Statistical and Algorithmic Learning

Alexander Gammerman

Computer Learning Research Centre
Department of Computer Science
Royal Holloway, University of London

IDEAL Conference
November 2018
Madrid

Jesus Manuel de la Cruz (1955-2017)

This talk is devoted to the memory of

Jesus Manuel de la Cruz
Full Professor of System Engineering and Automatic Control
Complutense University, Madrid

Jesus Manuel de la Cruz

In October 1992 he joined the Department of Computer Architecture and Automatic Control in Complutense University of Madrid. He was Dean of the Department from 1997 to 2001 His interest covered broad aspects of automatic control and its applications, real time control, optimization, statistical learning, and robotics. He contributed to six books and supervised 13 Ph.D. students and numerous M.Sc. students.

Content

1. Introduction
2. Part 1. Conformal Predictors
3. Part 2. Applications
4. References

Motivation

Statistical and Algorithmic Learning

- Statistical Learning: Statistics and Functional Analysis. It deals with the problem of finding a predictive function based on data.
- Algorithmic Learning: based on Algorithmic Information Theory (Shannon's information theory and Turing computability theory).

Motivation

- To develop classification and regression algorithms with reliable measures of confidence under a very general assumption is a motivation behind conformal prediction approach.
- This framework is founded on the principles of algorithmic randomness, transductive inference and hypothesis testing, and has several desirable properties for potential use in various real-world applications, such as the calibration of the obtained confidence values in an online setting.

Transduction

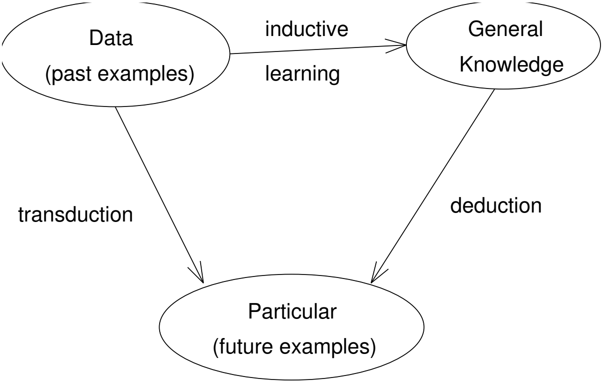


Figure: Transductive inference [Vapnik, 1995]

Introduction

Problem of Prediction: classification and regression

Classical techniques: small scale, low-dimensional data.

But conceptual and computational difficulties for high-dimensional and big data.

Also most ML systems output "bare" prediction only.
But how good the predictions are?

Require: measures of "reliability" of every prediction.

Problems and motivations

Lack of useful measures of confidence in their predictions.

- Bayesian methods?
- PAC-learning?

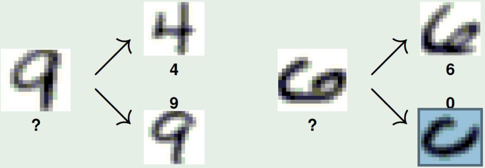
Features:

- validity
- online
- assumptions
- "wrapper" style ML algorithm
- individual prediction

Solution: **Conformal Predictors**

Hand-written digits: 16 × 16 pixels grayscale image and label

- 7291 training examples
- 2007 test examples



Examples

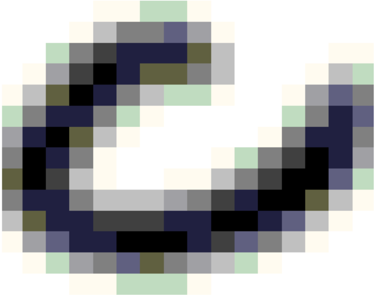


Figure: Example:unknown digit

Conformal predictors

:

- To “hedge” predictions: to complement with measures of their accuracy and reliability (“confidence machines” or “conformal predictors”).
- These measures are valid, informative, tailored to the individual object to be predicted.
- Automatic validity under the randomness assumption (the data are i.i.d.): they never overrate the accuracy and reliability of their predictions.
- Any classification or regression algorithm can be transformed into a conformal predictor.

Conformal Predictors and various extensions

Conformal Prediction (CP) in this talk is just abbreviation for a whole range of relevant algorithms that have been used in various fields. Among them:

- Transductive CP
- Inductive CP - for computational efficiency.
- Mondrian CP - for imbalanced data.
- On-line CP - for data which is updated in time.
- Transfer Learning CP - how to use acquired knowledge in an *old* domain to improve the efficiency and accuracy of learning in a *new* domain.
- On-line Compression Model - for assumptions other than i.i.d.
- Venn–Abers predictor - produces reliable two-sided probabilistic estimates instead of p-values.
- Ridge Regression Confidence Machine.

History: Algorithmic Learning in a Random World, Springer 2005; "Measures of Complexity", Festschrift for Alexey Chervonenkis, Springer, 2015; COPA Symposiums 2012–2017.

General Idea

For classification: try every possible label Y as a candidate for x_{l+1}'s label and see how well the resulting sequence

$$(x_1, y_1), \ldots, (x_l, y_l), (x_{l+1}, Y)$$

conforms to the randomness (or i.i.d.) assumption (if it does conform to this assumption, we will say that it is "random").

The ideal case: all Ys but one lead to sequences that are not random. We can then use the remaining Y as a *confident* (or hedged) **prediction** for y_{l+1}.

Problem of hedged prediction is intimately connected with the problem of **testing randomness**.

Different versions of the “universal” notion of randomness were defined by Kolmogorov, Martin-Löf and Levin based on the existence of universal Turing machines.

Martin-Löf (developing Kolmogorov’s earlier ideas) proved that there exists a smallest, to within a constant factor, randomness test.

Let **Z** be the set of all possible examples; as each example consists of an object and a label, $\mathbf{Z} = \mathbf{X} \times \mathbf{Y}$, where **X** is the set of all possible objects and **Y**, $|\mathbf{Y}| > 1$, is the set of all possible labels.

Martin-Löf's definition

A function $t : \mathbf{Z}^* \to [0,1]$ is a *randomness test* if

1. for all $\epsilon \in (0,1)$, all $n \in \{1,2,\dots\}$ and all probability distributions P on $\mathbf{Z}$,
$$P^n\{z \in \mathbf{Z}^n : t(z) \le \epsilon\} \le \epsilon; \tag{1}$$
2. t is upper semicomputable.

Validity: for example, if $t(z) \le 1\%$
then a rare event or data are not i.i.d.

Computability: the universal test is upper semi-computable.

Prediction with confidence and credibility

Once we have a randomness test t, we can use it for hedged prediction. 2 ways to package the results: 1) to predict each example with confidence and credibility:

- consider all possible values $Y \in \mathbf{Y}$ for the label y_{l+1};
- find the randomness level detected by t for every possible completion $(x_1, y_1), \ldots, (x_l, y_l), (x_{l+1}, Y)$;
- predict the label Y corresponding to a completion with the largest randomness level detected by t;
- output as the **confidence** in this prediction one minus the second largest randomness level detected by t;
- output as the **credibility** of this prediction the randomness level detected by t of the output prediction Y (i.e., the largest randomness level detected by t over all possible labels).

2) Another way is to use a "prediction set": choose a range of “confidence levels” $1 - \epsilon$, and for each of them specify a prediction set $\Gamma^{\epsilon} \subseteq \mathbf{Y}$

Intuition: choose a conventional "significance level", such as 1%.
If the confidence in our prediction is 99% or more and the prediction is wrong – a very rare event happened (the set of all data sequences with randomness level detected by t not exceeding 1%).

Intuitively, low credibility means that either the training set is non-random or the test object is not representative of the training set (say, in the training set we have images of digits and the test object is that of a letter).

Martin-Löf's definition of algorithmic randomness deficiency is a universal notion of the statistical notion of p-values.

Conformal predictor maps each data set

$$(x_1, y_1), \ldots, (x_l, y_l),$$

$l = 0, 1, \ldots$., each new example x_{l+1}, and each confidence level $1 - \epsilon \in (0, 1)$, into the prediction set

$$\Gamma^{\epsilon}(x_1, y_1, \ldots, x_l, y_l, x_{l+1}) := \{Y \in \mathbf{Y} : p_Y > \epsilon\},$$

where p_Y are defined by

$$p_Y := \frac{|\{i = 1, \ldots, l+1 : \alpha_i \geq \alpha_{l+1}\}|}{l+1}$$

with $\alpha_1, \ldots, \alpha_{l+1}$ being the nonconformity scores corresponding to the completion

$$(x_1, y_1), \ldots, (x_l, y_l), (x_{l+1}, Y).$$

Associating with each completion its p-value gives a randomness test. Therefore: for each l the probability of the event

$$y_{l+1} \in \Gamma^{\epsilon}(x_1, y_1, \dots, x_l, y_l, x_{l+1})$$

is at least $1 - \epsilon$.

In the case of classification we can summarize the prediction sets Γ^{ϵ} by two numbers: the confidence

$$\sup\left\{1 - \epsilon : |\Gamma^{\epsilon}| \le 1\right\}$$

and the credibility

$$\inf\left\{\epsilon : |\Gamma^{\epsilon}| = 0\right\}.$$

Smoothed CP

If the p-values are redefined as

$$p_Y := \frac{|\{i : \alpha_i > \alpha_{l+1}\}| + \eta\,|\{i : \alpha_i = \alpha_{l+1}\}|}{l+1},$$

where $i \in \{1, \ldots, l+1\}$ (i ranges over $\{1, \ldots, l+1\}$ and $\eta \in [0,1]$ is generated randomly from the uniform distribution on $[0,1]$, we obtain *smoothed conformal predictors.*

NCM: general definition

A nonconformity measure is a function that assigns to every data point a number: $\alpha_1, \dots, \alpha_l$, called nonconformity scores, such that: interchange of any two examples (x_i, y_i) and (x_j, y_j) leads to the interchange of the corresponding nonconformity scores (with all the other nonconformity scores unchanged).

The greater NCM value is, the stranger example is relative to the training set. As a simple example, one can consider the Nearest Neighbours conformity measure, where NCM is the average distance from the example to its nearest neighbours in the training set.

Intuitively the smaller p-value is, the stranger the observation is.

Examples of non-conformity measures

- **Support Vector Machine:** Extracting α_i from $\sum_{i=1}^{l+1} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l+1} \alpha_i \alpha_j y_i y_j K(x_i, x_j) \to \min_{\alpha_1,\dots,\alpha_{l+1}}$
- **k Nearest Neighbours:** $\alpha_i := \sum_{j=1}^{k} d_{ij}^{+} / \sum_{j=1}^{k} d_{ij}^{-}$ where d_{ij}^{+} is the jth shortest distance from x_i to other examples with the same label y_i, and d_{ij}^{-} – with a different label.
- **Nearest Centroid:** NCM of an example is its distance to the "average" feature vector of the same class examples divided by distance to the nearest "average" feature vector of another class.
- **Neural Networks:** NCM of an example is the ratio of the output layer unit corresponding to the true class divided by the highest of output layer units corresponding to the other classes.
- **Random Forest** is a set of decision trees voting for classes. The NCM of an example is a number of votes against the true class.

Main idea of SVM:

- *to map* the original set of vectors into a h/d feature space and
- then to construct a linear separating *hyperplane* (or a linear regression function)in this feature space.

The task: to find a separating hyperplane with a small number of errors and a large "margin".

Formally, this is done by finding the minimum of the objective function,

$$\frac{1}{2}(w \cdot w) + C\left(\sum_{i=1}^{l} \xi_i\right) \to \min, \tag{2}$$

subject to the constraints

$$y_i\left((x_i \cdot w) + b\right) \geq 1 - \xi_i,\ i = 1, \ldots, l.$$

Here C is a fixed positive constant (maybe ∞), w are weights, b is the intercept, and ξ_i are non-negative "slack variables".

The original setting can be replaced by “dual” setting: maximise a quadratic form:

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i, x_j) \to \max$$

under the constraints

$$0 \leq \alpha_i \leq C, \quad i = 1, 2, \ldots, l.$$

Here, K is the kernel and the values α_i, $i = 1, \ldots, l$, are the Lagrange multipliers corresponding to the training vectors and for each non-zero α_i there is a corresponding vector x_i. If x is a new vector, the prediction $\hat{y}$ is

$$\hat{y} = \mathrm{sign}\left(\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b \right).$$

These Lagrange multipliers can play role of NCM in this setting.

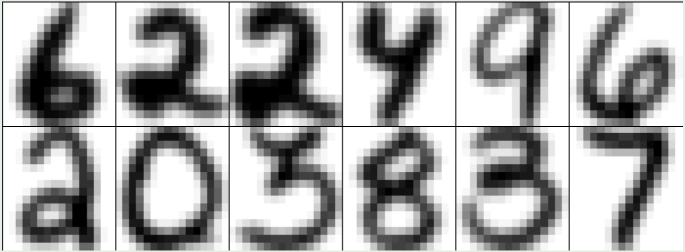
On-line prediction protocol

```
$\mathrm{Err}_0 := 0$;   $\mathrm{Mult}_0 := 0$;   $\mathrm{Emp}_0 := 0$;
FOR $n = 1, 2, \ldots$:
    Reality outputs $x_n \in \mathbf{X}$;
    Predictor outputs $\Gamma_n^\epsilon \subseteq \mathbf{Y}$ for all $\epsilon \in (0, 1)$;
    Reality outputs $y_n \in \mathbf{Y}$;
    $\mathrm{err}_n^\epsilon := \begin{cases} 1 & \text{if } y_n \notin \Gamma_n^\epsilon \\ 0 & \text{otherwise}, \end{cases} \quad \epsilon \in (0, 1)$;
    $\mathrm{Err}_n^\epsilon := \mathrm{Err}_{n-1}^\epsilon + \mathrm{err}_n^\epsilon, \quad \epsilon \in (0, 1)$;
    $\mathrm{mult}_n^\epsilon := \begin{cases} 1 & \text{if } |\Gamma_n^\epsilon| > 1 \\ 0 & \text{otherwise}, \end{cases} \quad \epsilon \in (0, 1)$;
    $\mathrm{Mult}_n^\epsilon := \mathrm{Mult}_{n-1}^\epsilon + \mathrm{mult}_n^\epsilon, \quad \epsilon \in (0, 1)$;
    $\mathrm{emp}_n^\epsilon := \begin{cases} 1 & \text{if } |\Gamma_n^\epsilon| = 0 \\ 0 & \text{otherwise}, \end{cases} \quad \epsilon \in (0, 1)$;
    $\mathrm{Emp}_n^\epsilon := \mathrm{Emp}_{n-1}^\epsilon + \mathrm{Emp}_n^\epsilon, \quad \epsilon \in (0, 1)$
END FOR.
```

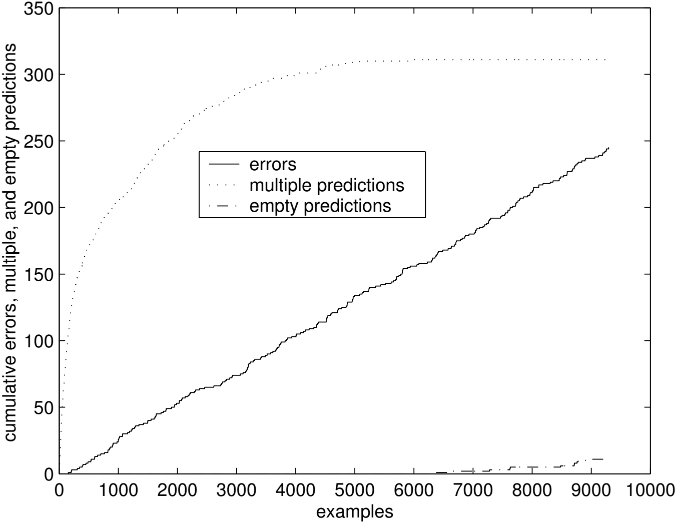
Examples: US Postal Service

Hand-written digits: 16 × 16 pixels grayscale image and label

- 7291+2007 examples



Efficiency – Uncertain (Mult) Predictions for USPS at 97.5% confidence



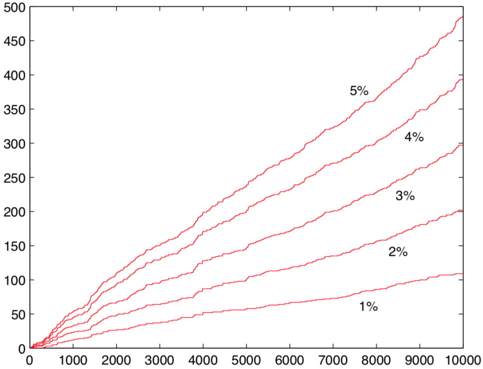


Figure: MNIST data

Applications

- Fault Diagnosis
- Homeland Security (anomaly detection)
- Plazma Physics (images classification)
- Network traffic
- Information security (bots detection)
- Environment (water and air pollution)
- Biology (plant promoter prediction; PPI, analysis microarrays)
- Pharmaceutical industry (compounds activities)
- Veterinary
- Electronic nose
- Medicine

Applications

- Abdominal Pain
- Ovarian Cancer
- Depression (diagnostic and prognosis)
- Child Leukemia
- Heart Diseases
- Osteoporosis prognosis

Gastroenterology[1]

A patient with acute abdominal pain:

APP	DIV	PPU	NAP	CHO	INO	PAN	RCO	DYS	true label
1.2%	0.4%	0.2%	2.8%	5.7%	0.9%	1.4%	0.5%	80.6%	DYS

At 95% the prediction set is multiple, {cholecystitis, dyspepsia}.

At 90%, the prediction set narrows down to {dyspepsia}

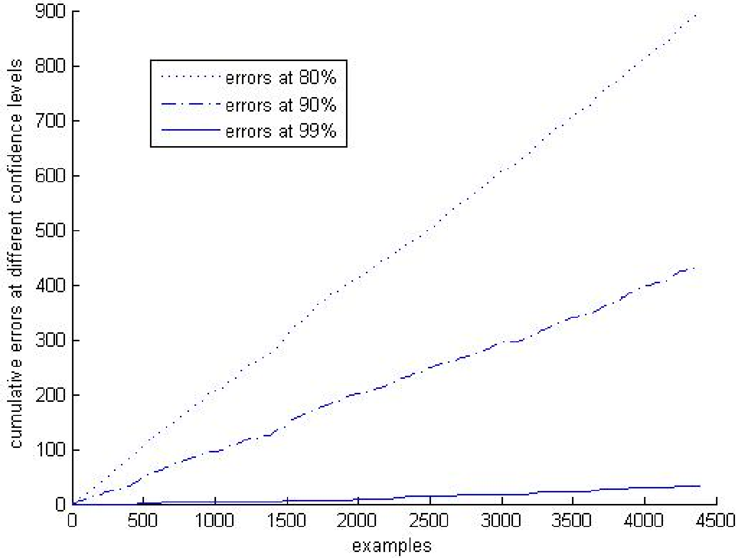
At 99% the prediction set widens to {appendicitis, non-specific abdominal pain, cholecystitis, pancreatitis, dyspepsia}.

Prediction (diagnosis): Dispepsia
Confidence = 94.3% (high confidence – all alternatives are unlikely).
Credibility = 80.6% (low credibility – whole situation is suspect).
Overall results: Accuracy of 9-class diagnostic is 74%.

http : //turing.cs.rhul.ac.uk/ ∼ leo/apex/

[1] A.Gammerman and A.R.Thatcher. Bayesian diagnostic probabilities; Meth of Inf in Med, v.30, No.1; Papadopoulos, H., Gammerman, A., Vovk, V. Reliable Diagnosis of Acute Abdominal Pain with Conformal Prediction. Eng Intel Sys, 17(2–3), 127–137. CRL Publishing (2009).

Gastroenterology (validity)



Inductive and Mondrian Conformal Predictor

ICP for ”big data": training set is divided to a proper training set and a validation set. The proper training set is used only to calculate NCM scores (αs) of calibration and testing examples. Then *p*-values are calculated using only these αs.

MCP: Variation of CP allows to have separate guarantees of the errors of different types. CP prediction set covers the true label with probability $1 - \epsilon$. In Mondrian CP: if the true label is 1, then the prediction set contains 1 with probability $1 - \epsilon_1$; if the true label is 0, then the prediction set contains 1 with probability $1 - \epsilon_0$.

Conclusions

Advantages of the hedging technique:

- it gives provably *valid* measures of confidence, in the sense that they never overrate the accuracy and reliability of the predictions;
- it does not make any additional assumptions about the data beyond the IID assumption (the examples are independent and identically distributed);
- it allows to estimate the confidence in the prediction of individual examples;
- conformal predictors can be used as *region predictors*, allowed to output a range of labels as their prediction, so that one can control the number of erroneous predictions by selecting a suitable confidence level;
- the well-calibrated prediction regions produced by conformal predictors can be used in both on-line and off-line modes of learning, as well as in several intermediate modes, allowing, for example, “slow” and “lazy” teachers.

References

- Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, 2005.
- A.Gammerman and V.Vovk. Hedging Prediction in Machine Machine Learning. *The Computer Journal,, v.50, No.2, pp. 151-163, 2007*.
- Nouretdinov, I., Costafreda, S.G., Gammerman, A., Chervonenkis, A., Vovk, V., Vapnik, V., Fu, C.H.Y. Machine learning classification with confidence: Application of transductive conformal predictors to MRI-based diagnostic and prognostic markers in depression. *Neuroimage*; 56(2):809-13, 2011.
- Paolo Toccaceli; Ilia Nouretdinov; Alexander Gammerman. Conformal Predictors for Compound Activity Prediction. In: "Conformal and Probabilistic Prediction with Applications"LNAI Proceedings. Vol. 9653, pp. 51-66; Springer, 2016. *Lecture Notes in Computer Science; Vol. 9653*.