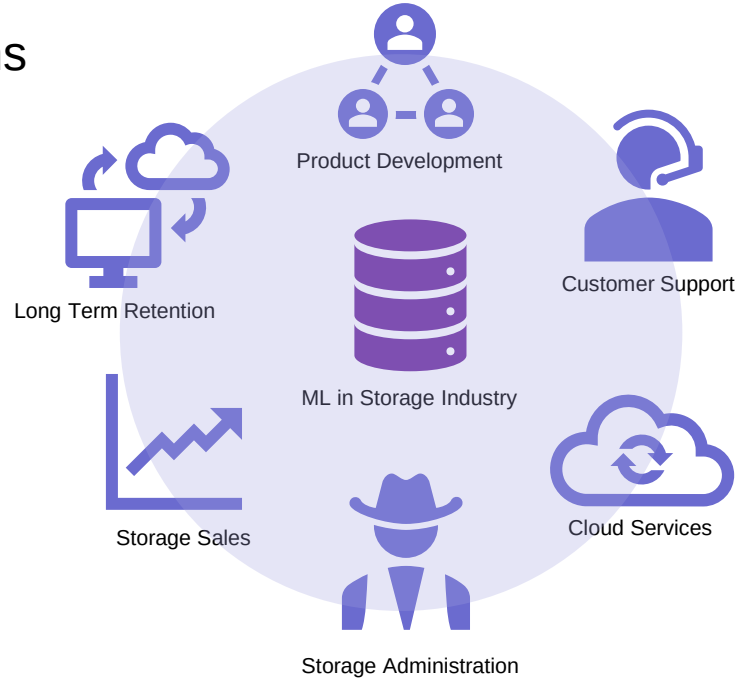
SDC 19
SNIA INDIA
May 23-24, 2019
Bangalore, India
STORAGE DEVELOPER CONFERENCE

Understanding the Reliability of Predictions Made by Machine Learning

Rahul Vishwakarma, Supriya Kannery
DELL EMC

Motivation

- Drive Failure and Disk Full Predictions
- Predictions' Reliability
- Standard vs Conformal



Introduction to Conformal Prediction

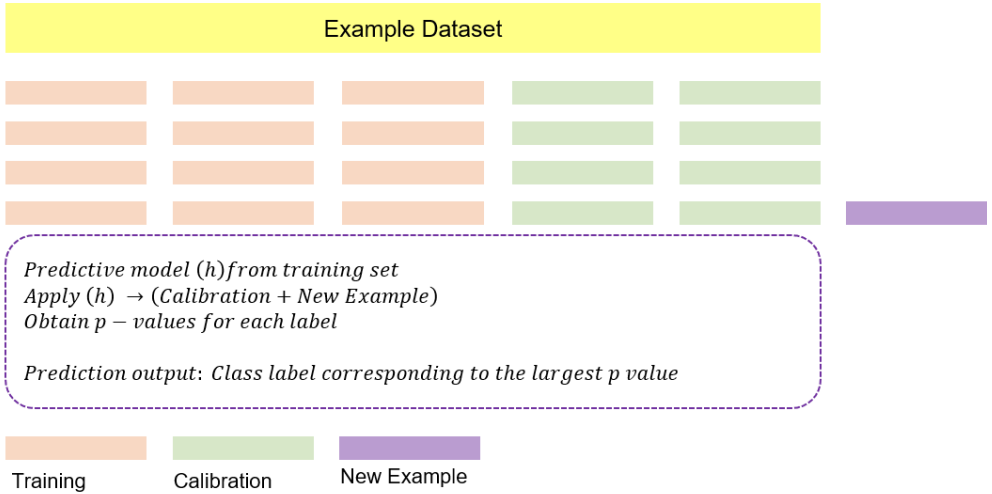
Prediction complemented with valid confidence measures

- Confidence – Indication of confidence of predictions
- Credibility – Quality of data for making the prediction

Goal – High confidence predictions along with a credibility that is not too low

Conformal Prediction Framework

- Predictive model (h)
- Calibration set
- Non-conformity score (A)



Disk Drive Failure Prediction

sn	class	s1	s3	s5	s7	s9	s187	s189	s194	s195	s197	sr5	sr197
17178	NORMAL	0.419355	0.555556	1	0.317073	0.789474	1	1	-0.6	-0.13979	1	-0.99953	-1
12815	NORMAL	0.419355	0.555556	1	0.390244	0.578947	1	0.898	-0.65714	-0.09677	1	-0.99953	-1
14164	FAIL	0.225806	0.555556	1	0.390244	0.578947	1	1	-0.31429	-0.11828	1	-1	-1
10641	NORMAL	0.290323	0.555556	1	0.390244	0.578947	1	1	-0.6	-0.1828	1	-1	-1
16015	?	0.290323	0.703704	1	0.317073	0.789474	1	1	-0.6	-0.09677	1	-0.99953	-1

SMART parameters for Disk Drive
Source: http://pan.baidu.com/share/link?shareid=189977&uk=4278294944

Given an example set $z_i = (x_i, y_i)$, $i = 1, 2, 3, \ldots, n-1$
Independent variable $s1, \ldots, sr197 = x_i \in R^d$
Dependent variable $class = y_i \in Y$
Predict $class$ of x_n

Try each class label $c \in Y$ as prediction for x_n
Measure randomness of sequence $z_1 = (x_i, y_i)$, . . ., $z_{n-1} = (x_{n-1}, y_{n-1})$, $z_n = (x_n, c)$

Conformal Classification

Calculate Calibration scores

$\alpha_1 = A(x_1, y_1 = NORMAL, h) = 1 - P_h(NORMAL, x_1) = 0.10$
$\alpha_2 = A(x_2, y_2 = FAILED, h) = 1 - P_h(FAILED, x_1) = 0.14$
. . .
$\alpha_n = A(x_n, y_n = NORMAL, h) = 1 - P_h(NORMAL, x_n) = 0.64$

Calculate p-value for each possible class label I_j

$\alpha_{n+1} = A(x_{n+1}, y_{n+1} = I_j$
$P_{I_j} = |\{\alpha_i : \alpha_i \geq \alpha_{n+1}\}|/(n+1)$

$\alpha_{n+1} = A(x_{n+1}, \acute{y} = FAILED) = 1 - P_h(FAILED, x) = 0.85$ $\quad \boldsymbol{P_{FAILED} = 0.03}$
$\alpha_{n+1} = A(x_{n+1}, \acute{y} = NORMAL) = 1 - P_h(NORMAL, x) = 0.15$ $\quad \boldsymbol{P_{NORMAL} = 0.75}$

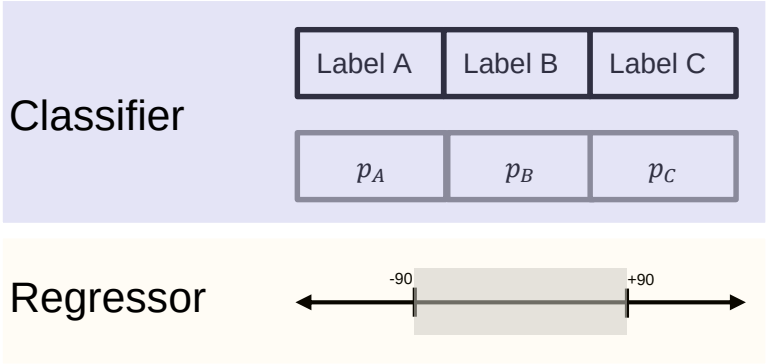
$\acute{y} = NORMAL$

Conformal prediction for Disk Drive failure

Label	*Confidence*	*Credibility*
NORMAL	0.752179	0.777801
NORMAL	0.613903	0.538756
FAILED	0.68458	0.97442
...	...	...

SDC 19
SNIA INDIA

Regression



Regression

Mode $y = \alpha + \beta x + \epsilon$ where $\epsilon \sim \xi(\mu, \sigma)$
Data points $(x_1 y_1), \ldots, (x_n, y_n)$

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

$$\hat{\alpha} + \hat{\beta} x \pm t_{n-2}^{\epsilon/2} \, s \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\Sigma_{i=1}^{n} (x - \bar{x})^2} \right)^{1/2}$$

For new data x_{n+1} $(1 - \epsilon)$ confidence interval for y_{n+1}

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

Conformal prediction for Disk Full

min	max	truth	size
11.46	39.50	25.00	28.04
19.57	47.61	50.00	28.04
17.91	45.95	26.40	28.04
2.40	30.44	13.80	28.04
6.76	34.80	13.60	28.04
13.21	41.25	13.60	28.04
2.40	30.44	23.20	28.04
17.88	45.92	16.10	28.04
2.40	30.44	19.40	28.04
11.82	39.86	26.50	28.04
7.37	35.41	22.50	28.04
…	…	…	…

Time Series

$Given\colon\ a_1, a_2, a_3, \dots, a_{i-1}\ where\ a_i \subset R^K$
$Predict\colon\ a_i$

Introducing exchangeability

$A \forall T+1 \le i \le n\ :\ z_i = (x_i, y_i) := \big((a_{i-T}, \dots, a_{i-1}), a_i\big)\ where\ n\ is\ length\ of\ time\ series$

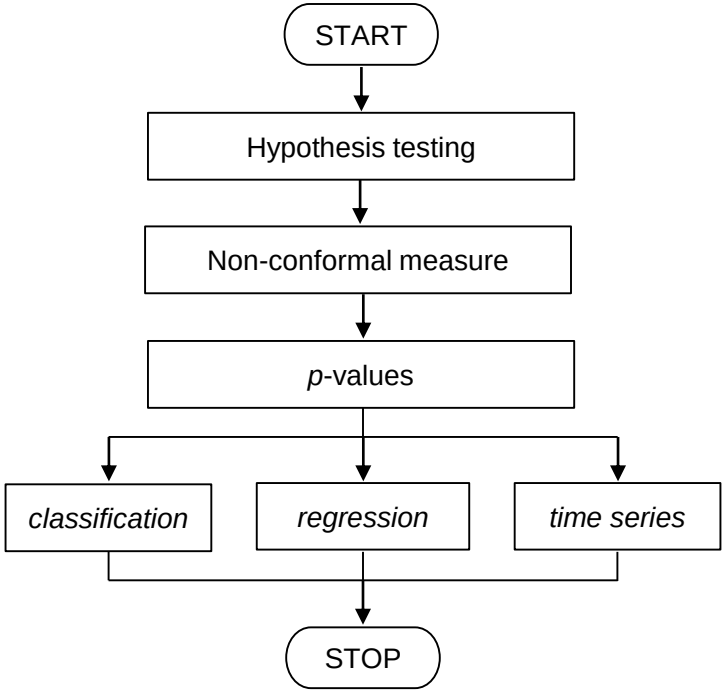
Example of transformed data

$If\ n = 6\ and\ T = 2$

$\{z_1, z_2, z_3, z_4\} = \{((a_1, a_2), a_3), (a_2, a_3), a_4), (a_3, a_4), a_5), (a_4, a_5), a_6), \}$

SDC 19
SNIA INDIA

Summary



References

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

G. Shafer and V. Vovk, “A tutorial on conformal prediction,” The Journal of Machine Learning Research, vol. 9, pp. 371–421, 2008.

Balasubramanian, V., Ho, S. and Vovk, V. (2014). Conformal Prediction for Reliable Machine Learning.