Entropic Evaluation of Classification
A hands-on, get-dirty introduction

Website: `http://gpm.webs.tsc.uc3m.es/resources/tutorials/ettutorial/`

Francisco J. Valverde-Albacete[1] Carmen Peláez-Moreno[1]

[1] Departamento de Teoría de la Señal y de las Comunicaciones, Universidad Carlos III de Madrid, Spain

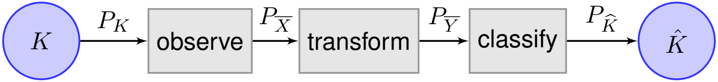
08/07/2018 / IJCNN 2018 - Rio de Janeiro, Brazil

Outline

Outline

Take Home Lessons

- Supervised Classification as an Information Channel



- It is possible to *assess* this system by entropic means. We have
  - mathematical tools, the balance equations and,
  - data exploratory tools, the entropy triangles and diagrams.
- You can study the system,
  - end-to-end (CBET), or
  - its sources (SMET)
  - its subsystems (CMET).
- Each tool provides different insight into the technologies implementing each subsystem.

From Barra da Tijuca to Copacabana... A soggy story

- Bob has been coming to Rio on and offs for quite some time.
- His favourite haunt is Copacabana, but he only gets hotels in Barra da Tijuca.
- In the morning he looks through the window and decides whether it will be **sunny, cloudy or rainy**, and dresses accordingly.
- Then he goes to Copacabana/Ipanema an tries to enjoy his stay in Rio, sometimes soaking wet in his Havaianas flip flops, sometimes roasting in his waterproof.
- How good of a weather predictor is Bob (or for that matter Alice, Carol, Dan, Eva and Frank)?
- Suppose they have had ample time to hone their weather-predicting skills.

Tourists as Weathervanes

- Here are the *confusion matrices* of their decision:
  - K (rows): the actual weather in Copacabana.
  - $\widehat{K}$ (columns): what X deems the weather is going to be.

$$a = \begin{bmatrix} 15 & 0 & 5 \\ 0 & 15 & 5 \\ 0 & 0 & 20 \end{bmatrix} \qquad b = \begin{bmatrix} 16 & 2 & 2 \\ 2 & 16 & 2 \\ 1 & 1 & 18 \end{bmatrix} \qquad c = \begin{bmatrix} 1 & 0 & 4 \\ 0 & 1 & 4 \\ 1 & 1 & 48 \end{bmatrix}$$

$$d = \begin{bmatrix} 15 & 0 & 0 \\ 0 & 18 & 0 \\ 0 & 0 & 27 \end{bmatrix} \qquad e = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 57 \end{bmatrix} \qquad f = \begin{bmatrix} 0 & 0 & 5 \\ 0 & 0 & 5 \\ 0 & 0 & 50 \end{bmatrix}$$

Their Representation in the CBET

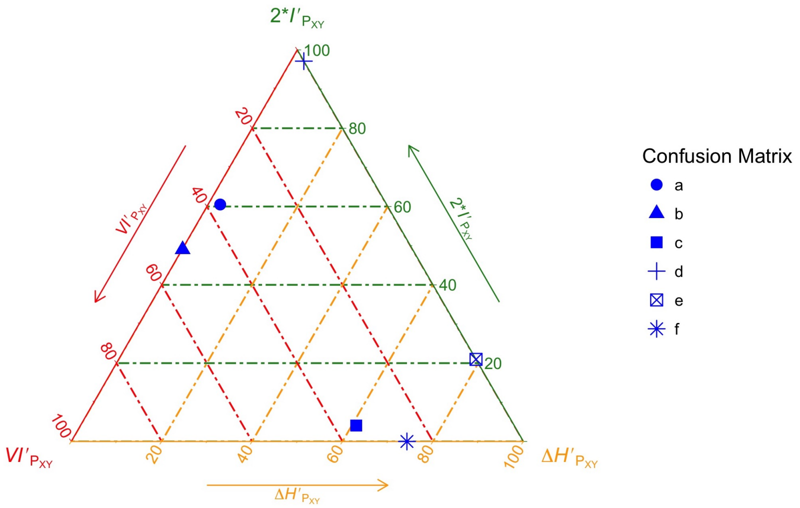
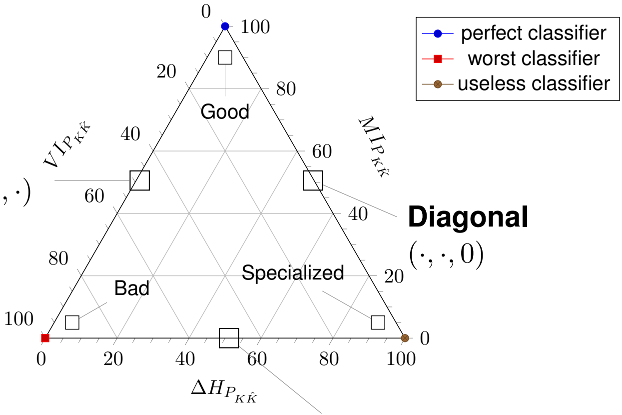


Figure: Assessment of Tourists as Weathervanes by entropic means

Interpretation of the CBET for Classification

Balanced

$$\begin{bmatrix} U_K = P_K \\ U_{\widehat{K}} = P_{\widehat{K}} \end{bmatrix} \Leftrightarrow (0, \cdot, \cdot)$$



Diagonal

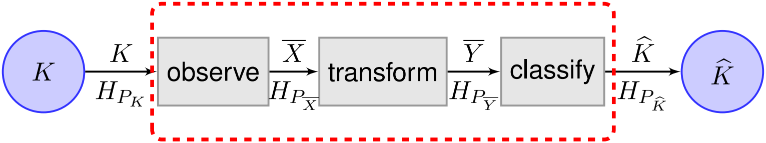
$(\cdot, \cdot, 0)$

Random

$$P_K \perp P_{\widehat{K}} \Leftrightarrow (\cdot, 0, \cdot)$$

End-to-End Evaluation

- For end-to-end evaluation we ignore all details within the outer hatched block:



- The whole chain is considered here as a single-input single-output (SISO) block with joint distribution $(K, \hat{K}) \sim P_{K\hat{K}}$

The proposal

In this tutorial we show you how to tackle the evaluation of such a system based on a simple heuristic.

The Transmitted Information Heuristic

The more “information´´ is sent from K to $\widehat{K}$
the better the classification system is.

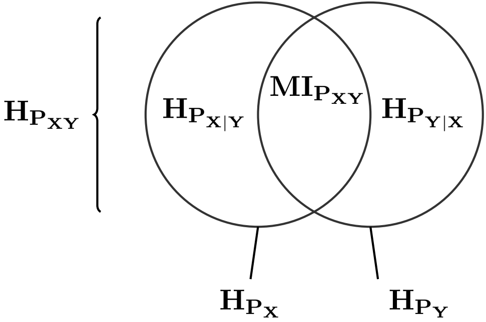
Outline

Conventional Entropy Diagram for Two Variables

- Recall the conventional entropy diagram of distribution P_{XY}.

$$H_{P_{XY}} = H_{P_{X|Y}} + MI_{P_{XY}} + H_{P_{Y|X}} \quad (1)$$

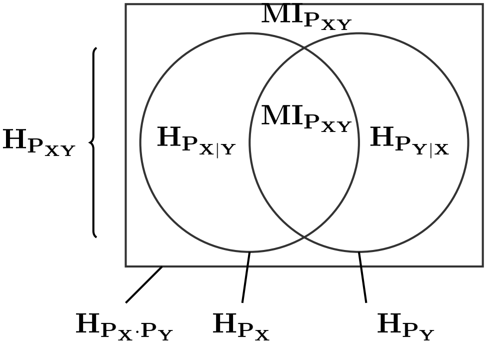
$$MI_{P_{XY}} = H_{P_X} - H_{P_{X|Y}} = H_{P_Y} - H_{P_{Y|X}} \quad (2)$$



But the Mutual Information appears once more...

- Consider a "virtual" distribution $Q_{XY} = P_X \cdot P_Y$ the we also have:

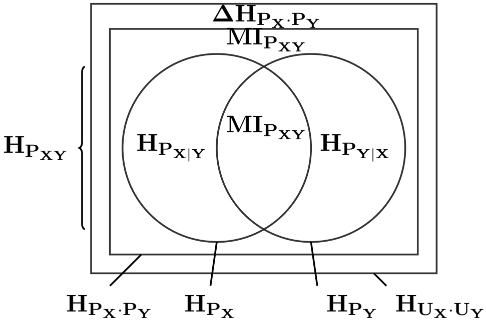
$$MI_{P_{XY}} = H_{P_X \cdot P_Y} - H_{P_{XY}} \tag{3}$$



Is this all of the Information around P_{XY}?

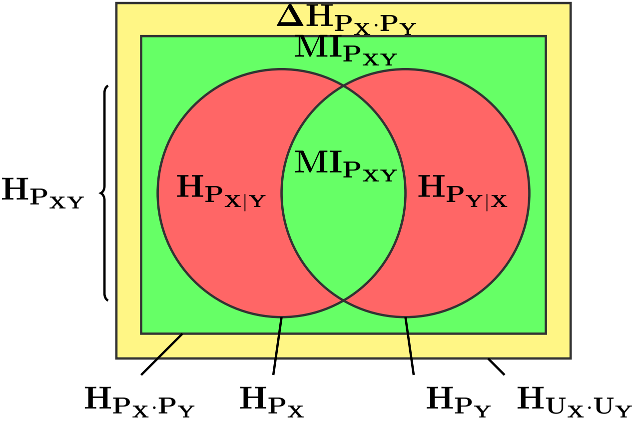
- NO! Consider the uniform joint distribution $U_{XY} = U_X \cdot U_Y$. Then:

$$H_{U_X \cdot U_Y} = H_{U_X} + H_{U_Y} \quad (4)$$



Extended Entropy Diagram for Two Variables

Our method is based in an exhaustive analysis of the information in $P_{K\widehat{K}}$.



Definition of Areas

- Consider two new distributions related to P_{XY}:
  - $P_X \cdot P_Y$, is defined on the same domain, and has similar marginals, but independent.
  - $U_X \cdot U_Y$, is defined on the same domain, with uniform marginals (hence independent).
- Then we may define:

  Non-realized entropy

$$\Delta H_{P_X \cdot P_Y} = H_{U_X \cdot U_Y} - H_{P_X \cdot P_Y} \tag{5}$$

  Mutual Information

$$MI_{P_{XY}} = H_{P_X \cdot P_Y} - H_{P_{XY}} \tag{6}$$

$$MI_{P_{XY}} = H_{P_X} - H_{P_{X|Y}} = H_{P_Y} - H_{P_{Y|X}} \tag{7}$$

  Variation of Information = Remanent Entropy = Residual Entropy

$$VI_{P_{XY}} = H_{P_{X|Y}} + H_{P_{Y|X}} \,. \tag{8}$$

The Total Entropy Balance Equation

The entropies concerning two variables can be partitioned as:

$$H_{U_X \cdot U_Y} = \Delta H_{P_X \cdot P_Y} + 2 * MI_{P_{XY}} + VI_{P_{XY}} \quad (9)$$
$$0 \leq \Delta H_{P_X \cdot P_Y}, MI_{P_{XY}}, VI_{P_{XY}} \leq H_{U_X \cdot U_Y}$$

Proof

$$\begin{aligned} H_{U_X \cdot U_Y} &= \Delta H_{P_X \cdot P_Y} + H_{P_X \cdot P_Y} \\ H_{P_X \cdot P_Y} &= H_{P_X} + H_{P_Y} \\ H_{P_X} &= H_{P_{X|Y}} + MI_{P_{XY}} \\ H_{P_Y} &= H_{P_{Y|X}} + MI_{P_{XY}} \\ H_{P_{X|Y}} + H_{P_{Y|X}} &= VI_{P_{XY}} \end{aligned}$$

Add up and cancel.

Outline

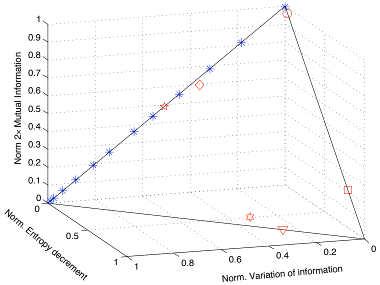
Normalization

Looking for a cardinality-independent representation, we carry out a normalization by $H_{U_{XY}}$.

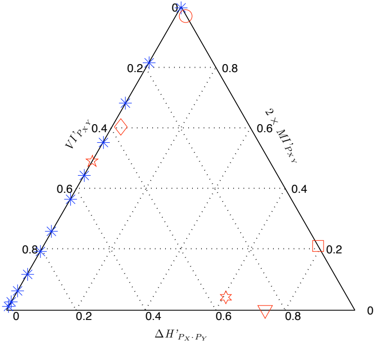
$$\begin{aligned} 1 &= \Delta H'_{P_X \cdot P_Y} + 2MI'_{P_{XY}} + VI'_{P_{XY}} \\ 0 &\leq \Delta H'_{P_X \cdot P_Y}, 2MI'_{P_{XY}}, VI'_{P_{XY}} \leq 1\,. \end{aligned} \tag{10}$$

- This is the equation of a **2-simplex** in $[\Delta H'_{P_X \cdot P_Y}, 2MI'_{P_{XY}}, VI'_{P_{XY}}]$ space!

The 2-simplex in $[\Delta H'_{P_X \cdot P_Y}, 2MI'_{P_{XY}}, VI'_{P_{XY}}]$ space



(a) The 2-simplex in three-dimensional, normalized entropy space $\Delta H'_{P_X \cdot P_Y} \times VI'_{P_{XY}} \times 2MI'_{P_{XY}}$.



(b) De Finetti entropy diagram or entropy triangle, a projection of the 2-simplex onto a two-dimensional space.

Examples of Confusion Matrices

$$a = \begin{bmatrix} 15 & 0 & 5 \\ 0 & 15 & 5 \\ 0 & 0 & 20 \end{bmatrix} \quad b = \begin{bmatrix} 16 & 2 & 2 \\ 2 & 16 & 2 \\ 1 & 1 & 18 \end{bmatrix} \quad c = \begin{bmatrix} 1 & 0 & 4 \\ 0 & 1 & 4 \\ 1 & 1 & 48 \end{bmatrix}$$

$$d = \begin{bmatrix} 15 & 0 & 0 \\ 0 & 18 & 0 \\ 0 & 0 & 27 \end{bmatrix} \quad e = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 57 \end{bmatrix} \quad f = \begin{bmatrix} 0 & 0 & 5 \\ 0 & 0 & 5 \\ 0 & 0 & 50 \end{bmatrix}$$

Figure: **Examples of synthetic confusion matrices with varied behavior**: a, b and c from [3], d a matrix whose marginals tend towards uniformity, e a matrix whose marginals tend to Kronecker's delta and f the confusion matrix of a majority classifier.

Their Representation in the CBET

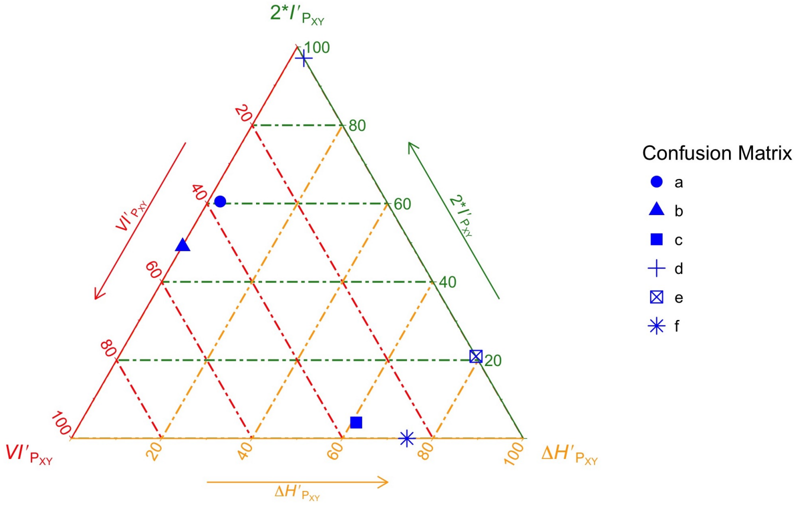
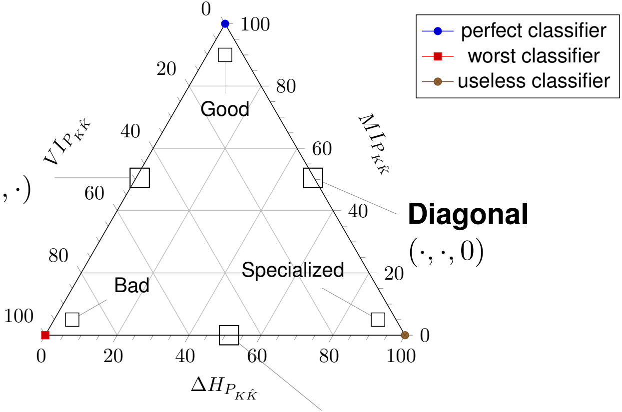


Figure: Representation of the previous confusion matrices in the CBET

Interpretation of the CBET for Classification

Balanced

$$\begin{bmatrix} U_K = P_K \\ U_{\widehat{K}} = P_{\widehat{K}} \end{bmatrix} \Leftrightarrow (0, \cdot, \cdot)$$



Diagonal

$(\cdot, \cdot, 0)$

Random

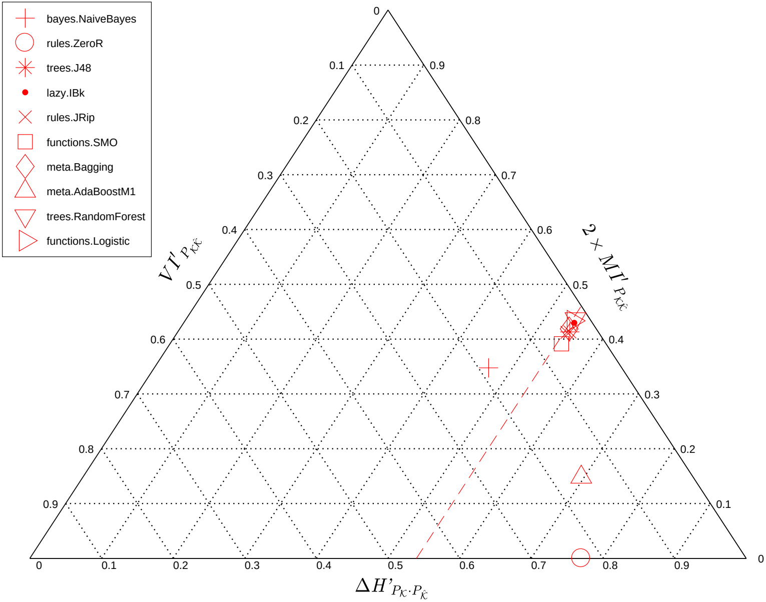
$P_K \perp P_{\widehat{K}} \Leftrightarrow (\cdot, 0, \cdot)$

Exploring one Database with many Classifierss

Consider the UCI anneal dataset as a classification task.

- This task has $k = 6$ classes, but it is highly imbalanced and has empty classes.
- We carry out the process with 10 classifiers provided by Weka and plot the confusion matrices on the CBET.
- This is the envisioned use for people who need to solve a specific task using classification.

Aggregated CBET for anneal

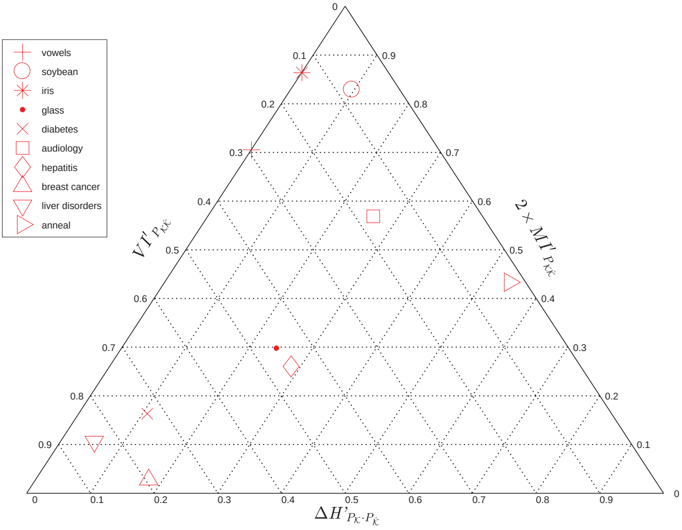


Exploring many Databases with one Classifier

Consider the *functions.logistic* classifier of Weka implementing multiclass logistic regression.

- We obtained confusion matrices for 10 different UCI tasks and plotted them in the ET.
- This is the type of use we envision for classification algorithm research and design.

Aggregated CBETs for Logistic Regression (in Weka)

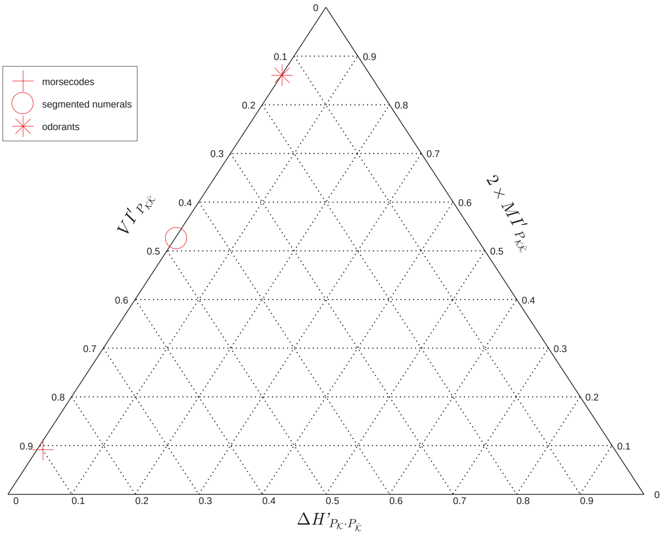


Analysis of Human Performance in Perceptual Classification

In the following examples we present the performance of humans in several perceptual tasks:

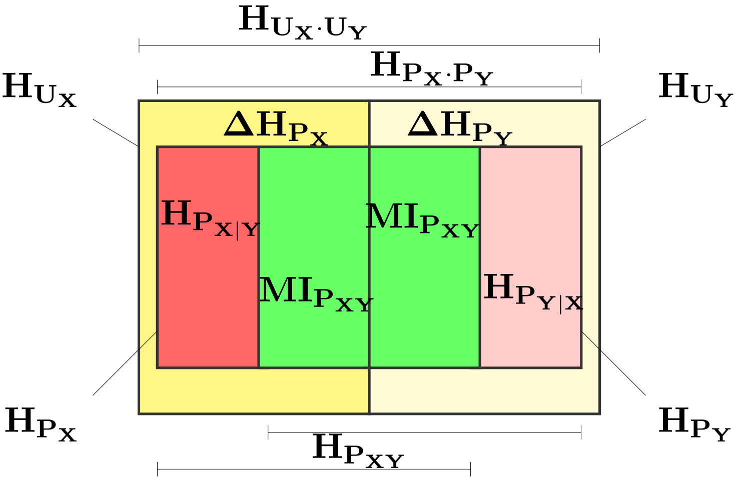
- **Visual distinction of segmented numeral digits in humans.** In this task from [1], the $k = 10$ seven-segment digits of digital displays were presented to testers under different conditions designed to guarantee that confusions would appear.
- **Acoustic-perceptual distinction of morse-codes in humans.** Likewise, in [2] human testers would listen to $k = 36$ Morse codes for English letters and digits and provide the transcodified symbol under tests conditions that induced confusions.
- **Odorant confusions in healthy humans.** In [8], $k = 10$ different odorants selected on the basis that they could be identified by a population were presented to 10 female and 10 male subjects and their confusions noted down.

Aggregated CBET in Perceptual Tasks



Outline

The I-diagram can be split in two



Splitting the Balance Equations

- This means the balance equation can be *split in two:*

$$H_{U_X} = \Delta H_{P_X} + MI_{P_{XY}} + H_{P_{X|Y}} \quad H_{U_Y} = \Delta H_{P_Y} + MI_{P_{XY}} + H_{P_{Y|X}}$$
$$0 \leq \Delta H_{P_X}, MI_{P_{XY}}, H_{P_{X|Y}} \leq H_{U_X} \quad 0 \leq \Delta H_{P_Y}, MI_{P_{XY}}, H_{P_{Y|X}} \leq H_{U_Y}$$

- And similarly normalized by H_{U_X} and H_{U_Y}:

$$1 = \Delta H'_{P_X} + MI'_{P_{XY}} + H'_{P_{X|Y}} \quad 1 = \Delta H'_{P_Y} + MI'_{P_{XY}} + H'_{P_{Y|X}}$$
$$0 \leq \Delta H'_{P_X}, MI'_{P_{XY}}, H'_{P_{X|Y}} \leq 1 \quad 0 \leq \Delta H'_{P_Y}, MI'_{P_{XY}}, H'_{P_{Y|X}} \leq 1$$

Splitting the Balance Equations

- This means the balance equation can be *split in two:*

$$H_{U_X} = \Delta H_{P_X} + MI_{P_{XY}} + H_{P_{X|Y}} \quad H_{U_Y} = \Delta H_{P_Y} + MI_{P_{XY}} + H_{P_{Y|X}}$$
$$0 \leq \Delta H_{P_X}, MI_{P_{XY}}, H_{P_{X|Y}} \leq H_{U_X} \quad 0 \leq \Delta H_{P_Y}, MI_{P_{XY}}, H_{P_{Y|X}} \leq H_{U}$$

- And similarly normalized by H_{U_X} and H_{U_Y}:

$$1 = \Delta H'_{P_X} + MI'_{P_{XY}} + H'_{P_{X|Y}} \quad 1 = \Delta H'_{P_Y} + MI'_{P_{XY}} + H'_{P_{Y|X}}$$
$$0 \leq \Delta H'_{P_X}, MI'_{P_{XY}}, H'_{P_{X|Y}} \leq 1 \quad 0 \leq \Delta H'_{P_Y}, MI'_{P_{XY}}, H'_{P_{Y|X}} \leq 1$$

Splitting the Entropy Triangle

- This means we can also split the CBET in two.
- These describe the *marginal fractions* of entropy when the normalization is done with H_{U_X} and H_{U_Y} respectively

$$F_X(P_{XY}) = [\Delta H'_{P_X}, MI'_{P_{XY}}, H'_{P_{X|Y}}] \qquad (11)$$
$$F_Y(P_{XY}) = [\Delta H'_{P_Y}, MI'_{P_{XY}}, H'_{P_{Y|X}}]$$

- But they can be represented side by side in the same diagram!

Splitting the Entropy Triangle

- This means we can also split the CBET in two.
- These describe the *marginal fractions* of entropy when the normalization is done with H_{U_X} and H_{U_Y} respectively

$$F_X(P_{XY}) = [\Delta H'_{P_X}, MI'_{P_{XY}}, H'_{P_{X|Y}}] \quad (11)$$
$$F_Y(P_{XY}) = [\Delta H'_{P_Y}, MI'_{P_{XY}}, H'_{P_{Y|X}}]$$

- But they can be represented side by side in the same diagram!

Examples of Confusion Matrices (again!)

Recall the confusion matrices we visited above:

$$a = \begin{bmatrix} 15 & 0 & 5 \\ 0 & 15 & 5 \\ 0 & 0 & 20 \end{bmatrix} \quad b = \begin{bmatrix} 16 & 2 & 2 \\ 2 & 16 & 2 \\ 1 & 1 & 18 \end{bmatrix} \quad c = \begin{bmatrix} 1 & 0 & 4 \\ 0 & 1 & 4 \\ 1 & 1 & 48 \end{bmatrix}$$

$$d = \begin{bmatrix} 15 & 0 & 0 \\ 0 & 18 & 0 \\ 0 & 0 & 27 \end{bmatrix} \quad e = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 57 \end{bmatrix} \quad f = \begin{bmatrix} 0 & 0 & 5 \\ 0 & 0 & 5 \\ 0 & 0 & 50 \end{bmatrix}$$

Split Diagram for Confusion Matrices

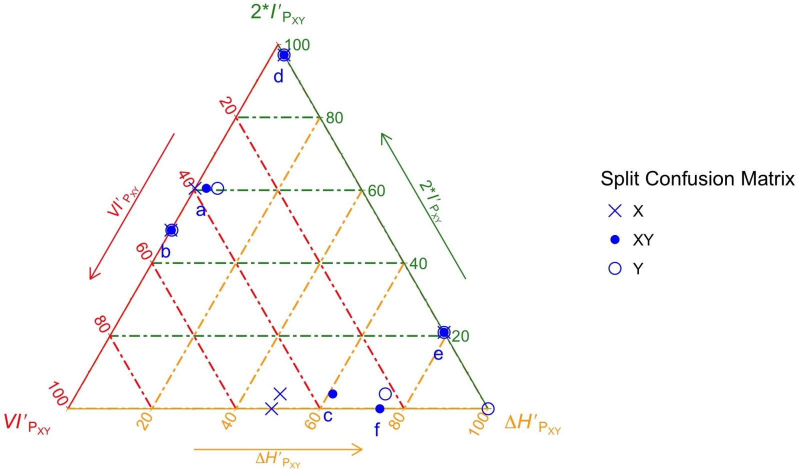


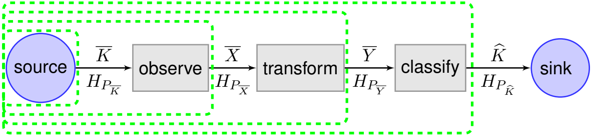
Figure: Representation of the previos confusion matrices in the split CBET

Outline

Outline

Entropy Balances for Sources

We can also try to investigate sources with entropy balance equations and entropy triangles



- A source should be as random as possible.
- How many variables does it have?
- How much privative information does each variable carry? And joint information?

Entropy Decomposition of a Source $X \sim P_{\overline{X}}$

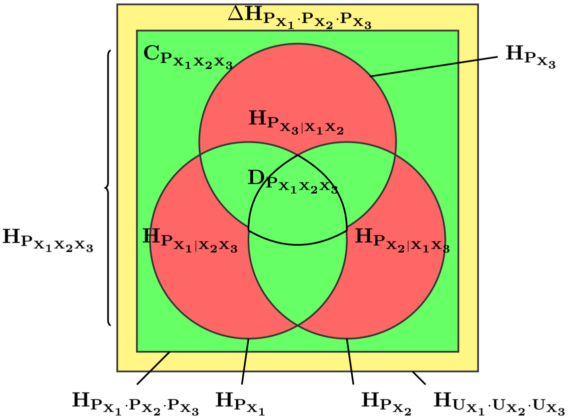


Figure: **(Color online) Extended entropy diagram of a trivariate distribution.** The bounding rectangle is the joint entropy of uniform (hence independent) distributions U_{X_i} of the same cardinality as distribution P_{X_i}. The green area is the sum of the multiinformation (total correlation) $C_{P_{\overline{X}}}$ and the dual total correlation $D_{P_{\overline{X}}}$.

The Balance equation for a Source

We can find an aggregate balance equation

$$H_{U_{\overline{X}}} = \Delta H_{\Pi_{\overline{X}}} + M_{P_{\overline{X}}} + VI_{P_{\overline{X}}} \qquad (12)$$
$$0 \leq \Delta H_{\Pi_{\overline{X}}}, M_{P_{\overline{X}}}, VI_{P_{\overline{X}}} \leq H_{U_{\overline{X}}}$$

The roles of these are analogues of those in the previous equation:

- The *divergence from uniformity*

  $$\Delta H_{\Pi_{\overline{X}}} = H_{U_{\overline{X}}} - H_{\Pi_{\overline{X}}}$$

  where $\Pi_{\overline{X}} = \Pi_j P_{X_j}$.
- The *remanent information* is the addition of informations in only one variable:

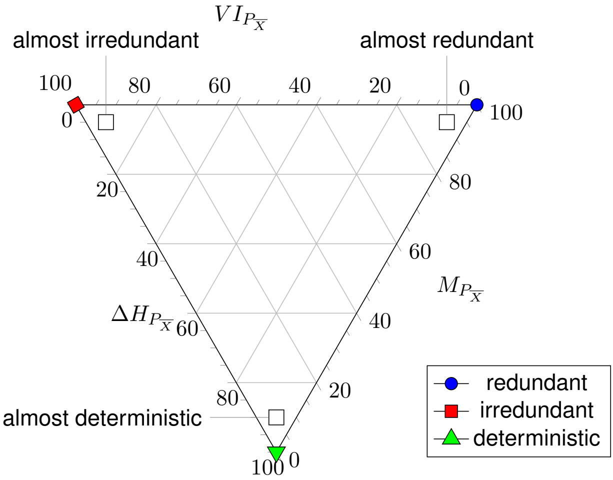
  $$VI_{P_{\overline{X}}} = \sum_{i=1}^{n} H_{P_{X_i|X_i^c}}$$

- The *bound information*, that in any subset of variables.

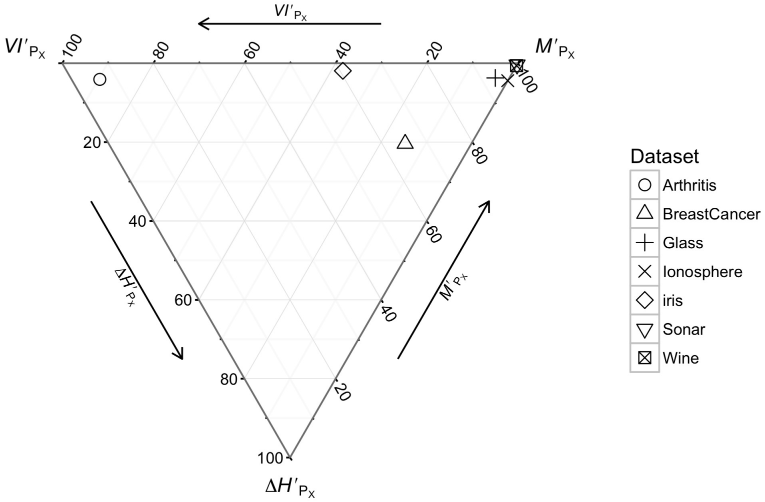
  $$M_{P_{\overline{X}}} = C_{P_{\overline{X}}} + D_{P_{\overline{X}}} = H_{\Pi_{\overline{X}}} - VI_{P_{\overline{X}}}$$

  .

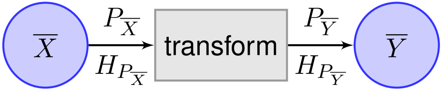
The Source Triangle



Some Examples on UCI Tasks



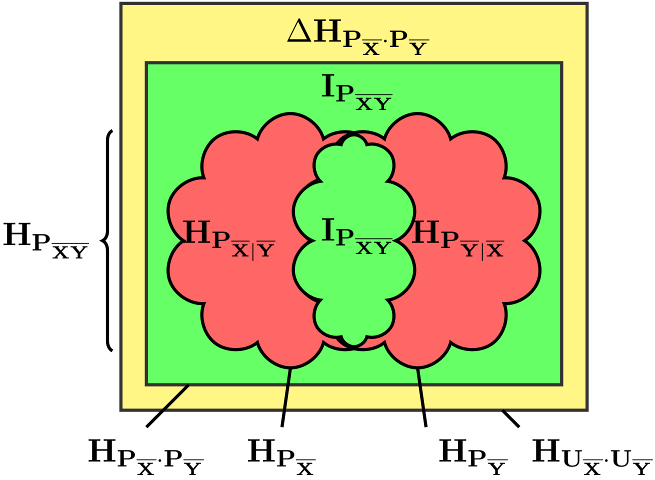
We can also look at transformations



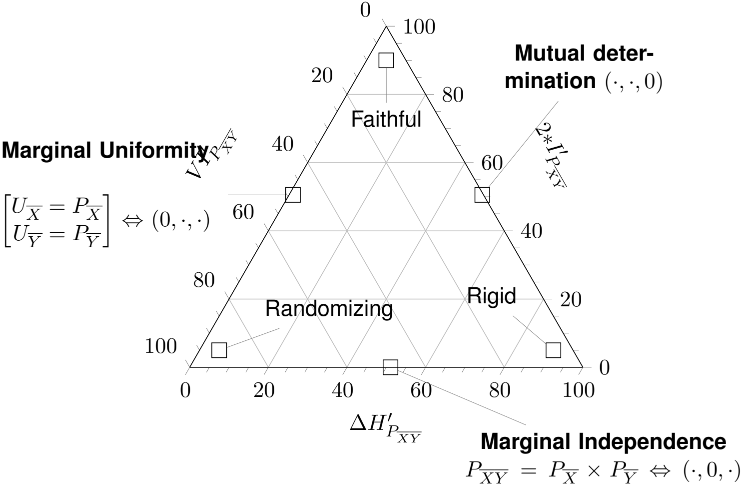
Transformations have features of both previous cases:

- They have an input and output, like the binary case.
- They are multivariable, like the source case.

The Entropy Diagram for a Transformation

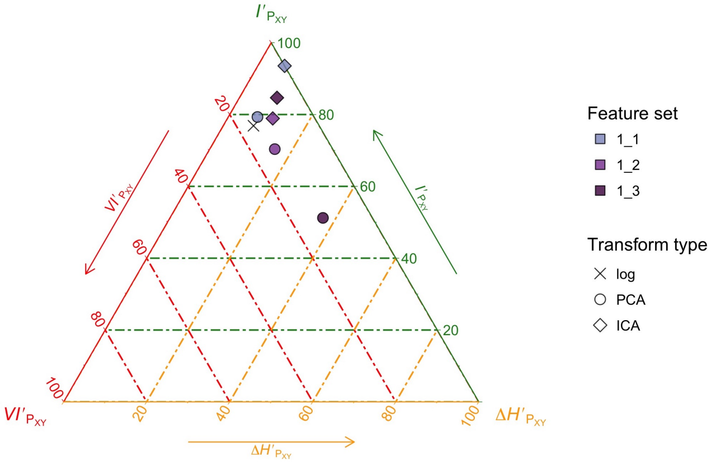


The Channel Multivariate Entropy Triangle



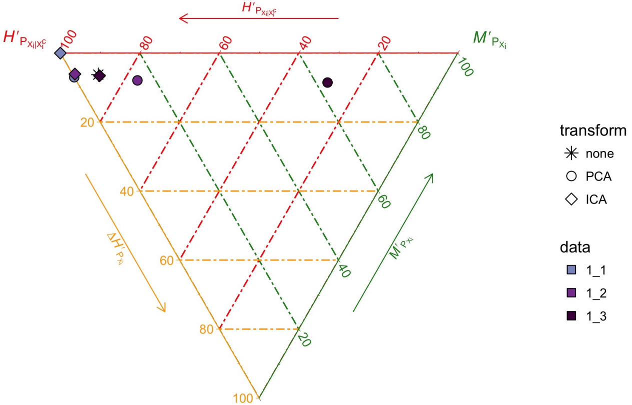
An Application: the Study of Feature Transformation

Comparing PCA & ICA on `Arthritis`



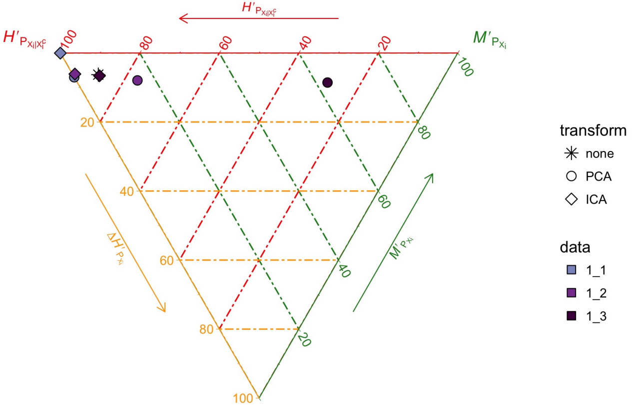
Using the SMET to Analize the Quality of the Features

- The figure analyzes the informational content of the feature sources $\overline{Y}_i$
- ICA_1 and ICA_2 and FCA_1 have more info per feature than the original database.
- ICA_1 has balanced the database! (Better for classifiers).



Using the SMET to Analize the Quality of the Features

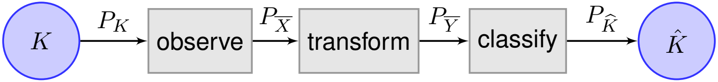
- The figure analyzes the informational content of the feature sources $\overline{Y}_i$
- ICA_1 and ICA_2 and FCA_1 have more info per feature than the original database.
- ICA_1 has balanced the database! (Better for classifiers).



Outline

Take Home Lessons

- Supervised Classification as an Information Channel



- It is possible to *assess* this system by entropic means. We have
  - mathematical tools, the balance equations and,
  - data exploratory tools, the entropy triangles and diagrams.
- You can study the system,
  - end-to-end (CBET), or
  - its sources (SMET)
  - its subsystems (CMET).
- Each tool provides different insight into the technologies implementing each subsystem.

G Keren and S Baggen. “Recognition models of alphanumeric characters.”. In: *PERCEPT PSYCHOPHYS* 29.3 (1981), pp. 234–246.

E.Z Rothkopf. “A measure of stimulus similarity and errors in some paired-associate learning tasks”. In: *J EXP PSYCHOL* 53.2 (1957), pp. 94–101.

V Sindhwani et al. “Feature selection in MLPs and SVMs based on maximum output information”. In: *IEEE Transactions on Neural Networks* 15.4 (2004), pp. 937–948.

Francisco J Valverde-Albacete and C Peláez-Moreno. “The Evaluation of Data Sources using Multivariate Entropy Tools”. In: *Expert Systems with Applications* 78 (2017), pp. 145–157. DOI: 10.1016/j.eswa.2017.02.010.

Francisco J. Valverde-Albacete and Carmen Peláez-Moreno. "100% classification accuracy considered harmful: the normalized information transfer factor explains the accuracy paradox". In: *PLOS ONE* (2014), pp. 1–10. DOI: 10.1371/journal.pone.0084217.

Francisco J. Valverde-Albacete and Carmen Peláez-Moreno. "Assessing Information Transmission in Data Transformations with the Channel Multivariate Entropy Triangle". In: *Entropy* 20.7 (2018). ISSN: 1099-4300. DOI: 10.3390/e20070498.

Francisco J. Valverde-Albacete and Carmen Peláez-Moreno. "Two information-theoretic tools to assess the performance of multi-class classifiers". In: *Pattern Recognition Letters* 31.12 (2010), pp. 1665–1671. DOI: 10.1016/j.patrec.2010.05.017.

H N Wright. "Characterization of olfactory dysfunction". In: *ARCH OTOLARYNGOL* 113.2 (Feb. 1987), pp. 163–168.