University of Tokyo, Komaba Campus

Thursday, May 27, 2004

Conformal Prediction

Glenn Shafer

Rutgers University

Newark, New Jersey, USA

This document has been designed to be viewed two pages at a time—pp. 2-3 together, pp. 4-5 together, and so on. (Select "Continuous - Facing" from the "View" menu in Adobe Acrobat Reader.)

These slides were revised after the lecture, on June 1, 2004.

Conformal Prediction

Although machine-learning methods often work well, the performance guarantees proven for them are typically too asymptotic to be useful.

Conformal predictors, which perform equally well, come developed with simple and useful measures of confidence.

Reference: Chapter 2 of *Algorithmic learning in a random world*, by Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. Springer, to appear.

Outline of the Lecture

1. A short history of prediction intervals in statistics

2. The task of machine learning

3. Confidence prediction

4. Situating conformal prediction in statistical learning theory

5. From a nonconformity measure to a conformal predictor

6. The generality of the method

1. A short history of prediction intervals

- In 1908, Student discovered the t distribution and applied it to significance testing.

- In the early 30s, Fisher used the t distribution for "fiducial inference".

- In 1937, Neyman got confidence intervals by removing the mystical and arbitrary.

William Gossett
("Student")
1876–1937

R. A. Fisher
(Founder of modern statistics)
1890-1962

Less often remembered: Fisher also derived prediction intervals.

- Sample $x_1, \ldots, x_n$ from $\mathcal{N}(\mu, \sigma)$.
- Calculate $\overline{x}$ and s.
- Sample x from $\mathcal{N}(\mu, \sigma)$.

- **1933.** $1 - \epsilon$ **confidence interval for** μ**:**

$$\overline{x} \pm t_{n-1}^{\epsilon/2} \frac{s}{\sqrt{n}}$$

$$\left(\text{Because } \sqrt{n}\,\frac{\overline{x} - \mu}{s} \text{ is } t \text{ with } n-1 \text{ df.}\right)$$

- **1935.** $1 - \epsilon$ **prediction interval for** x**:**

$$\overline{x} \pm t_{n-1}^{\epsilon/2}\, s\sqrt{1 + \frac{1}{n}}$$

$$\left(\text{Because } \sqrt{\frac{n}{n+1}}\,\frac{\overline{x} - x}{s} \text{ is } t \text{ with } n-1 \text{ df.}\right)$$

Tolerance intervals

After 1940, mathematical statisticians

- went with Neyman,
- emphasized parameters,
- ignored Fisher's predictive concept.

Inspired by Shewhart's work on quality control, Wilks formulated the notion of a *tolerance interval*. It was studied in the 40s, 50s, and 60s by Wald, Wolfowitz, Tukey, Fraser, Kemperman, and Guttman.

Walter Shewhart
(Founder of quality control)
1891–1967

Samuel S. Wilks
(Brought statistics to Princeton)
1906–1962

Say you are sampling from a distribution P on $\mathbf{Z}$.

- You do not know P exactly. You know $P \in \mathcal{P}$.
- From $z_1, \ldots, z_n$, you want to predict a new observation z.
- You use a mapping Γ from $\mathbf{Z}^n$ to subsets of $\mathbf{Z}$.
- Γ is a (δ, ϵ)-*tolerance region* if

$$P^n\{P[\Gamma(z_1, \ldots, z_n)] \geq 1 - \epsilon\} \geq 1 - \delta$$

for all $P \in \mathcal{P}$.

Starting in the late 1950s, Fisher's idea of a prediction interval gained a little traction.

- He advertised it in his influential 1956 book.
- Some attention was paid to a simplified concept of tolerance region that formalizes and generalizes Fisher's idea:

Γ is an $(1 - \delta)$-*expectation tolerance region* if

$$P^{n+1}\{z \in \Gamma(z_1, \ldots, z_n)\} \geq 1 - \delta$$

for all $P \in \mathcal{P}$.

- Formulas for prediction intervals began to appear in textbooks on regression.

In mathematical statistics, prediction intervals still receive little attention.

This is occasionally deplored in the pages of the *American Statistician*.

Confidence and prediction intervals for linear regression with normal errors

- **Model:** $y = \alpha + \beta x + e,$ where $e \sim \mathcal{N}(\mu, \sigma)$

- **Data:** $(x_1, y_1), \ldots, (x_n, y_n)$

- $(1 - \epsilon)$ **confidence interval for** $\alpha + \beta x$

$$\widehat{\alpha} + \widehat{\beta} x \pm t_{n-2}^{\epsilon/2} \, s \left(\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right)^{1/2}$$

- **Additional data:** x_{n+1}

- $(1 - \epsilon)$ **prediction interval for** y_{n+1}

$$\widehat{\alpha} + \widehat{\beta} x_{n+1} \pm t_{n-2}^{\epsilon/2} \, s \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \right)^{1/2}$$

Unpublished fact: The prediction errors are independent.

Let H_n be the event that the prediction of y_n is a hit—i.e., that y_n is in the prediction interval based on

$$(x_1, y_1), \ldots, (x_{n-1}, y_{n-1}), x_n$$

When $n-1$ is equal to 3 or more, we can get a regression line and non-zero residual variance s^2 from $(x_1, y_1), \ldots, (x_{n-1}, y_{n-1})$. So it makes sense to talk about H_n for n equal to 4 or more.

Theorem. The events $H_4, H_5, \ldots$ are mutually independent.

No one has ever published this simple result.

How neglected prediction intervals are!!!

2. The task of machine learning

Reality outputs $(x_1, y_1), (x_2, y_2), \ldots$.

$x_i \in \mathbf{X}$, the *object space*

$y_i \in \mathbf{Y}$, the *label space*

$\mathbf{Z} := \mathbf{X} \times \mathbf{Y}$, the *example space*

object x label y

example $z = (x, y)$

- Here an object is a 16×16 matrix, with each entry chosen from 31 shades of gray. So $\mathbf{X}$ has $31^{16\times16} \approx 10^{382}$ elements.
- $Y = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$.

$z_1, \ldots, z_n$ is the same as $(x_1, y_1), \ldots, (x_1, y_n)$

The task: Predict each label after seeing its object.

- From x_1, predict y_1.
- From $(x_1, y_1), x_2$, predict y_2.
- From $(x_1, y_1), (x_2, y_2), x_3$, predict y_3.
- Etc.

Assumption: Randomness.

Reality chooses the examples independently from some probability distribution Q on $\mathbf{Z}$.

(The z_i are independent & identically distributed.)

No assumptions about Q.

Usually independence can be weakened to exchangeability.

3. Prediction with confidence

$$\mathbf{Z} = \mathbf{X} \times \mathbf{Y}$$

Write $\mathbf{Z}^*$ for the set of all finite sequences of elements of $\mathbf{Z}$:

$$\mathbf{Z}^* = \cup_{n=0}^{\infty} \mathbf{Z}^n$$

A *level* $(1-\epsilon)$ *confidence predictor* is a mapping

$$\Gamma : \mathbf{Z}^* \times \mathbf{X} \to 2^{\mathbf{Y}}.$$

After observing old examples $z_1, \ldots, z_{n-1}$ and the new object x_n, we predict that x_n's label y_n will be in the subset

$$\Gamma(z_1, \ldots, z_{n-1}, x_n)$$

of the label space $\mathbf{Y}$.

Write H_n for the event that this prediction is correct, and call H_n. This event is the predictor's *hit* on the nth round.

A $(1-\epsilon)$ confidence predictor is

- *exactly valid* if its hits are independent and all happen with probability $(1-\epsilon)$.

- *conservatively valid* if the probability that the predictions on rounds $n_1, \ldots, n_k$ are all hits is always at least $(1-\epsilon)^k$.

(The probability statements must be true no matter what probability distribution Q on $\mathbf{Z}$ governs the data.)

In our forthcoming book, Vovk, Gammerman, and I demonstrate the existence of valid confidence predictors and explain how they are constructed.

Valid confidence predictors are constructed from nonconformity measures

As Vovk, Gammerman, and I explain, a valid confidence predictor is obtained from a real-valued function A on strings of the form

$$(x_1, y_1), \ldots, (x_{n-1}, y_{n-1}), (x, y).$$

To get the confidence predictor,

- interpret $A((x_1, y_1), \ldots, (x_{n-1}, y_{n-1}), (x, y))$ as a measure of how different (x, y) is from $(x_1, y_1), \ldots, (x_{n-1}, y_{n-1})$;

- predict values of y_n that make (x_n, y_n) differ minimally from $(x_1, y_1), \ldots, (x_{n-1}, y_{n-1})$.

I will explain this more precisely later.

Because of the way they are constructed, valid confidence predictors are called *conformal predictors*.

The function A is called a *nonconformity measure*.

From a given nonconformity measure, we construct a $(1-\epsilon)$ confidence predictor Γ^ϵ for every $\epsilon \in [0,1]$, and they are nested in the natural way:

$$\Gamma^{\epsilon_1}(z_1,\ldots,z_{n-1},x_n) \subseteq \Gamma^{\epsilon_2}(z_1,\ldots,z_{n-1},x_n)$$

when $\epsilon_1 \geq \epsilon_2$.

The more confident you want to be, the larger the region.

4. Situating conformal prediction in statistical learning theory

Vladimir Vapnik, the founder of statistical learning theory, has recently emphasized the distinction between *induction* and *transduction*.

- **Induction:** Fix n. Sample $z_1, \ldots, z_n$. Use this sample to make a rule for predicting a new y from a new x.

  This rule might be hard wired into devices to be used on many new examples $z_{n+1}, \ldots, z_m$, each involving a y to be predicted from an x. (Example: give postal workers devices for reading zip codes.)

- **Transduction:** On each round n, use $z_1, \ldots, z_n$ and x_{n+1} to predict y_{n+1}.

  Perhaps in the course of making the prediction we formulate a rule for predicting a new y from n examples and a new x, but we do not store this rule for very long, because on the next round we update it to a rule for predicting a new y from $n+1$ examples and a new x.

Induction is more common in practice than transduction, for two reasons:

- Transduction requires a teacher.

  In order the postal worker's device to update its algorithm, it must be told the correct digit for each example that it predicts!

- Until now, estimating confidence for predictions required a batch approach: the *hold-out method*.

  1. Divide the data into two halves — an estimation set $z_1, \ldots, z_l$ and a test set $z_{l+1}, \ldots, z_n$.
  2. Calculate a prediction rule from the estimation set.
  3. Observe the success rate of the prediction rule on the test set.
  4. If the rule works 90% of the time on the test set, give it 90% confidence in future examples.

The usual practice in machine learning (PAC, VC) is to give a rule Γ for forming a prediction region $\Gamma(z_1, \ldots, z_{n-1}, x_n) \subseteq \mathbf{Y}$, together with and a upper bound on the probability that y will fail to be in the region. One then shows that this upper bound tends to zero as $n \to \infty$.

- The prediction regions often work well in practice; with reasonable values of n, there is a very low probability of y not being in the region.

- But the theoretical upper bounds on this probability are typically useless (larger than 1) for these reasonable values of n.

Bottom line: Vapnik's statistical learning theory does not give degrees of confidence. For this, one must fall back on the naive method of the hold-out estimate.

Our new theory of conformal predictors solves the problem of confidence prediction. And the practical performance of our predictors is as good as those of statisical learning theory.

But this does not eliminate the need to use induction rather than transduction when there is no teacher to correct each new prediction after it is made. In these situations, batch methods must be used, and the hold-out estimate will probably be competitive.

Vladimir Vapnik, the founder of statistical learning theory, did not demonstrate the existence of valid confidence predictors.

Our understanding of how to construct them is due mainly to Vladimir Vovk.

Vladimir Vapnik
(Ph.D., Moscow, 1964)

Vladimir Vovk
(Ph.D., Moscow, 1988)

5. From a nonconformity measure to a conformal predictor

Recall that a *bag* (or multiset) is a collection of elements in which repetition is allowed.

(A bag is different from a set, because elements may be repeated. But it is like a set in that its elements are not ordered.)

Write $\mathcal{B}$ for the set of all finite bags of elements of $\mathbf{Z}$.

Definition: A *nonconformity measure* is a function $A : \mathcal{B} \times \mathbf{Z} \to \mathbb{R}$.

We interpret $A(B, z)$ as the degree to which z is strange with respect to the bag B.

Formally, any function $A : \mathcal{B} \times \mathbf{Z} \to \mathbb{R}$ qualifies as a nonconformity measure and can be used in our theory, but in practice we choose A so that a large value of $A(B, z)$ indicates that z is strange relative to B.

As always, $\mathbf{Z} = \mathbf{X} \times \mathbf{Y}$ and $z = (x, y)$.

Example: Regression

Suppose $\mathbf{X} = \mathbf{Y} = \mathbb{R}$. Then a useful nonconformity measure is

$$A(B, z) = |y - \widehat{\alpha} - \widehat{\beta} x|,$$

where $\widehat{\alpha} + \widehat{\beta} x$ is the least-squares line based on the examples in the bag B.

Any other method of obtaining a regression line can also be used.

Example: Classification

Suppose $\mathbf{X} = \mathbb{R}^k$, while $\mathbf{Y}$ is finite. Then a useful nonconformity measure is

$$A(\{z_1, \ldots, z_n\}, z) := \frac{\min_{i:y_i=y} d(x_i, x)}{\min_{i:y_i \neq y} d(x_i, x)},$$

where d is Euclidean distance.

Any other way of scoring a proposed classifcation can also be used.

How to construct a 95% confidence region for y_n from

- a nonconformity measure A
- old examples $z_1, \ldots, z_{n-1}$, and
- a new object x_n.

1. Consider separately each $y \in \mathbf{Y}$.
2. Let B be the bag consisting of $z_1, \ldots, z_{n-1}$ together with (x_n, y).
3. For $i = 1, \ldots, n$, let B^{-i} be the bag obtained by removing z_i. (For the moment, (x_n, y) will serve as z_n.)
4. Set $W_i := A(B^{-i}, z_i)$.
5. Set
$$p_y := \frac{\#\{i = 1, \ldots, n \mid W_i \geq W_n\}}{n} .$$
Call this the p-*value* for y. It is the fraction of the elements in B that are at least as strange relative to the others as (x_n, y).
6. Include y in the confidence region if and only if $p_y > 0.05$.

Lemma

Suppose $W_1, \ldots, W_n$ are exchangeable random variables. Set

$$U := \#\{i = 1, \ldots, n : W_i \geq W_n\}$$

$$= \begin{array}{l}\text{number of the } W \text{ at least} \\ \text{as large as the last one}\end{array}$$

Then

$$\mathbb{P}\{\frac{U}{n} > \epsilon\} \geq 1 - \epsilon$$

for all $\epsilon \in [0, 1]$.

Proof: The random variable U has the possible values $1, \ldots, n$, each with equal probability. So if

$$\frac{i}{n} \leq \epsilon < \frac{i+1}{n},$$

then

$$\mathbb{P}\{\frac{U}{n} > \epsilon\} = \mathbb{P}\{\frac{U}{n} > \frac{i}{n}\} = \mathbb{P}\{U > i\}$$

$$= \frac{n-i}{n} = 1 - \frac{i}{n} \geq 1 - \epsilon.$$

6. The Generality of the Method

Any method of statistical prediction, together with any way of measuring its error, produces a nonconformity measure and therefore a conformal predictor:

- Let D be a way of predicting y from a bag of old examples and a new object x.
- Suppose $\mathsf{dist}(y_1, y_2)$ is some metric on $\mathbf{Y}$.
- Set $A(B, (x, y) := \mathsf{dist}(D(B, x), y)$.

The nonconformity approach is universal: Any method of obtaining valid nested confidence regions arises from a nonconformity measure. This is true because the p-values themselves define a nonconformity measure.

If we believe the examples are being generated by a certain model (Gaussian errors, say), then we may want to use a nonconformity measure based on method of prediction that is optimal for that model (least squares, say).

This will be efficient (the confidence regions will be small) if the proposed model is right.

It will be valid even if the model is not right. (We assume only exchangeability.)