INDO-UK WORKSHOP ON
CONFORMAL PREDICTION FOR RELIABLE MACHINE LEARNING

Introduction to Conformal Prediction

15 Dec 2015

Vineeth N Balasubramanian
Department of Computer Science and Engineering
Indian Institute of Technology, Hyderabad

Uncertainty Estimation: An Overview

Sources of Uncertainty

Data Uncertainty

- Is the data noisy?
- Are any data values missing?
- Are data attributes correlated?

Model Uncertainty

- How typical of the training points is the given test data point?
- Was there a bias in the training dataset?
- What portion of the possible universe of datasets was provided for training?

Algorithm Uncertainty

- Different algorithms have different assumptions
- Heuristics in algorithms
- Choice of parameters in algorithms

Uncertainty Estimation

Approaches

Probabilistic

- Data is modeled as probability distributions

Statistical

- E.g. Use model errors to build confidence intervals

Simulation/Resampling

- Monte Carlo methods

Fuzzy

- Non-probabilistic methodology based on membership functions

Evidence-based

- Dempster-Shafer theory, Dezert-Smarandache theory, Possibility theory

Heuristic

- E.g. Distance of a data point from the decision hyperplane

Uncertainty Estimation: An Overview

Representations

Probability as confidence

- Most common approach

Confidence intervals

- Most popular in statistical analysis

Credible intervals

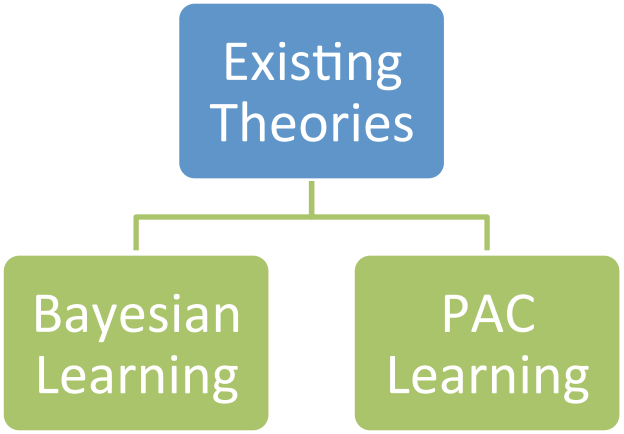
- Bayesian confidence intervals

Gamesman intervals

- Output prediction interval contains the true output a large fraction of the time, and this fraction can be set by the user.

Uncertainty Estimation

Existing Theories



$$P(A|B) = P(B|A)P(A)/P(B)$$

The learner needs to select a generalization function such that with high probability (*probably*), the function will have low generalization error (*approximately*)

Assume that data set $\mathcal{D}_n$, consisting of n data points, was generated from some true θ^*, then under some regularity conditions, as long as $p(\theta^*) > 0$

$$\lim_{n\to\infty} p(\theta|\mathcal{D}_n) = \delta(\theta - \theta^*)$$

C is **PAC-learnable** by L using H if for all $c \in C$, distributions D over X, ε such that $0 < \varepsilon < \frac{1}{2}$, and δ such that $0 < \delta < \frac{1}{2}$, learner L will with probability at least $(1 - \delta)$ output a hypothesis $h \in H$ such that $error_D(h) \leq \varepsilon$, in time that is polynomial in $\frac{1}{\varepsilon}$, $\frac{1}{\delta}$, n, and $size(C)$.

Uncertainty Estimation

Limitations of Existing Theories

Bayesian methods

- Have strong underlying assumptions
- Provide asymptotic guarantees

PAC learning

- Cannot be applied to each test data point individually
- Not very useful in practice
- Provide asymptotic guarantees

Uncertainty Estimation

Existing Theories

- Littlestone and Warmuth's theorem
  - Error bound for SVM

$$b \leq \frac{1}{l-d}\left(d \ln \frac{el}{d} + \ln \frac{l}{\delta}\right)$$

    - Where "d" is the number of support vectors
  - USPS dataset experiment (for one of 10 classifiers)
    - For binary classifier

$$\frac{1}{l-d}\left(d \ln \frac{el}{d} + \ln \frac{l}{\delta}\right) \approx \frac{1}{7291-274} 274 \ln \frac{7291e}{274} \approx 0.17,$$

    - For multi-class classifier

$$\frac{1}{l-d}\left(d \ln \frac{el}{d} + \ln \frac{l}{\delta}\right) \approx \frac{1}{7291-1677} 1677 \ln \frac{7291e}{1677} \approx 0.74$$

Nouretdinov, Vovk, Vyugin, Gammerman. Pattern Recognition and Density Estimation under the General i.i.d. Assumption. COLT 2001, 337-353.

Uncertainty Estimation

Desiderata

Validity

- The number of errors made by the system should be 1-t, if the confidence value is given as t.

Optimality

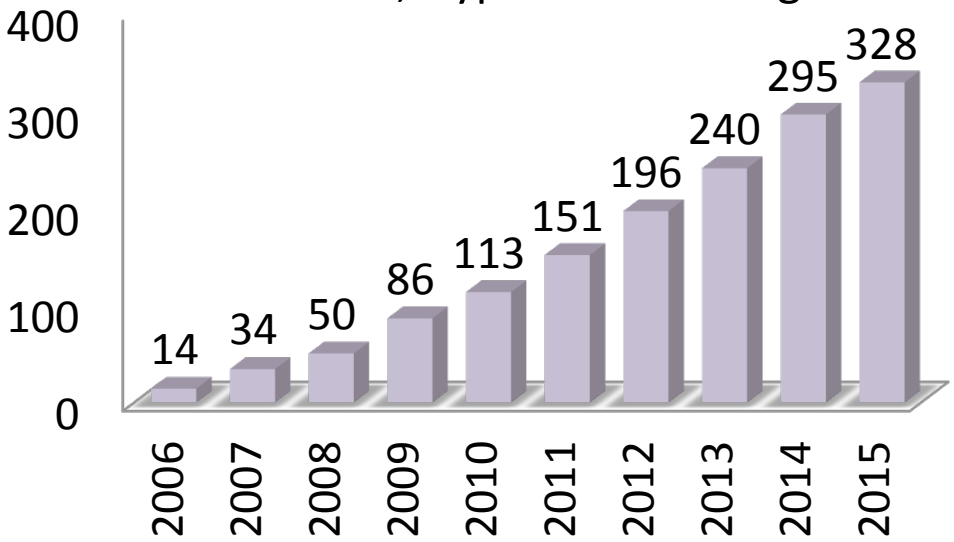
- Prediction regions should have as narrow a width as possible

Generalizability

- Should be extensible to all kinds of classification/ regression algorithms, as well as multiple classifier/regressor systems

Conformal Prediction

- Developed by Vovk, Shafer, Gammerman
- Based on algorithmic randomness, Transductive inference, Hypothesis testing



Conformal Prediction

In a Nutshell

- Given an error probability ε, together with a method that makes a prediction Y of a label y, it produces a set of labels, typically containing y with probability 1-ε
- CP has a theoretical guarantee for an **online setting** (in which the labels are predicted successively, each one being revealed before the next is predicted), and empirical validity for offline settings

V. Vovk, Online Confidence Machines are Well-Calibrated, FOCS 2002.

Conformal Predictions Framework

Classification Tasks

Compute non-conformity scores

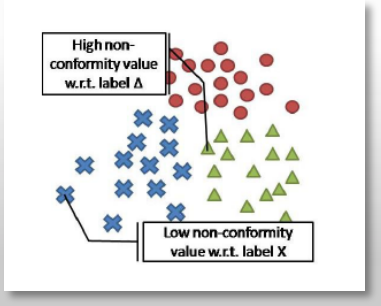
Compute p-value for every hypothesis

Output the hypotheses whose p-values satisfy the confidence level

Can be defined suitably for any classifier

$$\alpha_i^y = \frac{\sum_{j=1}^{k} D_{ij}^{y}}{\sum_{j=1}^{k} D_{ij}^{-y}}$$

K-NN



$$p(\alpha_{n+1}^{y_p}) = \frac{count\left\{i : \alpha_i^{y_p} \geq \alpha_{n+1}^{y_p}\right\}}{n+1}$$

Non-conformity measure

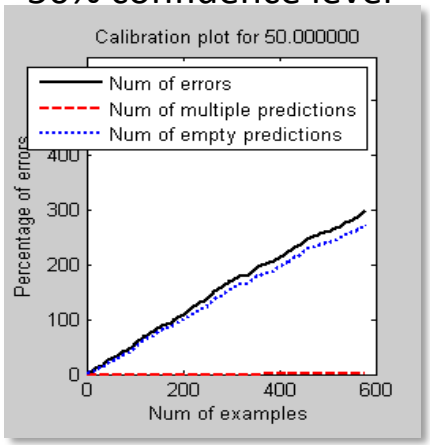
Conformal prediction regions with confidence level 1- ε

$$\Gamma_{1-\varepsilon} = \{y_i : P^{y_i} > \varepsilon, y_i \in Y\}$$

Conformal Prediction

Empirical Performance

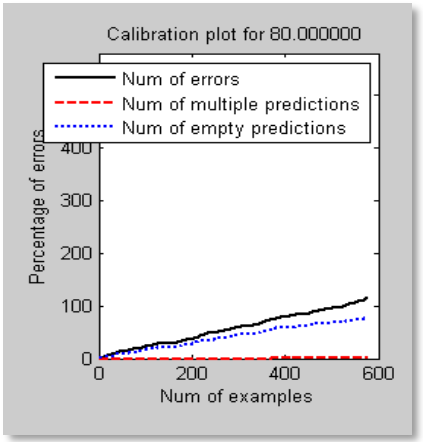
50% confidence level



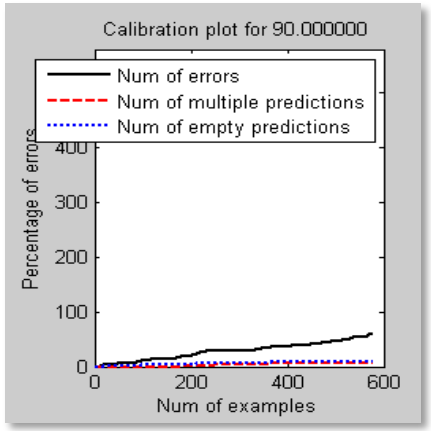
60% confidence level

70% confidence level

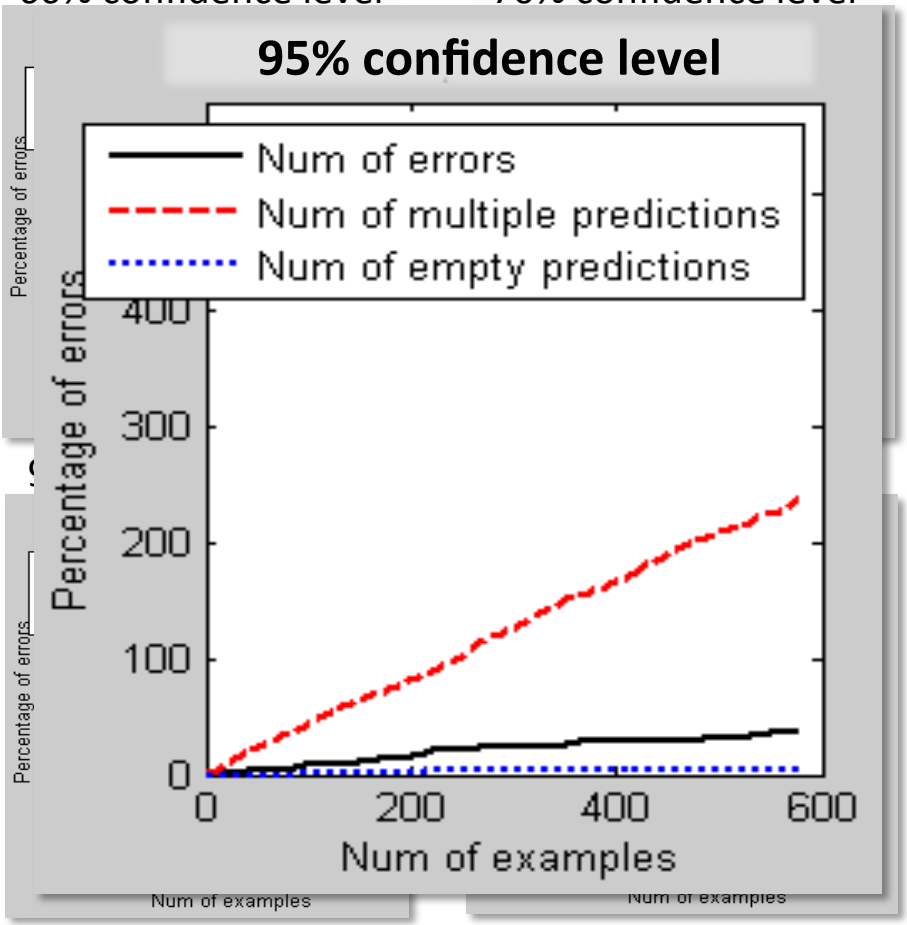
80% confidence level



90% confidence level



95% confidence level



Conformal Prediction

Motivation

- How good is your prediction y?
- How confident are you that the prediction Y for a new object is the correct label?
- If the label y is a number, how close do you think the prediction Y is to y?

The usual prediction goal

We want new predictions to perform as well as past predictions

Conformal Prediction

Can we ...

1) Allow a user to specify a confidence level or error rate so that a method cannot perform worse than the predefined level or rate before prediction

or

2) provide confidence/uncertainty level for all possible outcomes?

Conformal Prediction

A Brief History

- 1999: Machine learning applications of algorithmic randomness (VGS), ICML
- 2002: Transductive confidence machines for pattern recognition (PNVG), ECML; Inductive confidence machines for regression (PPVG), ECML; Online confidence machines are well-calibrated (V), FOCS.
- 2003: First adaptation (active learning, (HW), IJCNN) beyond supervised learning ...
- 2004: Algorithmic learning in a random world (VGS)

Conformal Prediction

Randomness

- Assumption: examples are generated independently from the same distribution
- A data sequence is said to be random with respect to a statistical model if a test does not detect any lack of conformity between the two.

Conformal Prediction

Randomness

- Kolmogorov's algorithmic approach to complexity: formalizing the notion of a random sequence.
- Complexity of a finite string *z* can be measured by the length of the shortest program for a universal Turing machine that outputs the string *z*.

Conformal Prediction

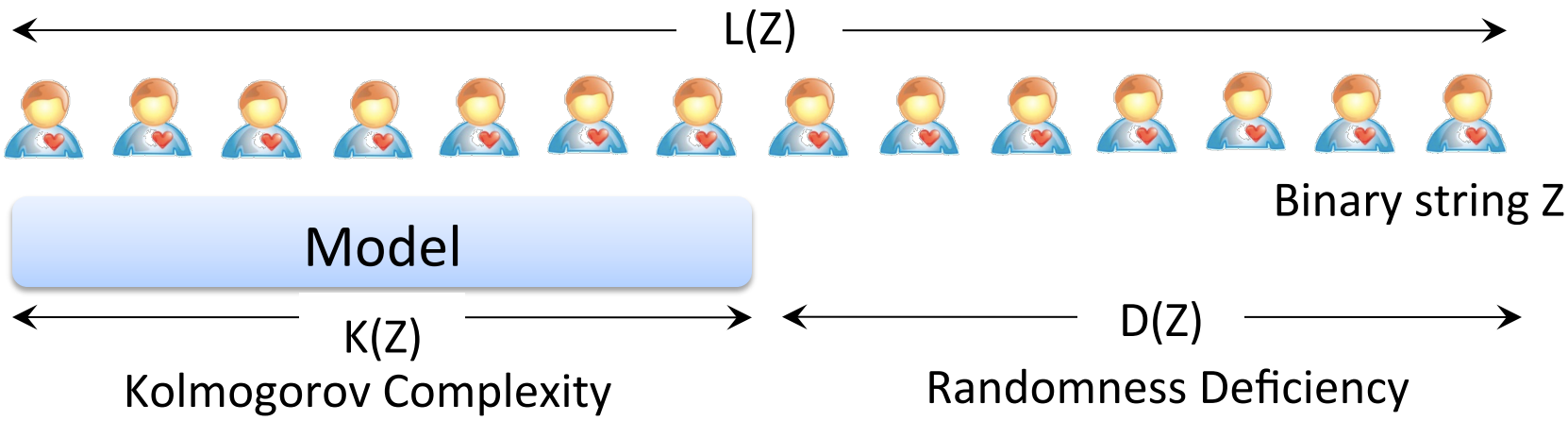
Randomness

- Two useful characteristics of Kolmogorov's notion of randomness
  - Applies to finite sequence
  - Provide degrees of randomness

- Alternative - Minimum Description Length (MDL) Principle: regularity (lacks of randomness) implies the ability to compress the given string.

Conformal Prediction

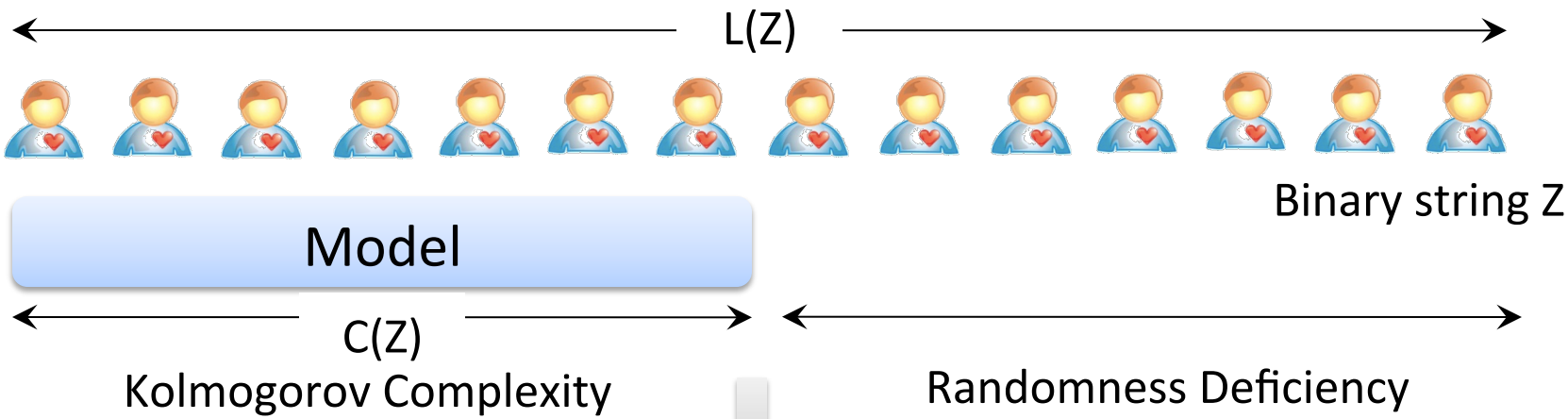
Background



- Connection between incompressibility and randomness
- *K(z)* is small (compressible), *D(z)* is high (lack of randomness)

Conformal Prediction

Background



$$P\left\{z \in Z^n : t'(z) \leq 2^{-m}\right\} \leq 2^{-m}$$

Conformal Prediction

Background

Definition: Let P_n be a set of computable probability distributions in a sample space X^n containing elements made up of n data points. A function t: $X^n \rightarrow \mathbf{N}$, the set of natural numbers N including ∞, is a Martin-Lof test for randomness if

- t is enumerable; and
- for all $n \in N$ and $m \in N$ and $P \in P_n$,

$$P\{\mathbf{x} \in X^n: t(\mathbf{x}) \geq m\} \leq 2^{-m}.$$

Martin Lof test for randomness

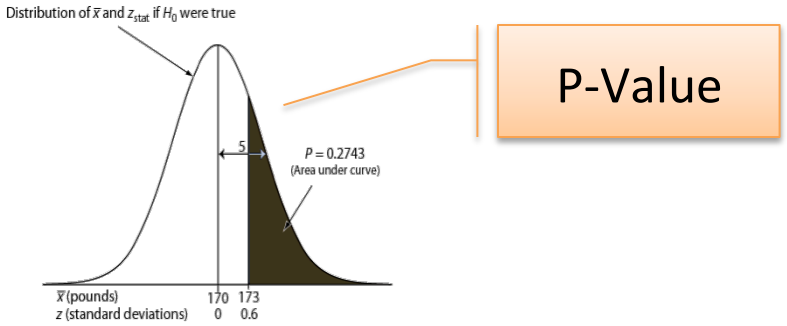
$$P\left\{z \in Z^n : t'(z) \leq 2^{-m}\right\} \leq 2^{-m}$$

Conformal Prediction

Hypothesis Testing: A Review

- Null/Alternate Hypothesis, Evaluate Test Statistic, Compute P-Value
- Example
  - In the 1970s, 20–29 year old men had a mean body weight μ of 170 kgs. Standard deviation σ was 40 kgs. We test whether mean body weight in the population now differs.
  - **Null hypothesis** $H_{0:}$ μ = 170 ("no difference")
  - The **alternative hypothesis** can be either $H_{a:}$ μ > 170 (**one-sided test**) or $H_{a:}$ μ ≠ 170 (**two-sided test**)

$$z_{\text{stat}} = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}} = \frac{173 - 170}{5} = 0.60$$



Conformal Prediction

Background

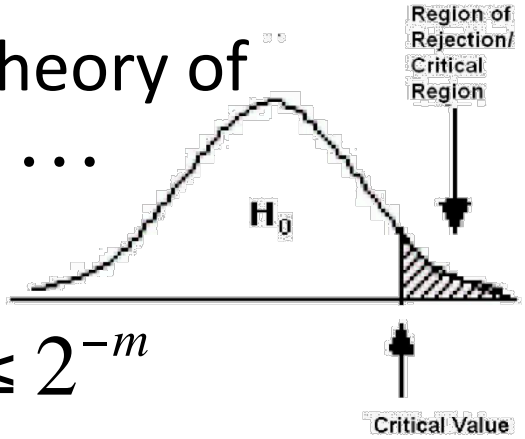
- Constructing critical regions used in the theory of hypothesis testing such that $C_1 \supseteq C_2 \supseteq \cdots$

  where $C_m = \{x : t(x) \geq m\}$

  and the critical function is $P\{x \in C_m\} \leq 2^{-m}$

  for all n at a fixed m

  with the significance level (size of the test or Type I error or false negative) $\alpha = 2^{-m}$



Conformal Prediction

Background

- Using the Martin-Lof randomness test definition, one can reconstruct the critical regions in the theory of hypothesis. By transform the test t using $f(a) = 2^{-a,}$ one gets

Definition: Let P_n be a set of computable probability distributions in a sample space X^n containing elements made up of n data points. A function $t : X^n \rightarrow (0,1]$ is a p–value function if for all $n \in N$, $P \in P_n$ and $r \in (0,1]$,

$$P\{\mathbf{x} \in X^n : t(\mathbf{x}) \leq r\} \leq r$$

- Equivalent to the statistical notion of p-value, a measure on how well the data support or discredit a null hypothesis.

Conformal Prediction

Prediction via Hypothesis Testing

- An example x is assigned a label y.
- Hypothesis Test:
  - Null: The data sequence S U {(x, y)} is random in the sense that they are generated independently from the same distribution
  - Alternate: The data sequence S U {(x, y)} is not random.

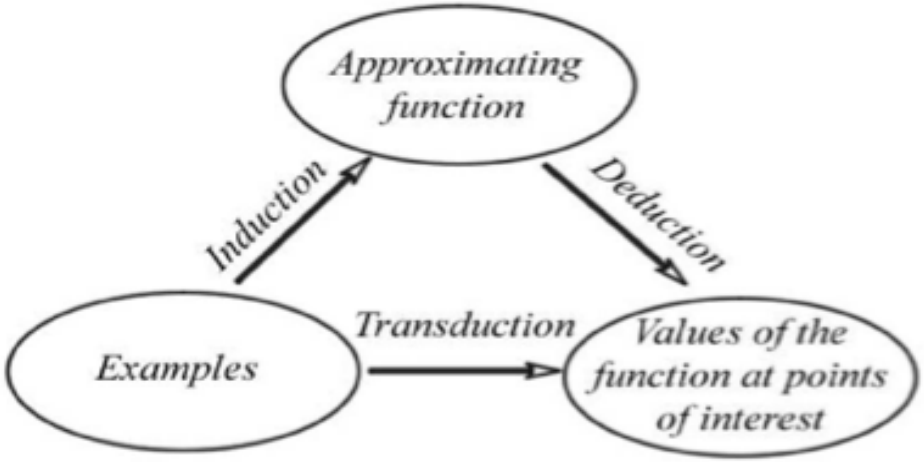
Conformal Prediction

Prediction via Hypothesis Testing

- Classification problem: Given a sequence labeled data, S and a test object x
- Consider all possible label values for y
- Find the randomness level detected by t for each possible label y, on the data sequence, S U {(x, y)}
- Predict the label Y that has the largest randomness level detected by t.
- Transductive Confidence Machine, 2002

Conformal Prediction

Connection to Transductive Inference



Using the older examples *S* directly to predict label based on the randomness level for the new object *x* without deriving a general rule.

Conformal Prediction

p-Values (randomness level from t)

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

- Approximation for randomness level since t is only upper semi-computable.

Conformal Prediction

Non-conformity measures

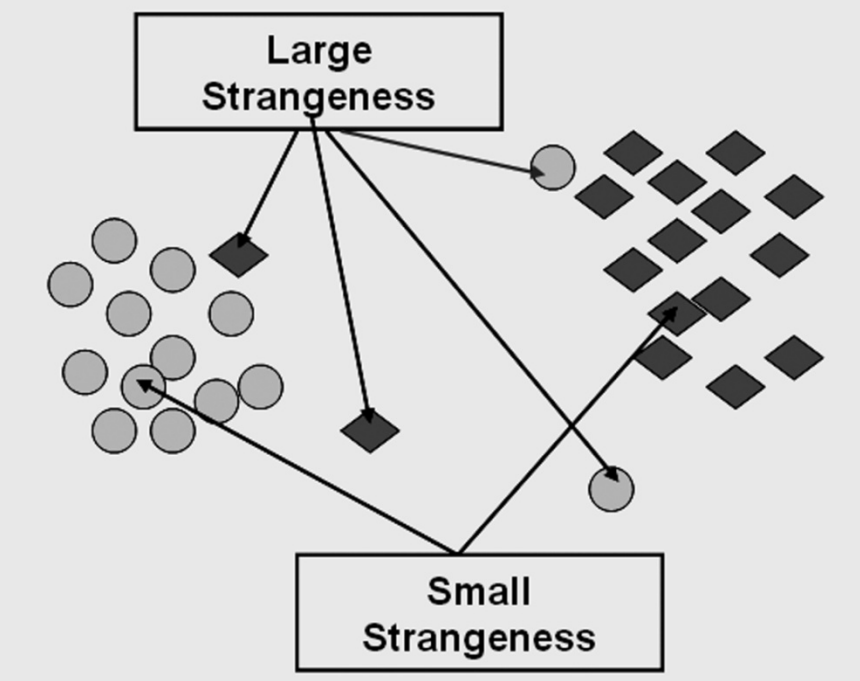
... and strangeness measure

$$\alpha_i$$

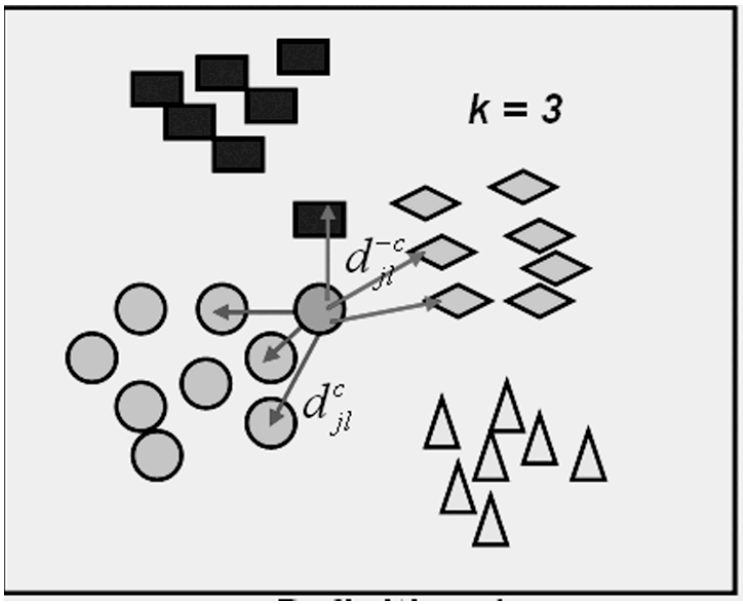
a measure on how “strange” a data point (xi, yi) is compared to the rest in the data points in sequence S

Conformal Prediction

Non-conformity measures: Example - kNN

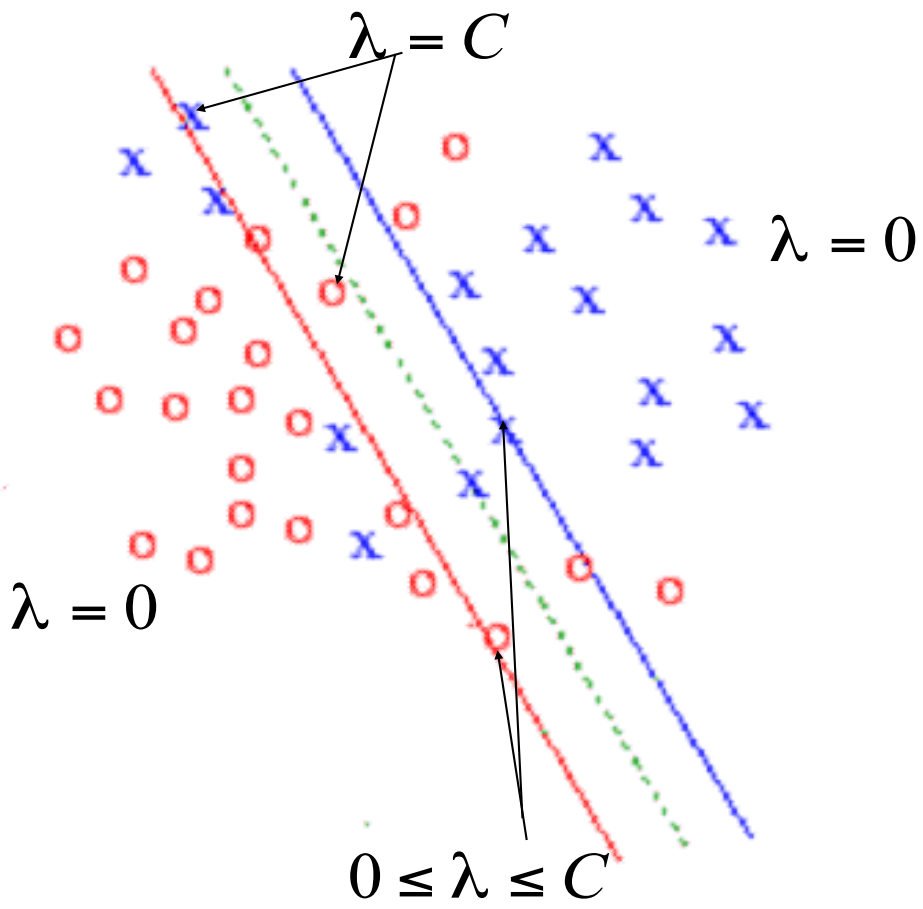


$$\alpha_i = \frac{\sum_{j=1}^{k} d_{ij}^{y}}{\sum_{j=1}^{k} d_{ij}^{\neg y}}$$



Conformal Prediction

Non-conformity measures: Example - SVM



Conformal Prediction

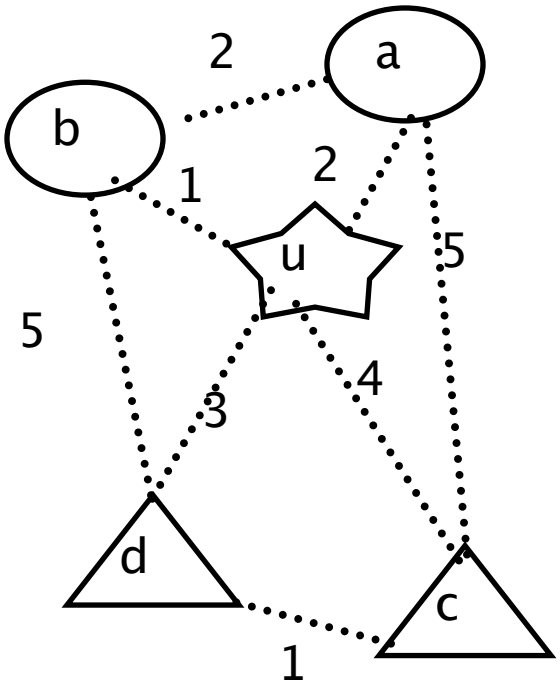
Non-Conformity Measures

Classifier	Non-conformity measure	Description
k-NN	$\frac{\sum_{j=1}^{k} D_{ij}^{y}}{\sum_{j=1}^{k} D_{ij}^{-y}}$	Ratio of the sum of the distances to the k nearest neighbors belonging to the same class as the hypothesis y, and the sum of the distances to the k nearest neighbors belonging to all other classes [38] [57] [58].
Support Vector Machines	Lagrange multipliers, or $e^{-ad_i^m}$	A suitable function of the distance of a data point from the hyperplane [38] [59] [60] [61].
Neural networks	$\frac{\sum_{y' \in Y: y' \neq y} o_{y'}}{o_y + \gamma}$	Ratio of the sum of the output values of all output neurons except the winning neuron and the output value of the winning neuron itself. γ is a parameter that can be varied [38] [62] [49] [63] [64].
Logistic regression	$\begin{cases} 1+\exp^{-w.x} & ,y=1 \\ 1+\exp^{w.x} & ,y=0 \end{cases}$	Reciprocal of the estimated probability of the observed y given the observed x for a given data instance. w is the weight vector typically computed using Maximum Likelihood Estimation [38].
Boosting	$\sum_{t=1}^{T} \alpha_t B_t(x,y)$	Weighted sum of the individual non-conformity measures of each of the weak classifiers B_t, and α_t are the weights learnt by the boosting algorithm [38]
Random forests	$\frac{out_{raw} - \overline{out_{raw}}}{\sigma}$, where $out_{raw}(i) = \frac{nsample}{\overline{p(i)}}$, and $\overline{p(i)} = \sum_j \lvert prox(i,j) \rvert^2$	Scaled outlier measure of an observed x with respect to label $y \in Y$, and other data instances belonging to the same class [65]. $nsample$ is the number of samples in the class under consideration, and $prox(i,j)$ is the similarity between two data instances in a random forest.

Conformal Prediction

Illustration of p-values and prediction

Let u = o,
$\alpha_a = 2/5$
$\alpha_b = 1/5$
$\alpha_c = 1/4$
$\alpha_d = 1/3$
$\alpha_u = 1/3$
$p_{u=o} = 3/5$



Let u = Δ,
$\alpha_a = 2/2$
$\alpha_b = 2/1$
$\alpha_c = 1/5$
$\alpha_d = 1/5$
$\alpha_u = 3/1$
$P_{u=\Delta} = 1/5$

Conformal Predictors

A conformal predictor maps a data sequence *S* and a new object *x* and each confidence level 1-ε in $(0,1)$ to the prediction set

$$\Gamma^{\varepsilon}(S,x) = \{y \in Y : p_y > \varepsilon\}$$

Conformal Predictors

Region Predictors

A "prediction set" $\Gamma^{\varepsilon} \subseteq Y$ consisting of the set of labels deemed possible at the confidence level 1 - ε.

Consider the nested prediction sets:

$$\Gamma^{\varepsilon_1} \subseteq \Gamma^{\varepsilon_2} \text{ when } \varepsilon_1 \geq \varepsilon_2$$

Conformal Predictors

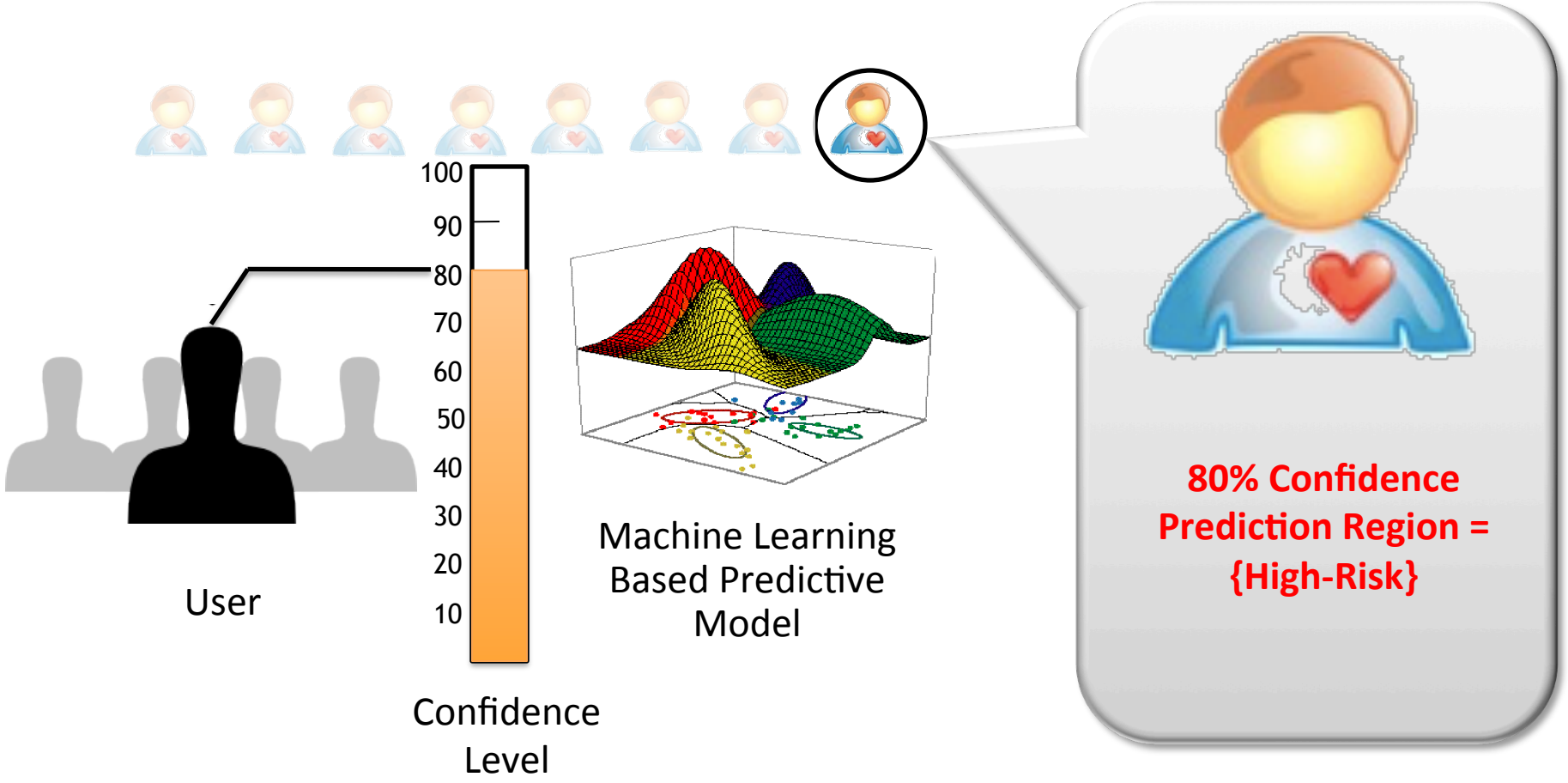
Illustration of Region Predictor

$p_{l=1}$= 0.3, $p_{l=2}$= 0.2, $p_{l=3}$= 0.7, $p_{l=4}$= 0.9, $p_{l=5}$= 0.4,
$p_{l=6}$= 0.6, $p_{l=7}$= 0.7, $p_{l=8}$= 0.8, $p_{l=9}$= 0.5, $p_{l=0}$= 0.8.

$$\Gamma^{0.85} = \{4\},$$
$$\Gamma^{0.75} = \{4,8,0\},$$
$$\Gamma^{0.65} = \{4,8,0,3,7\},$$
$$\Gamma^{0.55} = \{4,8,0,3,7,6\}.$$

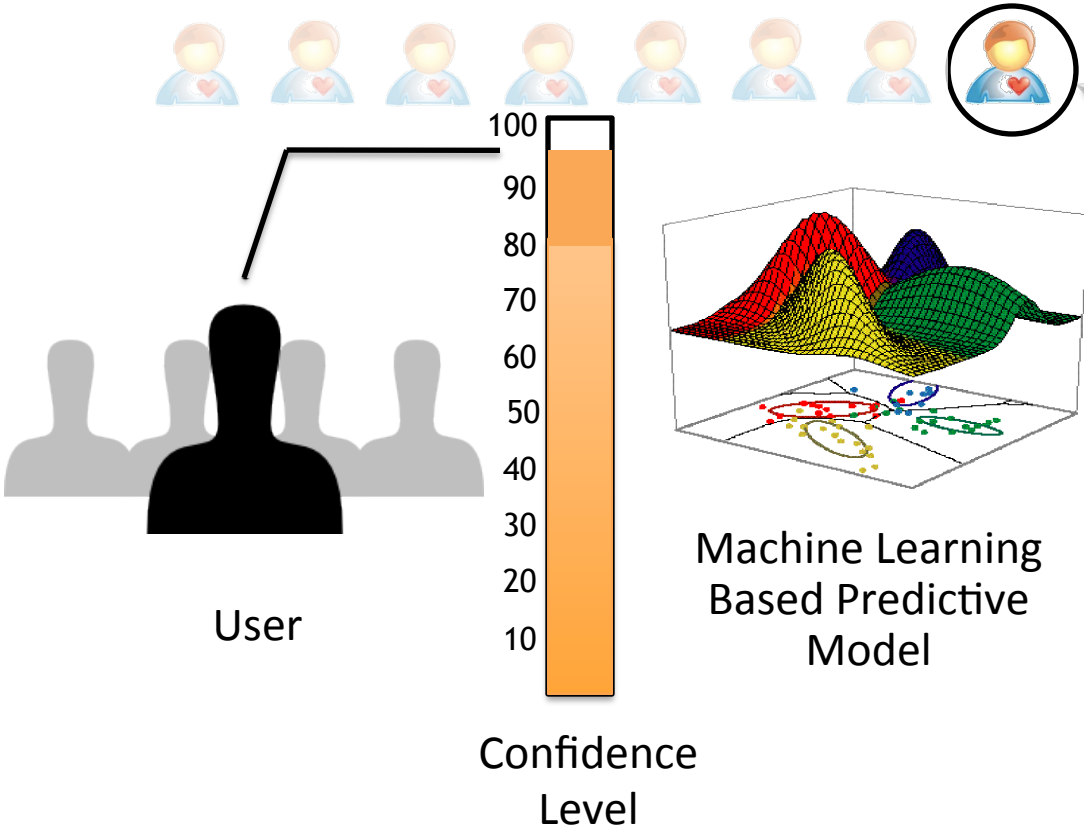
Conformal Prediction

An Illustration



Conformal Prediction

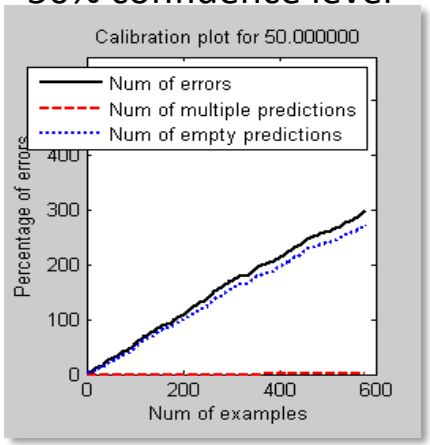
An Illustration



Conformal Prediction

Empirical Performance

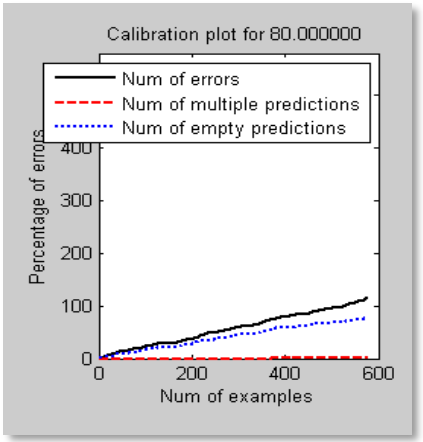
50% confidence level



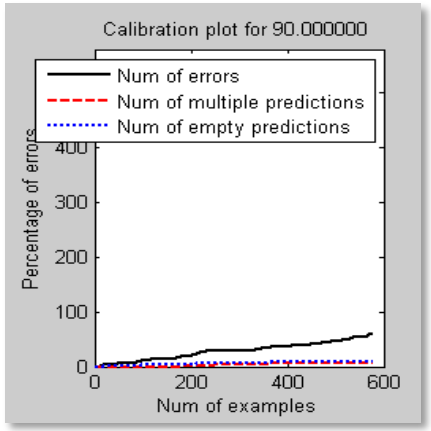
60% confidence level

70% confidence level

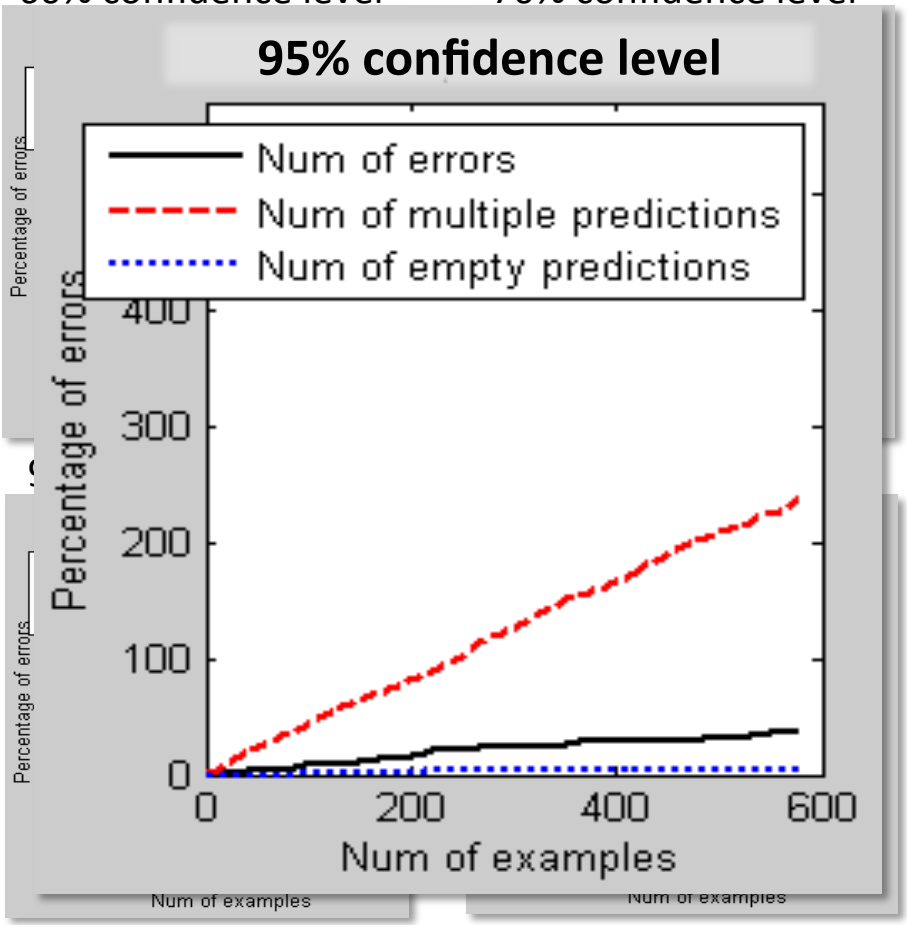
80% confidence level



90% confidence level



95% confidence level



Conformal Prediction

Summary

Randomness (Algorithmic) ⇔ Hypothesis Testing (Statistical)

↓

Non-conformal Measure (Strangeness)

↓

P-values (Prediction)

↙ ↘

Classification Regression

Conformal Predictions Framework

Classification Tasks

Compute non-conformity scores

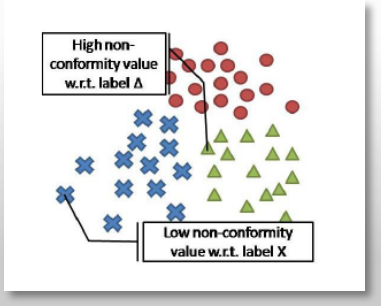
Compute p-value for every hypothesis

Output the hypotheses whose p-values satisfy the confidence level

Can be defined suitably for any classifier

$$\alpha_i^y = \frac{\sum_{j=1}^{k} D_{ij}^{y}}{\sum_{j=1}^{k} D_{ij}^{-y}}$$

K-NN



$$p(\alpha_{n+1}^{y_p}) = \frac{count\left\{i : \alpha_i^{y_p} \geq \alpha_{n+1}^{y_p}\right\}}{n+1}$$

Non-conformity measure

Conformal prediction regions with confidence level 1- ε

$$\Gamma_{1-\varepsilon} = \{y_i : P^{y_i} > \varepsilon, y_i \in Y\}$$

Conformal Prediction

Prediction Goals

- A predictor is **valid** (or well-calibrated) if its frequency of prediction error does not exceed ε at a chosen confidence level 1- ε in the long run.
- A predictor is **efficient** (or perform well) if the prediction set (or region) is as small as possible (tight).

Conformal Prediction

Criteria for Efficiency

- ε- free
  - Sum of p-values (small value preferred)
  - Sum of p-values apart from the largest one.
  - Second largest p-value.
- ε- dependent
  - Number of labels (small value preferred)
  - Total number of multiple label prediction set from the test sequence.
  - Average number of test examples having multiple predictions.

Working Paper 11. "Criteria of efficiency for conformal prediction" by Vladimir Vovk, Valentina Fedorova, Ilia Nouretdinov, and Alex Gammerman.

Conformal Prediction

Observations

- A data sequence has a corresponding sequence of non-conformity scores, in such a way that interchanging any 2 data points leads to the interchange of their corresponding non-conformity score.
- The data sequence is “represented” by the non-conformity score sequence

Conformal Prediction

Exchangeability

The variables $z_{1,\ldots,}\ z_N$ are exchangeable if for every permutation τ of the integer 1,...,N, the variables $w_{1,\ldots,}\ w_N$ where

$$w_i = z_{\tau(i)}$$

have the same joint probability distribution as $z_{1,\ldots,}\ z_N$.

Conformal Prediction

Examples of Exchangeability Models

- Identically and Independently Distributed Model (IID)
- Sampling without replacement from a finite population (not independent).

Conformal Prediction

Why is it useful?

- Lemmas
  - Lemma 1: The sequences of non-conformal scores for the data generated from a source satisfying the exchangeability assumption is exchangeable.
  - Lemma 2: P-values from the conformal predictor on data generated from a source satisfying the exchangeability assumption are independent and uniformly distributed on [0,1].
- Theorem
  - A transductive conformal predictor is valid in the sense that the probability of error that a correct label

$$y^c \notin \Gamma^\varepsilon(S, x)$$

at confidence level 1-ε never exceeds ε, with the error at successive prediction trials not independent (conservative), and the error frequency is close to ε in the long run.

Vovk, Online Confidence Machines are Well-Calibrated, 2002.

Conformal Prediction

Prediction Outputs

- Point Predictors
  - Confidence: 1 - second largest randomness level (aka p-value)
  - Credibility: The largest randomness level over all labels.
  - Low credibility implies either the training set is non-random (biased) or the test object is not representative of the training set.
- Region Predictors
  - Nested prediction sets

Conformal Prediction

Theorem

When a small amount of randomization is added to the prediction process (smoothed), the previous theorem holds true. Moreover, the error at successive prediction trials is independent and the long run error frequency converges to ε (well calibrated).

$$P_y = \frac{|\{i:\alpha_i > \alpha_{n+1}\}| + \eta_{n+1}|\{i: \alpha_i = \alpha_{n+1}\}|}{n+1}$$

where i ranges over {1,...,n+1} and $\eta \in [0,1]$ is generated randomly from U[0,1].

Conformal Prediction

Theorem

Suppose the predictor receives a feedback at the end of step $n_1, n_2, \ldots$ such that $n_1 < n_2 \ldots$ and a feedback is the label of one of the objects the predictor has seen and predicted. The probability of error that a correct label

$$y^c \notin \Gamma^\varepsilon(S, x)$$

at confidence level 1-ε approaches ε in probability, if and only if $n_k/n_{k-1} \to 1$ as k approaches infinity. (Lazy, Slow Teacher)

Conformal Prediction

Inductive Conformal Predictor

- Divide the training set *T* into: a 'proper training set' *Ts* and a 'calibration set' *C*
- Construct a prediction rule R from *Ts*
- Compute the non-conformity score for all examples in *C* using some measure of difference $\Delta(y, y')$ such that

$$\alpha_i = \Delta(y_i, R(x_i))$$

Conformal Prediction

Inductive Conformal Predictor

For each test object x_j, do the following:

1. For every possible label $y \in Y$, compute $\alpha_j = \Delta(y, F(x_j))$ and the P-value

$$P_y = \frac{|\{i \in I : \alpha_i \geq \alpha_j\}|}{|T| - |Ts| + 1}$$

where *I* is the index set for *C*.

2. Output the prediction set

$$\Gamma^{\varepsilon}(Ts \cup C, x) = \{y \in Y : p_y > \varepsilon\}$$

Conformal Prediction

Inductive Conformal Predictor

- Advantage:
  - computational efficiency
- Disadvantages:
  - possible loss of prediction efficiency
- Theorem
  - An inductive conformal predictor is valid in the sense that the probability of error that a correct label

$$y^c \notin \Gamma^\varepsilon(Ts \cup C, x)$$

  - at confidence level 1-ε never exceeds ε for each test object.

Conformal Prediction

Recent Developments

- Venn-Abers predictors (UAI 2014)
- Efficiency of conformalized Bayesian ridge regression (COLT 2014)
- Cross-Conformal predictors (Annals of Mathematics and AI)
- Plug-in martingales for testing exchangeability online (ICML 2012)
- etc.

Conformal Prediction

Adaptations

- Active Learning
- Model Selection
  - Define the critical threshold that creates one label in prediction set $\epsilon_{crit} = \min_{k \in K} \min_{y \in \{-1,+1\}} \frac{\left|\{\forall j : y_j^v f_k(x_j^v) \leq y f_k(x)\}\right|}{n}$
- Feature Selection
  - Strangeness Minimization: $\tilde{A}(z_i, (z_1, ..., z_n), S) = \sum_{j \in S} \phi(z_i, j, (z_1, ..., z_n))$
- Anomaly Detection
  - Strangeness Based Outlier Detection

Conformal Prediction

Active Learning

If the two highest p-values (p_+ and p_-) for a data point x_{n+1} (say, assigned label + and -) are too close in value, there is high uncertainty / ambiguity leading to "low confidence" in whether the data point should be labeled + or -.

"Closeness'':

$$I(x_{n+1}) = | p_+ - p_- |$$

Selection Criterion:

$$I(x_{n+1}) < \varepsilon$$

Stopping Criterion:

$$\left| error(x_1, \ldots x_{n+1}) - error(x_1, \ldots x_n) \right| < \gamma$$

Conformal Prediction

Active Learning

- Relationship with KL divergence (expected discrimination information between H_0 and H_1)

$$KL(f(Z \mid H_0) \| f(Z \mid H_{1|})) \leq \log P(H_0 \mid Z) - \log P(H_1 \mid Z)$$

- Relationship to QBC
  - Shannon entropy $I(x_{n+1}) = \log N - KL(f(Z \mid H_0) \| f(Z \mid H_{1|}))$
  - Threshold ε = upper bound for discrimination information (KLD) = lower bound for information gain in QBC

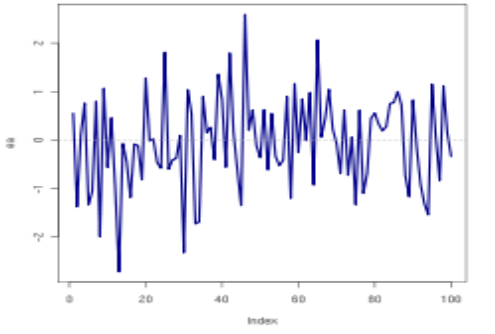
Thank you for your patience and attention!

Questions?

Conformal Prediction

Assumptions and Impact

- i.i.d
  - How to test?
- Exchangeability
  - Weaker than i.i.d.
- Martingales
  - $E(X_{n+1} \mid X_1,\ldots,X_n) = X_n$



Randomized Power Martingale

$$M_n^{\epsilon} = \prod_{i=1}^{n} \epsilon\, p_i^{\epsilon-1}$$

Uses the CP framework to test the exchangeability of a given sequence

Assumptions

Eg. Bag-of-words model

Exchangeability

$$P(z_1, z_2, \ldots) \in Z^\infty : (z_1, z_2, z_n) \in E = P(z_1, z_2, \ldots) \in Z^\infty : (\pi(z_1), \pi(z_2), \ldots, \pi(z_n)) \in E$$

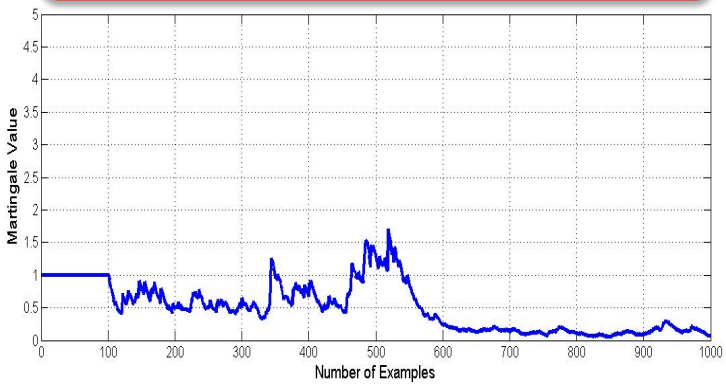
Test

Randomized Power Martingale

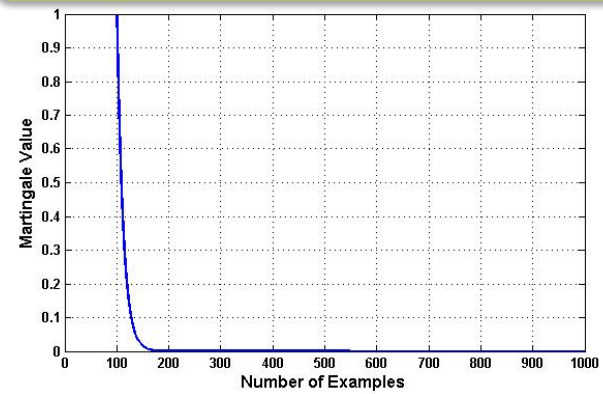
$$M_n^\varepsilon = \prod_{i=1}^{n} \varepsilon p_i^{\varepsilon - 1}$$

Testing Exchangeability of Datasets

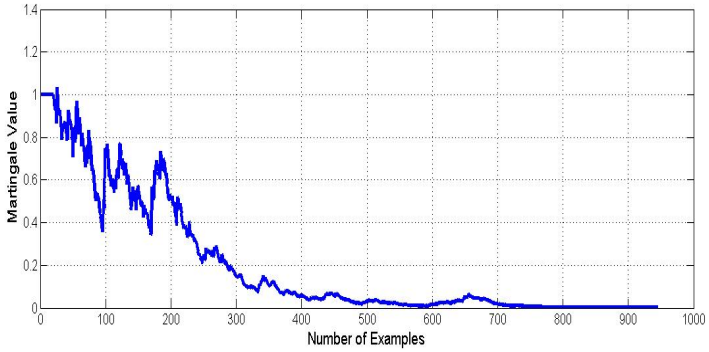
Cardiac Decision Support



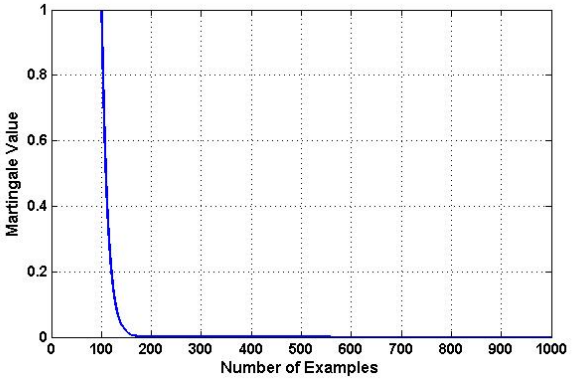
Head Pose Estimation



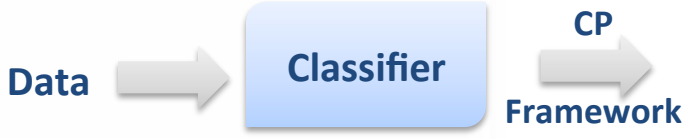
Multimodal Person Recognition



Saliency Prediction



Conformal Predictors for Regression



Class A	Class B	Class C	Class D	Class E
p_A	p_B	p_C	p_D	p_E

How to evaluate every hypothesis in an interval of values?

Conformal Predictors for Regression

Compute non-conformity scores

Compute p-value for every hypothesis

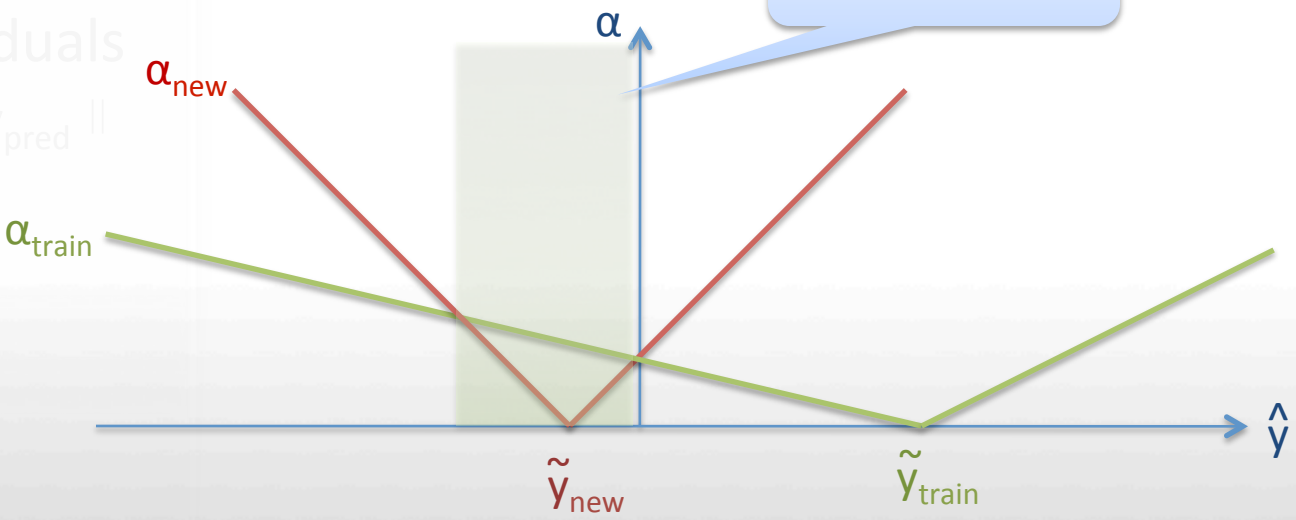
Output the hypotheses whose p-values satisfy the confidence level

$$p(\alpha_{n+1}^{y_p}) = \frac{count\left\{i : \alpha_i^{y_p} \geq \alpha_{n+1}^{y_p}\right\}}{n+1}$$

Constant p-value

Error residuals

$|| y_{actual} - y_{pred} ||$



Conformal Predictors for Regression

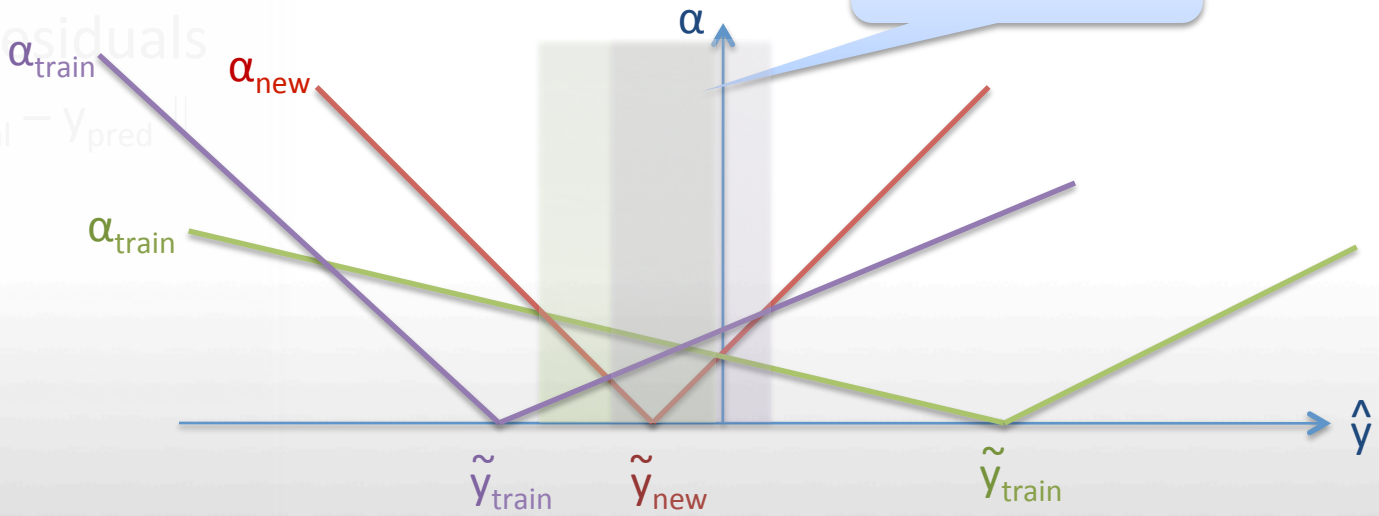
Compute non-conformity scores

Compute p-value for every hypothesis

Output the hypotheses whose p-values satisfy the confidence level

$$p(\alpha_{n+1}^{y_p}) = \frac{count\left\{i : \alpha_i^{y_p} \geq \alpha_{n+1}^{y_p}\right\}}{m}$$

Constant p-value



Error residuals

$|| y_{actual} - y_{pred} ||$

Conformal Predictors for Regression

Compute non-conformity scores

Compute p-value for every hypothesis

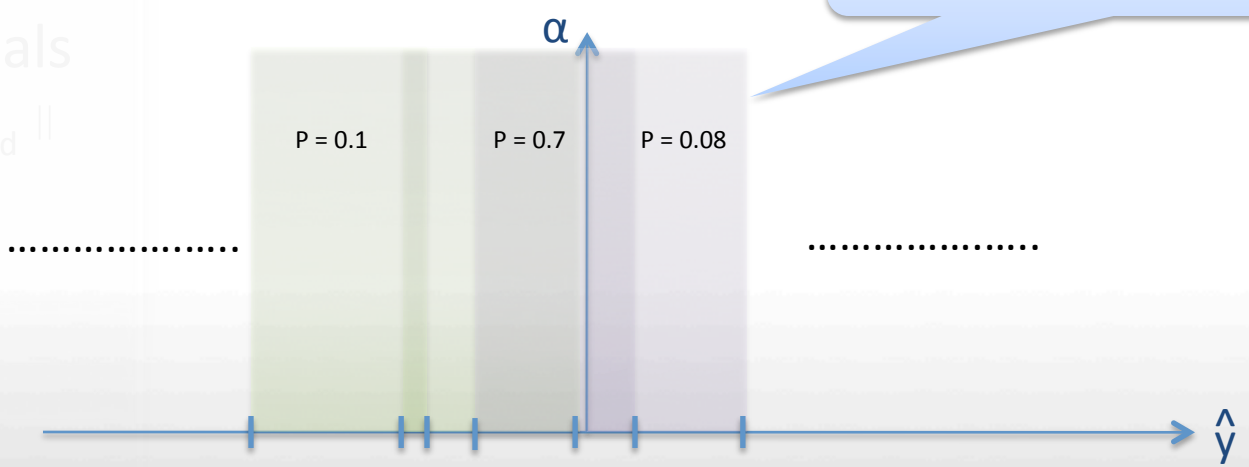
Output the hypotheses whose p-values satisfy the confidence level

$$p(\alpha_{n+1}^{y_p}) = \frac{count\left\{i : \alpha_i^{y_p} \geq \alpha_{n+1}^{y_p}\right\}}{m}$$

Constant p-value within each interval

Error residuals

$$\| y_{actual} - y_{pred} \|$$



Conformal Predictors for Regression

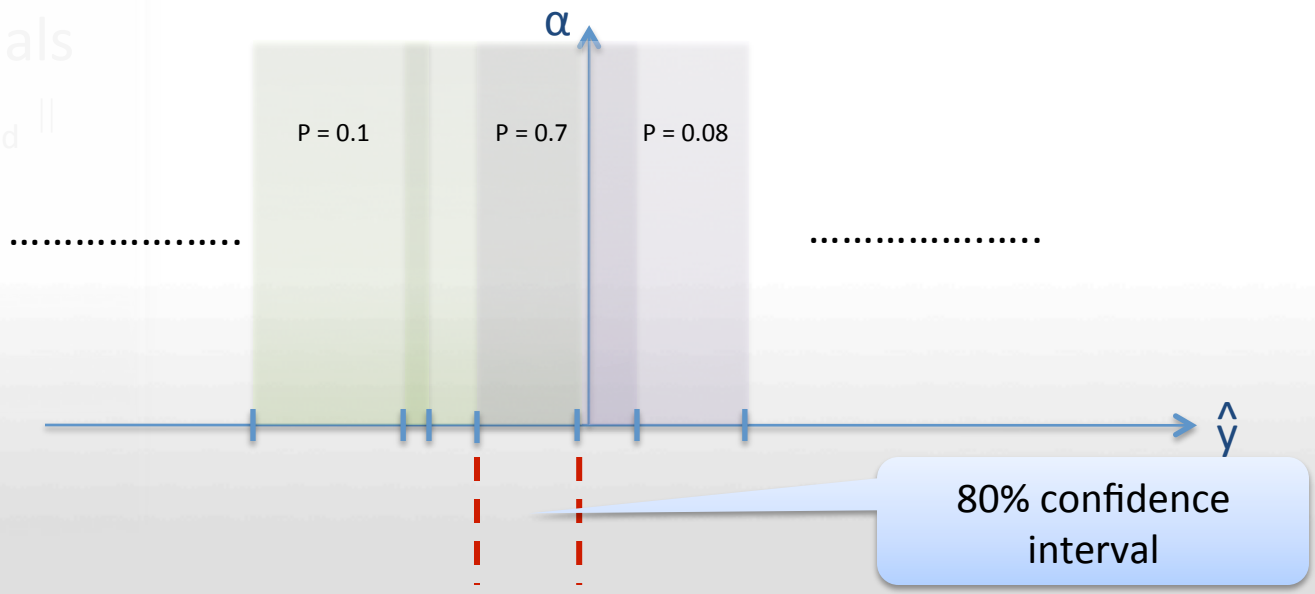
Compute non-conformity scores

Compute p-value for every hypothesis

Output the hypotheses whose p-values satisfy the confidence level

Error residuals

$|| y_{actual} - y_{pred} ||$



Conformal Prediction in Regression

Algorithm 2 Conformal Predictors for Regression

Require: Training set $T = \{(x_1, y_1), ..., (x_n, y_n)\}$, new example x_{n+1}, confidence level r, $X = x_1, x_2, \ldots, x_{n+1}$

1: Calculate $C = I - X(X'X + \alpha I)^{-1}X'$ (for ridge regression).
2: Let $A = C(y_1, y_2, \ldots, y_n, 0)' = (a_1, a_2, \ldots, a_{n+1})$
3: Let $B = C(0, 0, \ldots, 0, 1)' = (b_1, b_2, \ldots, b_{n+1})$
4: **for** $i = 1$ to $n + 1$, do **do**
5: Calculate u_i and v_i.
If $b_i \neq b_{n+1}$, then $u_i = min(\frac{a_i - a_{n+1}}{b_{n+1} - b_i}, \frac{-(a_i + a_{n+1})}{b_{n+1} + b_i})$; $v_i = max(\frac{a_i - a_{n+1}}{b_{n+1} - b_i}, \frac{-(a_i + a_{n+1})}{b_{n+1} + b_i})$
If $b_i = b_{n+1}$, then $u_i = v_i = \frac{-(a_i + a_{n+1})}{2b_i}$.
6: **end for**
7: **for** $i = 1$ to $n + 1$, do **do**
8: Compute S_i according to Equation 14 below.
9: **end for**
10: Sort $(-\infty, u_1, u_2, \ldots, u_{n+1}, v_1, \ldots, v_{n+1}, \infty)$ in ascending order, obtaining $\hat{y_0}, \ldots, \hat{y_{2n+3}}$
11: Output $\cup_i [\hat{y_i}, \hat{y_{i+1}}]$, such that $N(\hat{y_i} > r$, where $N(y_i) = \#S_j : [\hat{y_i}, \hat{y_{i+1}}] \subseteq S_j$, where $i = 0, \ldots, 2n$, and $j = 1, \ldots, n + 1$.

$$S_i = \begin{cases} [u_i, v_i] & \text{if } b_{n+1} > b_i \\ (-\infty, u_i] \cup [v_i, \infty) & \text{if } b_{n+1} < b_i \\ [u_i, \infty) & \text{if } b_{n+1} = b_i > 0 \text{ and } a_{n+1} < a_i \\ (-\infty, v_i] & \text{if } b_{n+1} = b_i > 0 \text{ and } a_{n+1} > a_i \\ \Re & \text{if } b_{n+1} = b_i = 0 \text{ and } |a_{n+1}| \leq |a_i| \\ \Phi & \text{if } b_{n+1} = b_i = 0 \text{ and } |a_{n+1}| > |a_i| \end{cases}$$