Groundwork for a New Approach to Knowledge Discovery

Certainty upon Empirical Distributions

Joan Garriga

Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
LSI-UPC

May 2011

Our Framework

- Knowledge: certainty about the outcome of a random event/variable.

Given,

- a relational (/transactional) domain;
- we have $\mathcal{X} = \{X^1, X^2, ..., X^m\}$, and a sample $\mathcal{D}$ of fixed size N;
- we does not assume any underlying distribution in the origin of the sample;
- we just observe some empirical distributions: for any $X^q \in \mathcal{X}$ we observe $\mathcal{P}(X^q)$ and $\mathcal{P}(X^q \mid \Pi(X^q))$.

Any such distribution expresses a degree of certainty about its outcome.
We want to measure this degree of certainty.

For the general case, we denote,

- $\Omega = \{e_1, e_2, \ldots, e_s\}$, a set of disjoint dependent events;
- $\mathcal{P} = (p_1, p_2, \ldots, p_s)$, a finite discrete probability distribution, (a vector of observed frequencies over Ω)

Shannon's Entropy

Shannon's Entropy is the most widely known measure of certainty, (uncertainty).

$$H(\mathcal{P}) = H(p_1, p_2, \ldots, p_n) = -\sum_{i=1}^{n} p_i logp_i$$

Attractive properties:

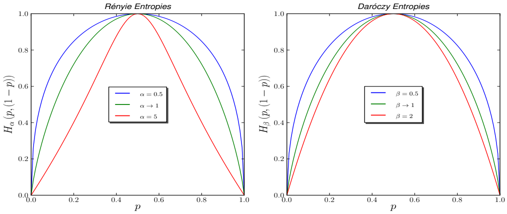
1. symmetry: $(p, 1-p)$ and $(1-p, p)$ have equal entropy;
2. normalization: a fair coin has entropy one, (more generally, max.entropy is $logn$);
3. monotonicity: the entropy of a coin, with bias p, goes to zero as p goes to zero;
4. subadditivity: $H(X, Y) \leq H(X) + H(Y)$, with equality for $X \perp Y$;
5. expansibility: $H(p_1, p_2, \ldots, p_n, 0, \ldots, 0) = H(p_1, p_2, \ldots, p_n)$;
6. composition:
   $H(t\, p_1, (1-t)\; p_1, p_2, \ldots, p_n) = H(p_1, p_2, \ldots, p_n) + p_1\, H(t, 1-t)$.

It is uniquely characterized by these properties.

A nice correspondance with a plausible axiomatic definition of knowledge.

Entropy generalizations

1. Rényi: $H_\alpha(\mathcal{P}) = H_\alpha(p_1, p_2, \ldots, p_n) = \frac{1}{1-\alpha} log \left(\sum_{i=1}^{n} p_i^\alpha\right)$
2. Daróczy: $H^\beta(\mathcal{P}) = \sum_{i=1}^{n} p_i \, u^\beta(p_i)$, where, $u^\beta(p_i) = \frac{2^{\beta-1}}{2^{\beta-1}-1}\left(1 - p_i^{\beta-1}\right)$



Entropy is a weighted mean value, where $-log\ p_i$ is the elementary entropy of e_i:

- higher values of α, β give more weight to the most probable events;
- lower values of α, β give more uniform weights to all possible values;
- $\alpha, \beta \to 1$ tend to Shannon's entropy, where the weight is just p_i

Entropy drawbacks

Generally known drawbacks:

- bias towards attributes with greater cardinalities
- undesired results with highly imbalanced frequencies

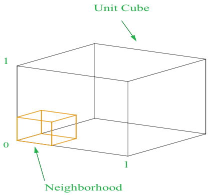
Particular questions, (the curse of dimensionality):

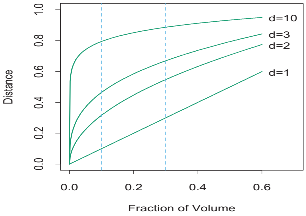
- the cardinality scaling of knowledge;
- the uncertainty of unseen events.

Is there any alternative, also plausible, axiomatic definition of knowledge?

The curse of dimensionality

- Let's figure our input space uniformly distributed in a p-dimensional unit hypercube.
- In order to capture a fraction r of the input space, we must consider an hypercubical fraction r of the unit volume.
- The expected edge length is $e_p = r^{1/p}$: a fraction of 10% yields an edge length $e_3 = 0.46$, but $e_{10} = 0.79$, and $e_{100} = 0.98$!!.
- The sample density is proportional to $N^{1/p}$: if N_1 is a dense sample for a single input problem, then $N_{10} = N_1^{10}$ is the sample size required for the same sampling density with 10 inputs.
- General case: given a fixed sample size (whatever large it is), the fraction captured is dramaticaly reduced, as the dimension of the input space increases.





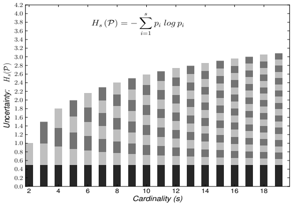
What are we really assuming when we undertake the *identical distribution* assumption?

The cardinality scaling of knowledge

Knowledge is akin to a notion of richness, related to the cardinality of Ω.

(The statistical significance of the observed frequencies is related to the dimension of the distribution).

- Let's figure a horse race beting example, with a fixed sample size N.
- Two runners (Tomcat and Apache): 50% of victories each.
- s runners (Tomcat and $(s-1)$ uniform competitors): we still observe a 50% of victories for Tomcat.
- We are always bound to loose 0.5 of our bets in the long run, but our epistemic state is quite different.



Does our initial state of knowledge (with two runners), evolve to an unbound uncertainty, as the number of runners increases?

Uncertainty of unseen events

- The dimension of the input space grows geometrically with the cardinalities of the input features.
- As soon as the complexity of the model involves just a few number of features, the sample will be very sparse.
- A great number of configurations will not be present in the sample: unseen events (observed frequency, $p_i = 0$.)

For any unseen event, entropy yields a puzzling result (!?):

$$H(0) = 0\, log \infty = -0\, log 0 = 0$$

At this point, one has to rely on *ad-hoc* smoothing procedures

Our proposal

Two main axiomatic intuitions:

- the minimum knowledge is given in the case of uniformity;
- knowledge is akin to a notion of richness, related to the cardinality of Ω.

Given:

- a set of disjoint dependent events $\Omega = \{e_1, e_2, \ldots, e_s\}$ with an observed distribution $\mathcal{P} = (p_1, p_2, \ldots, p_s)$;
- $\mid \Omega \mid = s$;
- $C_s = \frac{(s-1)}{s}$, and $U_s = \frac{1}{s}$, the *certainty* and *uncertainty* factors;

We take the distance to the most uninformative distribution,

$$\Delta(\mathcal{P}) = \sum_{j=1}^{n} (p_j - U_s)^2 \quad \equiv \quad L_2(\mathcal{P}, \mathcal{U})$$

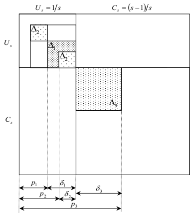
plus, our axiomatic requests, (knowledge can not be zero), i.e.,

$$K_s(\mathcal{P}) = U_s + \Delta(\mathcal{P}) = \sum_{j=1}^{n} p_j^2$$

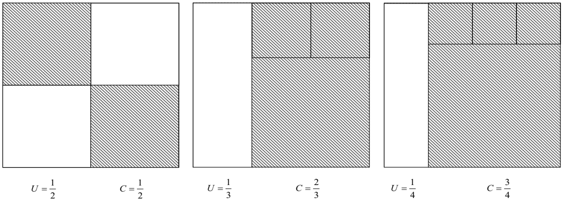
We get a direct measure of *Certainty*, or knowledge, conveyed by distribution $\mathcal{P}$

Geometric interpretation

Graphical depiction of knowledge conveyed by $\mathcal{P} = (p_1, p_2, p_3)$.

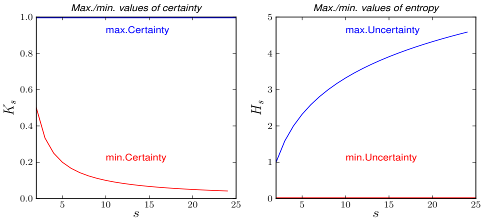


Areas of certainty and uncertainty for $s = (2, 3, 4)$.



Certainty properties I

- symmetry: $(p, 1-p)$ and $(1-p, p)$ convey equal certainty;
- normalization: the maximum certainty is one for any distribution;
- monotonicity, (with respect to deviation): the certainty of a coin, with bias p, goes to one as p goes to one; (but different from small information for small probabilities !!!)



Certainty properties II

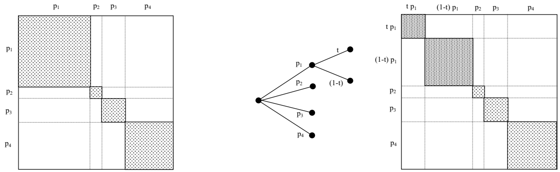
- expansibility:

$$K_{s+1}(p_1, p_2, \ldots, p_s, 0) = K_s(p_1, p_2, \ldots, p_s) = \sum_{j=1}^{s} p_j^2$$

(note the different offset, $U_{s+1} + \sum_{j=1}^{s+1}(p_j - U_{s+1})^2 = U_s + \sum_{j=1}^{s}(p_j - U_s)^2$).

- composition: given distributions $\mathcal{P} = (p_1, p_2, \ldots, p_s)$ and $\mathcal{T} = (t, 1-t)$, and their composition $\mathcal{Q} = (t\, p_1, (1-t)\, p_1, p_2, \ldots, p_s)$,

$$K_{s+1}(\mathcal{Q}) = K_s(\mathcal{P}) - p_1^2\,(1 - K_2(\mathcal{T}))$$



a) b)

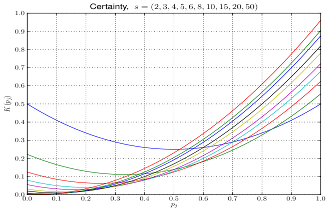
(remember, $H(\mathcal{Q}) = H(\mathcal{P}) + p_1\, H(\mathcal{T})$)

Elementary contributions to certainty of disjoint dependent events

$$\forall e_j \in \Omega, \quad p_j^2 = (U_s + (p_j - U_s))^2 = U_s^2 + (p_j - U_s)^2 + 2\,U_s\,(p_j - U_s)\ ,\ \text{that is,}$$

$$K_s(p_j) = U_s^2 + (p_j - U_s)^2 \quad \Longrightarrow \quad K_s(\mathcal{P}) = \sum_{j=1}^{s} K_s(p_j)$$

- explicitely dependent on s: $\lim_{s\to\infty} K_s(p_j) = p_j^2$;
- coherent minimum at equiprobability: $K_s(U_s) = U_s^2$;
- not null values for unseen events, (continuous at zero): $K_s(0) = U_s^2 + U_s^2$;
- not one values for completely biased events: $K_s(1) = U_s^2 + C_s^2$
- minimum certainty: $K_s(\mathcal{U}) = U_s$;
- maximum, (not absolute), certainty: $K_s(1, 0, \ldots, 0) = K_s(1) + (s-1)\,K_s(0) = 1$



Conditional Certainty

Given, $(\mathcal{Q} \mid e_i) = (q_{1i}, q_{2i}, \ldots, q_{si})$, e_i drawn from $\mathcal{P} = (p_1, p_2, \ldots, p_r)$;

$$K_s(\mathcal{Q} \mid e_i) = U_s + \sum_{j=1}^{s}(q_{ji} - U_s)^2 = \sum_{j=1}^{s} q_{ji}^2 \quad \Longrightarrow \quad \mathcal{Q} \perp e_i, \quad K_s(\mathcal{Q} \mid e_i) = K_s(\mathcal{Q})$$

Clear expression of dependence in terms of conditional certainty:

- minimum certainty given at the point of uniformity: $K_s(\mathcal{Q} \mid e_i) = U_s$;
- at the point of independence we have: $K_s(\mathcal{Q} \mid e_i) = K_s(\mathcal{Q}) \geq U_s$;
- from that point on, $K_s(\mathcal{Q} \mid e_i) \geq K_s(\mathcal{Q})$, and we may begin to consider a relation of dependence;
- in case of absolute dependence, $K_s(\mathcal{Q} \mid e_i) = 1$.

With respect to the whole distribution $\mathcal{P}$, it makes sense to consider:

$$K(\mathcal{Q} \mid \mathcal{P}) = \sum_{i=1}^{r} p_i \, K_s(\mathcal{Q} \mid e_i) = \sum_{i=1}^{r} p_i \sum_{j=1}^{s} q_{ji}^2$$

⥀ (different from composition: different cardinality of the final outcome !!)

Joint Certainty

Given, $(\mathcal{P}, \mathcal{Q}) = (p_{11}, \dots, p_{1s}, \dots, p_{r1}, \dots, p_{rs})$, and its Ufactor, $U_{rs} = U_r\, U_s$,

$$K(\mathcal{P}, \mathcal{Q}) = U_{rs} + \sum_{i,j}^{r,s} (p_{ij} - U_{rs})^2 = \sum_{i,j}^{r,s} p_{ij}^2$$

- minimum certainty given at the point of uniformity: $K_{rs}(\mathcal{P}, \mathcal{Q}) = U_{rs}$;
- in case of independence,(multiplicativity):

$$K_{rs}(\mathcal{P}, \mathcal{Q}) = K_r(\mathcal{P})\, K_s(\mathcal{Q});$$

- from that point on: $K_{rs}(\mathcal{P}, \mathcal{Q}) = \sum_i^r p_i^2\, K_s(\mathcal{Q} \mid e_i) = \sum_i^r p_i K_s^i(\mathcal{Q} \mid P)$
- in case of dependence, (supermultiplicativity):

$$\forall e_i,\ K_s(\mathcal{Q} \mid e_i) \geq K_s(\mathcal{Q}) \text{ , therefore,}$$

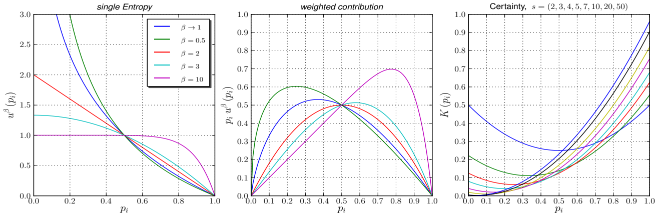
$$K_{rs}(\mathcal{P}, \mathcal{Q}) = \sum_{i=1}^{r} p_i^2\, K_s(\mathcal{Q} \mid e_i) \geq \sum_{i=1}^{r} p_i^2\, K_s(\mathcal{Q}) = K_r(\mathcal{P})\, K_s(\mathcal{Q}),$$

- in case of absolute dependence, $\forall e_i$, $K_s(\mathcal{Q} \mid e_i) = 1$, and $K_{rs}(\mathcal{P}, \mathcal{Q}) = K_r(\mathcal{P})$.

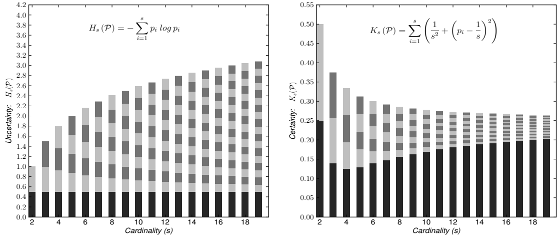
The algebra of Certainty

We build up an alternative hypothesis, that:

- offers a comprehensible insight of knowledge, (plausible axiomatic definition);
- has a consistent algebraic structure, (certainty is not a weighted mean);
- satisfies a set of consistent properties;
- is cardinality dependent, (stronger implications of expansibility);
- yields not null values for unseen events, (native smoothing);



Our former example

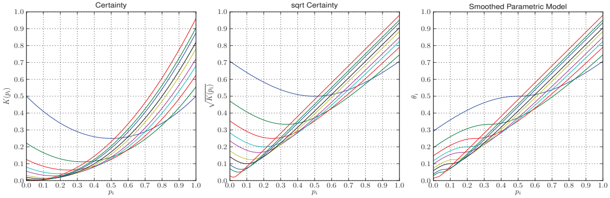


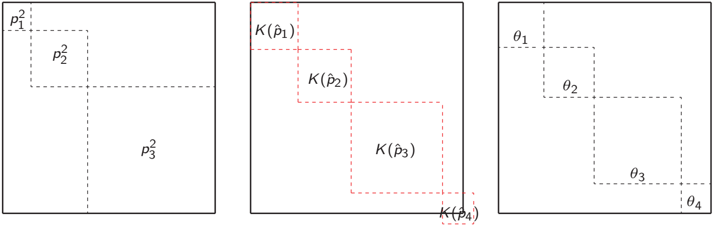
Given a fixed sample size N, and $\mathcal{P} = (0.5, \frac{0.5}{s-1}, \ldots, \frac{0.5}{s-1})$ of increasing dimension:

- as the cardinality increases, certainty decreases, and each competitor's contribution is less, (up to here, correctly expressed by both);
- the difference: our certainty is increasingly due to Tomcat's chances;
- at the limit (ideal situation), we just have the amount contributed by Tomcat, (the competitors contributions are null because their chances vanish);
- Tomcat's victory seems amaizingly guaranteed, but our certainty can not be one, because Tomcat's chances are less than one.

We judge this a more comprehensible description of our epistemic state, than a state of unbound uncertainty.

Native smoothing





$\mathcal{P}=(0.14,0.26,0.60,0)$

$K(\mathcal{P})=\sum_{i=1}^{4} p_i^2 = 0.4489$

$\hat{\mathcal{P}}=(0.22,0.25,0.43,0.14) \geq 1$

$K(\hat{\mathcal{P}})=\sum_{i=1}^{4} K(\hat{p}_i) = 0.4489$

$\Theta=(0.21,0.24,0.41,0.14)$

$K(\Theta) \neq K(\mathcal{P})$

Empirical Evaluation on Decision Trees

Base measure implemented in ID3 and C4.5 induction tree algorithms:

- Entropic Gain (Quinlan, 1986), $H(class) - H(class \mid attr.)$;

Other common measures based on Shannon's Entropy:

- the μ coefficient of Theil, (Theil, 1970), $\frac{H(class) - H(class|attr.)}{H(class)}$;
- the gain-ratio (Quinlan, 1993), $\frac{H(class) - H(class|attr.)}{H(attr.)}$;
- the Kvalseth coefficient (Kvalseth, 1987), $\frac{2(H(class) - H(class|attr.))}{H(attr.) + H(class)}$

Our implementation of certainty:

- at each node check for attributes yielding, $K_s(class \mid attr.) \geq K_s(class)$;
- among them choose the one with greater *utility*, i.e.,

$$Utl(attr.) = (1 - (K_r(attr.) - U_r)) \frac{1}{2}(K_s(class \mid attr.) + K_r(attr. \mid class))$$

Experimental Results

DataBase	setSize	attr.	Clssf.	tree	treeSize	nodes	leaves	nullLvs.	%uncovered	%correct
BreastCancer	683	10	10fld	ID3	211	21	190	95	50.00	91.65
			10fld	C4.5	61	6	55	14	25.45	93.41
			10fld	Crt.	51	5	46	4	8.70	95.46
SegmentChallenge	1500	20	10fld	ID3	390	44	346	193	55.78	93.92
			10fld	C4.5	213	23	190	102	53.68	94.93
			10fld	Crt.	174	28	146	49	33.56	91.73
OpticalDígits	5620	65	10fld	ID3	11493	676	10817	7582	70.09	44.11
			10fld	C4.5	4023	241	3782	2334	61.71	63.02
			10fld	Crt.	1769	104	1665	333	20.00	54.02
	1797		testSet	C4.5	3010	177	2833	1737	61.31	56.82
	1797		testSet	Crt.	1225	72	1153	198	17.17	54.26
penDigits	10992	17	10fld	ID3	5798	527	5271	2955	56.06	86.69
			10fld	C4.5	2366	215	2151	1068	49.65	89.16
			10fld	Crt.	1805	164	1641	342	20.84	86.85
	3498		testSet	C4.5	1915	174	1741	910	52.27	84.08
	3498		testSet	Crt.	1288	117	1171	227	19.39	81.76
letterRecognition	20000	17	10fld	ID3	30561	1910	28651	21832	76.20	73.53
			10fld	C4.5	13409	838	12571	9033	71.86	77.73
			10fld	Crt.	4929	308	4621	2294	49.64	72.66
Soybean	562	36	10fld	ID3	50	51	116	31	26.72	83.77
			10fld	C4.5	69	22	47	10	21.28	91.81
			10fld	Crt.	149	59	90	3	3.33	89.15
CarEvaluation	1728	7	10fld	ID3	408	112	296	0	0.00	89.35
			10fld	C4.5	182	51	131	0	0.00	92.36
			10fld	Crt.	213	58	155	0	0.00	94.21
			trainSet	C4.5	182	51	131	0	0.00	96.30
			trainSet	Crt.	213	58	155	0	0.00	96.30
Nursery	12960	9	10fld	ID3	1159	320	839	0	0.00	98.19
			10fld	C4.5	511	152	359	0	0.00	97.05
			10fld	Crt.	1031	274	757	0	0.00	96.37
			trainSet	C4.5	511	152	359	0	0.00	98.13
			trainSet	Crt.	1031	274	757	0	0.00	98.59