of Machine Learning Research - Proceedings Track*, vol. 25, pp. 475–490, 2012

Define a mapping function $K(z_i) = \kappa_i$

Examples

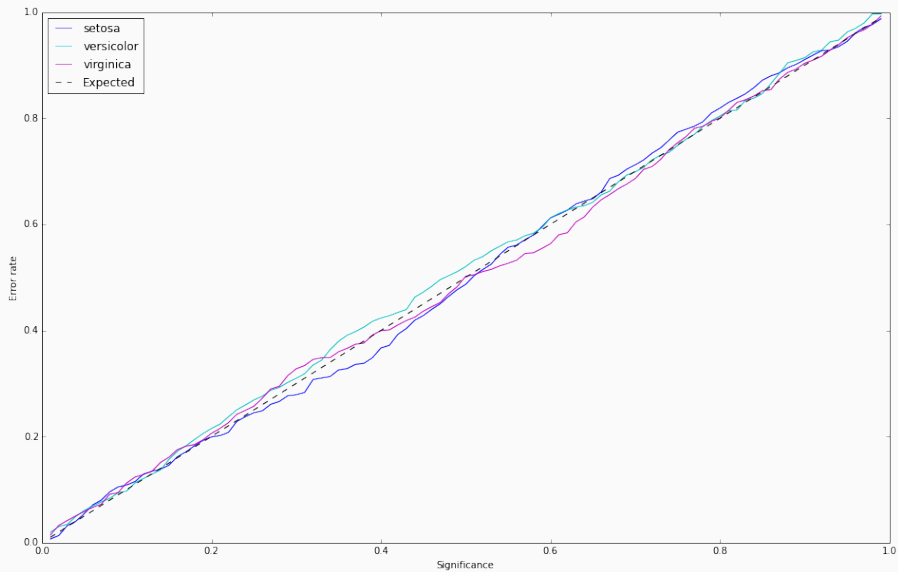
$$K(z_i) = y_i \quad (1)$$

$$K(z_i) = \begin{cases} 1 & \text{if } x_{i,1} < 50 \\ 2 & \text{if } 50 \leq x_{i,1} < 100 \\ 3 & \text{otherwise} \end{cases} \quad (2)$$

Conditional p-value

$$p_i^{\tilde{y}} = \frac{|\{z_j \in Z_c : \alpha_j > \alpha_i^{\tilde{y}}\} \wedge K(z_i) = K(z_j)|}{|K(z_i) = K(z_j)| + 1} + \theta \frac{|\{z_j \in Z_c : \alpha_j = \alpha_i^{\tilde{y}}\} \wedge K(z_i) = K(z_j)|}{|K(z_i) = K(z_j)| + 1}, \theta \sim U[0, 1]$$

CONDITIONAL CONFORMAL PREDICTION



The calibration set

Inductive conformal predictors need some data set aside for calibration? — How much?

25% ~ 33% are common choices, and provide a good balance between underlying model performance and calibration accuracy.⁴

Alternatives

Bagged ensembles can use out-of-bag examples for calibration.^{5 6}

⁴H. Linusson, U. Johansson, H. Boström, and T. Löfström, “Efficiency comparison of unstable transductive and inductive conformal classifiers,” in *Artificial Intelligence Applications and Innovations*. Springer, 2014, pp. 261–270

⁵U. Johansson, H. Boström, T. Löfström, and H. Linusson, “Regression conformal prediction with random forests,” *Machine Learning*, vol. 97, no. 1-2, pp. 155–176, 2014

⁶H. Boström, H. Linusson, T. Löfström, and U. Johansson, “Accelerating difficulty estimation for conformal regression forests,” *Annals of Mathematics and Artificial Intelligence*, pp. 1–20, 2017

The calibration set cont.

For an inductive conformal predictor to be exactly valid, it requires exactly $k\epsilon^{-1} - 1$ calibration instances.

- Otherwise, discretization errors come into play
 - (Rendering the conformal predictor conservatively valid)
- Of particular importance when calibration set is small
 - e.g. when using conditional conformal prediction

Alternatives

Interpolation of p-values can alleviate this problem.^{7 8}

⁷L. Carlsson, E. Ahlberg, H. Boström, U. Johansson, and H. Linusson, “Modifications to p-values of conformal predictors,” in *Statistical Learning and Data Sciences*. Springer, 2015, pp. 251–259

⁸U. Johansson, E. Ahlberg, H. Boström, L. Carlsson, H. Linusson, and C. Sönströd, “Handling small calibration sets in mondrian inductive conformal regressors,” in *Statistical Learning and Data Sciences*. Springer, 2015, pp. 271–280

OTHER SCENARIOS, TOPICS, SUGGESTED READING

Other scenarios for conformal prediction

- Anomaly detection with guaranteed maximum false positive rates.⁹
- Concept drift detection / i.i.d. checking with maximum false positive rates.¹⁰
- Rule extraction with guaranteed fidelity.¹¹
- Semi-supervised learning.¹²

⁹R. Laxhammar and G. Falkman, “Conformal prediction for distribution-independent anomaly detection in streaming vessel data,” in Proceedings of the First International Workshop on Novel Data Stream Pattern Mining Techniques. ACM, 2010, pp. 47–55

¹⁰V. Fedorova, A. Gammerman, I. Nourtdinov, and V. Vovk, “Plug-in martingales for testing exchangeability on-line,” in 29th International Conference on Machine Learning, 2012

¹¹U. Johansson, R. König, H. Linusson, T. Löfström, and H. Boström, “Rule extraction with guaranteed fidelity,” in Artificial Intelligence Applications and Innovations. Springer, 2014, pp. 281–290

¹²X. Zhu, F.-M. Schleif, and B. Hammer, “Semi-supervised vector quantization for proximity data,” in Proc. of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2013), Louvain-La-Neuve, Belgium, 2013, pp. 89–94

Nonconformity functions and underlying models

- H. Papadopoulos, V. Vovk, and A. Gammerman, “Regression conformal prediction with nearest neighbours,” *Journal of Artificial Intelligence Research*, vol. 40, no. 1, pp. 815–840, 2011
- U. Johansson, H. Boström, and T. Löfström, “Conformal prediction using decision trees,” in *International Conference Data Mining (ICDM)*. IEEE, 2013
- H. Papadopoulos, “Inductive conformal prediction: Theory and application to neural networks,” *Tools in Artificial Intelligence*, vol. 18, pp. 315–330, 2008
- U. Johansson, H. Boström, T. Löfström, and H. Linusson, “Regression conformal prediction with random forests,” *Machine Learning*, vol. 97, no. 1-2, pp. 155–176, 2014

Combined conformal predictors

- V. Vovk, “Cross-conformal predictors,” *Annals of Mathematics and Artificial Intelligence*, pp. 1–20, 2013
- L. Carlsson, M. Eklund, and U. Norinder, “Aggregated conformal prediction,” in *Artificial Intelligence Applications and Innovations*. Springer, 2014, pp. 231–240
- H. Papadopoulos, “Cross-conformal prediction with ridge regression,” in *Statistical Learning and Data Sciences*. Springer, 2015, pp. 260–270

Not (yet) proven valid

But seems to be working well in practice.

Application domains

- A. Lambrou, H. Papadopoulos, E. Kyriacou, C. S. Pattichis, M. S. Pattichis, A. Gammerman, and A. Nicolaides, “Assessment of stroke risk based on morphological ultrasound image analysis with conformal prediction,” in *Artificial Intelligence Applications and Innovations*. Springer, 2010, pp. 146–153
- D. Devetyarov, I. Nourtdinov, B. Burford, S. Camuzeaux, A. Gentry-Maharaj, A. Tiss, C. Smith, Z. Luo, A. Chervonenkis, R. Hallett et al., “Conformal predictors in early diagnostics of ovarian and breast cancers,” *Progress in Artificial Intelligence*, vol. 1, no. 3, pp. 245–257, 2012
- M. Eklund, U. Norinder, S. Boyer, and L. Carlsson, “The application of conformal prediction to the drug discovery process,” *Annals of Mathematics and Artificial Intelligence*, vol. 74, no. 1-2, pp. 117–132, 2015

Application domains

- I. Nourtdinov, S. G. Costafreda, A. Gammernan, A. Chervonenkis, V. Vovk, V. Vapnik, and C. H. Fu, “Machine learning classification with confidence: application of transductive conformal predictors to mri-based diagnostic and prognostic markers in depression,” *Neuroimage*, vol. 56, no. 2, pp. 809–813, 2011
- J. Vega, A. Murari, S. Dormido-Canto, R. Moreno, A. Pereira, A. Acero, and J.-E. Contributors, “Adaptive high learning rate probabilistic disruption predictors from scratch for the next generation of tokamaks,” *Nuclear Fusion*, vol. 54, no. 12, p. 123001, 2014

Suggested reading

- V. Vovk, A. Gammerman, and G. Shafer, Algorithmic learning in a random world. Springer, 2005
- www.alrw.net
- G. Shafer and V. Vovk, “A tutorial on conformal prediction,” The Journal of Machine Learning Research, vol. 9, pp. 371–421, 2008
- A. Gammerman, V. Vovk, and V. Vapnik, “Learning by transduction,” in Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1998, pp. 148–155
- A. Gammerman and V. Vovk, “Hedging predictions in machine learning the second computer journal lecture,” The Computer Journal, vol. 50, no. 2, pp. 151–163, 2007

Suggested reading cont.

- H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman, “Inductive confidence machines for regression,” in *Machine Learning: ECML 2002*. Springer, 2002, pp. 345–356
- H. Papadopoulos and H. Haralambous, “Reliable prediction intervals with regression neural networks,” *Neural Networks*, vol. 24, no. 8, pp. 842–851, 2011
- U. Johansson, H. Boström, T. Löfström, and H. Linusson, “Regression conformal prediction with random forests,” *Machine Learning*, vol. 97, no. 1-2, pp. 155–176, 2014

Nonconformist - conformal prediction in Python

Repository: <https://github.com/donlnz/nonconformist>

Docs: <http://donlnz.github.io/nonconformist/>

Available on PyPi







```
% pip install nonconformist
```






Questions, suggestions, feedback, contributions, etc.?







henrik.linusson@hb.se






QUESTIONS?






REFERENCES



-  V. Vovk, A. Gammerman, and G. Shafer, Algorithmic learning in a random world. Springer, 2005.
-  V. Vovk, V. Fedorova, I. Nourtdinov, and A. Gammerman, “Criteria of efficiency for conformal prediction,” 2014.
-  V. Vovk, “Conditional validity of inductive conformal predictors,” Journal of Machine Learning Research - Proceedings Track, vol. 25, pp. 475–490, 2012.
-  H. Linusson, U. Johansson, H. Boström, and T. Löfström, “Efficiency comparison of unstable transductive and inductive conformal classifiers,” in Artificial Intelligence Applications and Innovations. Springer, 2014, pp. 261–270.
-  U. Johansson, H. Boström, T. Löfström, and H. Linusson, “Regression conformal prediction with random forests,” Machine Learning, vol. 97, no. 1-2, pp. 155–176, 2014.
-  H. Boström, H. Linusson, T. Löfström, and U. Johansson, “Accelerating difficulty estimation for conformal regression forests,” Annals of Mathematics and Artificial Intelligence, pp. 1–20, 2017.

-  L. Carlsson, E. Ahlberg, H. Boström, U. Johansson, and H. Linusson, “Modifications to p-values of conformal predictors,” in *Statistical Learning and Data Sciences*. Springer, 2015, pp. 251–259.
-  U. Johansson, E. Ahlberg, H. Boström, L. Carlsson, H. Linusson, and C. Sönströd, “Handling small calibration sets in mondrian inductive conformal regressors,” in *Statistical Learning and Data Sciences*. Springer, 2015, pp. 271–280.
-  R. Laxhammar and G. Falkman, “Conformal prediction for distribution-independent anomaly detection in streaming vessel data,” in *Proceedings of the First International Workshop on Novel Data Stream Pattern Mining Techniques*. ACM, 2010, pp. 47–55.
-  V. Fedorova, A. Gammerman, I. Nouretdinov, and V. Vovk, “Plug-in martingales for testing exchangeability on-line,” in *29th International Conference on Machine Learning*, 2012.
-  U. Johansson, R. König, H. Linusson, T. Löfström, and H. Boström, “Rule extraction with guaranteed fidelity,” in *Artificial Intelligence Applications and Innovations*. Springer, 2014, pp. 281–290.

-  X. Zhu, F.-M. Schleif, and B. Hammer, “Semi-supervised vector quantization for proximity data,” in Proc. of European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2013), Louvain-La-Neuve, Belgium, 2013, pp. 89–94.
-  H. Papadopoulos, V. Vovk, and A. Gammerman, “Regression conformal prediction with nearest neighbours,” *Journal of Artificial Intelligence Research*, vol. 40, no. 1, pp. 815–840, 2011.
-  U. Johansson, H. Boström, and T. Löfström, “Conformal prediction using decision trees,” in *International Conference Data Mining (ICDM)*. IEEE, 2013.
-  H. Papadopoulos, “Inductive conformal prediction: Theory and application to neural networks,” *Tools in Artificial Intelligence*, vol. 18, pp. 315–330, 2008.
-  U. Johansson, H. Boström, T. Löfström, and H. Linusson, “Regression conformal prediction with random forests,” *Machine Learning*, vol. 97, no. 1-2, pp. 155–176, 2014.
-  V. Vovk, “Cross-conformal predictors,” *Annals of Mathematics and Artificial Intelligence*, pp. 1–20, 2013.

-  L. Carlsson, M. Eklund, and U. Norinder, “Aggregated conformal prediction,” in Artificial Intelligence Applications and Innovations. Springer, 2014, pp. 231–240.
-  H. Papadopoulos, “Cross-conformal prediction with ridge regression,” in Statistical Learning and Data Sciences. Springer, 2015, pp. 260–270.
-  A. Lambrou, H. Papadopoulos, E. Kyriacou, C. S. Pattichis, M. S. Pattichis, A. Gammerman, and A. Nicolaidis, “Assessment of stroke risk based on morphological ultrasound image analysis with conformal prediction,” in Artificial Intelligence Applications and Innovations. Springer, 2010, pp. 146–153.
-  D. Devetyarov, I. Nouretdinov, B. Burford, S. Camuzeaux, A. Gentry-Maharaj, A. Tiss, C. Smith, Z. Luo, A. Chervonenkis, R. Hallett et al., “Conformal predictors in early diagnostics of ovarian and breast cancers,” Progress in Artificial Intelligence, vol. 1, no. 3, pp. 245–257, 2012.
-  M. Eklund, U. Norinder, S. Boyer, and L. Carlsson, “The application of conformal prediction to the drug discovery process,” Annals of Mathematics and Artificial Intelligence, vol. 74, no. 1-2, pp. 117–132, 2015.

-  I. Nourtdinov, S. G. Costafreda, A. Gammerman, A. Chervonenkis, V. Vovk, V. Vapnik, and C. H. Fu, “Machine learning classification with confidence: application of transductive conformal predictors to mri-based diagnostic and prognostic markers in depression,” *Neuroimage*, vol. 56, no. 2, pp. 809–813, 2011.
-  J. Vega, A. Murari, S. Dormido-Canto, R. Moreno, A. Pereira, A. Acero, and J.-E. Contributors, “Adaptive high learning rate probabilistic disruption predictors from scratch for the next generation of tokamaks,” *Nuclear Fusion*, vol. 54, no. 12, p. 123001, 2014.
-  G. Shafer and V. Vovk, “A tutorial on conformal prediction,” *The Journal of Machine Learning Research*, vol. 9, pp. 371–421, 2008.
-  A. Gammerman, V. Vovk, and V. Vapnik, “Learning by transduction,” in *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1998, pp. 148–155.
-  A. Gammerman and V. Vovk, “Hedging predictions in machine learning the second computer journal lecture,” *The Computer Journal*, vol. 50, no. 2, pp. 151–163, 2007.

-  H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman, “Inductive confidence machines for regression,” in *Machine Learning: ECML 2002*. Springer, 2002, pp. 345–356.
-  H. Papadopoulos and H. Haralambous, “Reliable prediction intervals with regression neural networks,” *Neural Networks*, vol. 24, no. 8, pp. 842–851, 2011.