Quality control of corpus annotation through reliability measures

Ron Artstein

Department of Computer Science
University of Essex
artstein@essex.ac.uk

ACL-2007 tutorial
24 June 2007

Thanks to EPSRC grant GR/S76434/01, ARRAU
(Anaphora Resolution and Underspecification)

Annotated corpora

Annotated corpora are needed for:

- Supervised learning – training and evaluation
- Unsupervised learning – evaluation
- Hand-crafted systems – evaluation
- Analysis of text

Quality control:

- Annotations need to be correct.

Correctness and reliability

Systems are evaluated with respect to a standard

- standard taken to be **correct**

During corpus creation, no standard exists

- As a minimum, annotation should be **reliable**
- Qualitative evaluation also necessary

Reliability and agreement

Reliability = **consistency**

- Needs to be measured on the same text
- Different annotators

If independent annotators mark a text the same way,

- they have internalized the same scheme (instructions)
- will apply it consistently to new data
- annotations might be correct

Reliability studies

Reliability data

- Sample of the corpus
- Multiple annotators

Annotators must work **independently**

- Otherwise we can't compare them

Results **do not generalize** from one domain to another

- Annotators internalized a scheme for newswire corpus
- They may apply it differently to email corpus

Measuring agreement

Agreement measures are not hypothesis tests

- Evaluating magnitude, not existence/lack of effect
- Not comparing two hypotheses
- No clear probabilistic interpretation

Observed agreement

Observed agreement: proportion of items on which 2 coders agree.

Detailed Listing

Item	Coder 1	Coder 2
a	Boxcar	Tanker
b	Tanker	Boxcar
c	Boxcar	Boxcar
d	Boxcar	Tanker
e	Tanker	Tanker
f	Tanker	Tanker
	⋮	⋮

Contingency Table

	Boxcar	Tanker	Total
Boxcar	41	3	44
Tanker	9	47	56
Total	50	50	100

Agreement: $\frac{41+47}{100} = 0.88$

Chance agreement

Some agreement is expected by chance alone.

- Two coders randomly assigning "Boxcar" and "Tanker" labels will agree half of the time.
- The amount expected by chance varies depending on the annotation scheme and on the annotated data.

Meaningful agreement is the agreement **above chance**.

- Similar to the concept of "baseline" for system evaluation.

Correction for chance

How much of the observed agreement is above chance?

	A	B	Total
A	44	6	50
B	6	44	50
Total	50	50	100

Total

44	6
6	**44**

88

=

Chance

6	6
6	**6**

12

+

Above

38	0
0	**38**

76

Agreement: $88/100$
Due to chance: $12/100$
Above chance: $76/100$

Correction for chance

How much of the observed agreement is above chance?

	A	B	C	D	Total
A	22	1	1	1	25
B	1	22	1	1	25
C	1	1	22	1	25
D	1	1	1	22	25
Total	25	25	25	25	100

Correction for chance

$$\underset{\textbf{88}}{\overset{\textit{Total}}{\begin{array}{|cccc|}\hline \mathbf{22} & 1 & 1 & 1 \\ 1 & \mathbf{22} & 1 & 1 \\ 1 & 1 & \mathbf{22} & 1 \\ 1 & 1 & 1 & \mathbf{22} \\ \hline \end{array}}} = \underset{\textbf{4}}{\overset{\textit{Chance}}{\begin{array}{|cccc|}\hline \mathbf{1} & 1 & 1 & 1 \\ 1 & \mathbf{1} & 1 & 1 \\ 1 & 1 & \mathbf{1} & 1 \\ 1 & 1 & 1 & \mathbf{1} \\ \hline \end{array}}} + \underset{\textbf{84}}{\overset{\textit{Above}}{\begin{array}{|cccc|}\hline \mathbf{21} & 0 & 0 & 0 \\ 0 & \mathbf{21} & 0 & 0 \\ 0 & 0 & \mathbf{21} & 0 \\ 0 & 0 & 0 & \mathbf{21} \\ \hline \end{array}}}$$

Agreement:	*88/100*
Due to chance:	*4/100*
Above chance:	*84/100*

Correction for chance

	A	B	Total
A	44	6	50
B	6	44	50
Total	50	50	100

	A	B	C	D	Total
A	22	1	1	1	25
B	1	22	1	1	25
C	1	1	22	1	25
D	1	1	1	22	25
Total	25	25	25	25	100

Agreement: 88/100
Due to chance: 12/100
Above chance: 76/100

Agreement: 88/100
Due to chance: 4/100
Above chance: 84/100

Expected agreement

Observed agreement (A_o)*: proportion of actual agreement*
Expected agreement (A_e)*: expected value of* A_o

Amount of agreement above chance: $A_o - A_e$
Maximum possible agreement above chance: $1 - A_e$

Proportion of agreement above chance attained: $\dfrac{\mathbf{A_o - A_e}}{\mathbf{1 - A_e}}$

Expected agreement

Big question: how to calculate the amount of agreement expected by chance (A_e)?

S: same chance for all coders and categories

Number of category labels: $\mathbf{q}$

Probability of one coder picking a particular category q_a: $\frac{1}{\mathbf{q}}$

Probability of both coders picking a particular category q_a: $\left(\frac{1}{\mathbf{q}}\right)^2$

Probability of both coders picking the same category:

$$A_e^S = \mathbf{q} \cdot \left(\frac{1}{\mathbf{q}}\right)^2 = \frac{1}{\mathbf{q}}$$

Are all categories equally likely?

	A	B	Total
A	44	6	50
B	6	44	50
Total	50	50	100

	A	B	C	D	Total
A	44	6	0	0	50
B	6	44	0	0	50
C	0	0	0	0	0
D	0	0	0	0	0
Total	50	50	0	0	100

$$A_o = 0.88$$
$$A_e = \frac{1}{2} = 0.5$$
$$S = \frac{0.88-0.5}{1-0.5} = 0.76$$

$$A_o = 0.88$$
$$A_e = \frac{1}{4} = 0.25$$
$$S = \frac{0.88-0.25}{1-0.25} = 0.84$$

π: different chance for different categories

Total number of judgments: **N**

Probability of one coder picking a particular category q_a: $\frac{\mathbf{n_{qa}}}{\mathbf{N}}$

Probability of both coders picking a particular category q_a: $\left(\frac{\mathbf{n_{qa}}}{\mathbf{N}}\right)^2$

Probability of both coders picking the same category:

$$A_e^\pi = \sum_q \left(\frac{\mathbf{n}_q}{N}\right)^2 = \frac{1}{N^2} \sum_q \mathbf{n}_q^2$$

Comparison of S and π

	A	B	C	Total
A	44	6	0	50
B	6	44	0	50
C	0	0	0	0
Total	50	50	0	100

$$A_o = 0.88$$
$$S = \frac{0.88-1/3}{1-1/3} = 0.82$$
$$\pi = \frac{0.88-0.5}{1-0.5} = 0.76$$

	A	B	C	Total
A	77	1	2	80
B	1	6	3	10
C	2	3	5	10
Total	80	10	10	100

$$A_o = 0.88$$
$$S = \frac{0.88-1/3}{1-1/3} = 0.82$$
$$\pi = \frac{0.88-0.66}{1-0.66} \approx 0.65$$

We can prove that for any sample: $A_e^\pi \geq A_e^S \qquad \pi \leq S$

Prevalence

Is the following annotation reliable?

Two annotators disambiguate 1000 instances of the word **love**:

- emotion
- zero (as in tennis)

Each annotator found:

- 995 instances of 'emotion'
- 5 instances of 'zero'

The annotators marked **different** instances of 'zero'. **Agr: 99%!**

	emotion	*zero*	*Total*
emotion	*990*	*5*	*995*
zero	*5*	*0*	*5*
Total	*995*	*5*	*1000*

$$A_o = 0.99$$

$$S = \frac{0.99 - .5}{1 - .5} = 0.98$$

$$\pi = \frac{0.99 - 0.99005}{1 - 0.99005} \approx -0.005$$

Prevalence

When one category is dominant:

- High agreement **does not indicate** high reliability
- π measures agreement on the rare category

Therefore, π is a good indicator of reliability.

Individual annotator bias

Different annotators have different interpretations of the instructions (bias/prejudice).

Does this affect expected agreement?

κ: different chance for different coders

Total number of items: $\mathbf{i}$

Probability of coder c_x *picking a particular category* q_a*:* $\frac{\mathbf{n}_{c_x q_a}}{\mathbf{i}}$

Probability of both coders picking category q_a*:* $\frac{\mathbf{n}_{c_1 q_a}}{\mathbf{i}} \cdot \frac{\mathbf{n}_{c_2 q_a}}{\mathbf{i}}$

Probability of both coders picking the same category:

$$A_e^{\kappa} = \sum_q \frac{\mathbf{n}_{c_1 q}}{\mathbf{i}} \cdot \frac{\mathbf{n}_{c_2 q}}{\mathbf{i}} = \frac{1}{\mathbf{i}^2} \sum_q \mathbf{n}_{c_1 q} \mathbf{n}_{c_2 q}$$

Comparison of π and κ

	A	B	C	Total
A	38	0	12	50
B	0	12	0	12
C	0	0	38	38
Total	38	12	50	100

	A	B	C	Total
A	17	0	40	57
B	0	26	0	26
C	0	0	17	17
Total	17	26	57	100

$$A_o = 0.88$$
$$\pi = \frac{0.88-0.4016}{1-0.4016} \approx 0.7995$$
$$\kappa = \frac{0.88-0.3944}{1-0.3944} \approx 0.8018$$

$$A_o = 0.6$$
$$\pi = \frac{0.6-0.3414}{1-0.3414} \approx 0.3927$$
$$\kappa = \frac{0.6-0.2614}{1-0.2614} \approx 0.4584$$

We can prove that for any sample: $A_e^\pi \geq A_e^\kappa$ $\quad \pi \leq \kappa$

Individual annotator bias

Different interpretations of the instructions = lack of reliability.

- π preferable to κ

High agreement entails small differences between coders.

- Small numerical difference between π and κ

Differences among coders are diluted when more coders are used.

- Small numerical difference between π and κ

Multiple coders

Multiple coders: Agreement is the proportion of agreeing **pairs**

Item	Coder 1	Coder 2	Coder 3	Coder 4	Pairs
a	Boxcar	Tanker	Boxcar	Tanker	2/6
b	Tanker	Boxcar	Boxcar	Boxcar	3/6
c	Boxcar	Boxcar	Boxcar	Boxcar	6/6
d	Tanker	Engine 2	Boxcar	Tanker	1/6
e	Engine 2	Tanker	Boxcar	Engine 1	0/6
f	Tanker	Tanker	Tanker	Tanker	6/6
g	Engine 1	Engine 1	Engine 1	Engine 1	6/6
	⋮	⋮	⋮	⋮	

Multiple coders

Numerical interpretation

- When 3 of 4 coders agree, only 3 of 6 pairs agree

Graphical representation

- Contingency table requires multiple dimensions. . .

Expected agreement

- The probability of agreement for an **arbitrary pair** of coders

K: multiple coders

Confusing terminology: K is a generalization of π.

Total number of judgments: $\mathbf{N}$

Probability of arbitrary coder picking a particular category q_a: $\frac{\mathbf{n_{q_a}}}{\mathbf{N}}$

Probability of two coders picking a particular category q_a: $\left(\frac{\mathbf{n_{q_a}}}{\mathbf{N}}\right)^2$

Probability of two arbitrary coders picking the same category:

$$\mathrm{A}_{\mathrm{e}}^{K} = \sum_q \left(\frac{\mathbf{n}_q}{N}\right)^2 = \frac{1}{N^2} \sum_q \mathbf{n}_q^2$$

Multiple coders – example

Item	Cod-1	Cod-2	Cod-3	Cod-4	Pairs
(a)	Box	Box	Box	Box	6/6
(b)	Box	Box	Box	Box	6/6
(c)	E-2	E-2	E-2	E-2	6/6
(d)	Tank	Tank	Tank	Tank	6/6
(e)	E-1	E-1	E-1	E-1	6/6
(f)	E-1	Box	E-1	E-1	3/6
(g)	Tank	Tank	Tank	Tank	6/6
(h)	Box	Box	Box	Box	6/6
(i)	Box	Box	Box	Box	6/6
(j)	Box	Box	E-1	Box	3/6
(k)	E-2	E-2	E-2	E-2	6/6
(l)	Box	Tank	Box	Box	3/6
(m)	E-1	E-1	E-1	E-1	6/6
(n)	Tank	Tank	Tank	Tank	6/6
(o)	E-1	E-1	E-1	E-1	6/6
(p)	E-2	E-2	E-2	Tank	3/6
(q)	Box	Box	Box	Box	6/6
(r)	Box	Box	Box	Box	6/6
(s)	E-1	E-1	Tank	E-1	3/6
(t)	Box	Box	Box	Box	6/6
(u)	Box	Box	Box	Box	6/6
(v)	E-1	E-1	E-1	E-1	6/6
(w)	Tank	Tank	Tank	Tank	6/6
(x)	Box	Box	Box	Box	6/6
(y)	Box	Box	Box	Tank	3/6

25 items, 100 judgments:
Box **46**, Tank **20**, E-1 **23**, E-2 **11**.

Observed agreement:

$$A_o = 132/150 = 0.88$$

Expected agreement:

$$A_e = .46^2 + .2^2 + .23^2 + .11^2 = 0.3166$$

$$K = \frac{0.88 - 0.3166}{1 - 0.3166} \approx 0.8244$$

Are all disagreements the same?

Some disagreements are more important than others

- **Boxcar/engine** more serious than **engine 1/engine 2**
- Depends on application

Need to count and weigh the disagreements

- Not only agreeing pairs
- Principled method of assigning weights

Agreement and disagreement

Observed disagreement: $D_o = 1 - A_o$
Expected disagreement: $D_e = 1 - A_e$

Chance-corrected **agreement***:*

$$1 - \frac{D_o}{D_e} = 1 - \frac{1 - A_o}{1 - A_e} = \frac{1 - A_e - (1 - A_o)}{1 - A_e} = \frac{A_o - A_e}{1 - A_e}$$

Weights

Three labels: Boxcar, Engine 1, Engine 2.

Three weights:

- **Identical judgments**: disagreement $= 0$ (agreement $= 1$)
- **Engine 1 / engine 2**: disagreement $= 0.5$ (agreement $= 0.5$)
- **Boxcar / engine**: disagreement $= 1$ (agreement $= 0$)

Weight table:

	Box	*E-1*	*E-2*
Box	*0*	*1*	*1*
E-1	*1*	*0*	*0.5*
E-2	*1*	*0.5*	*0*

Weighted kappa κ_w

Observed disagreement:

	Box	E-1	E-2	
Box	29	1	0	30
E-1	1	39	10	50
E-2	0	10	10	20
	30	50	20	100

$$\bullet \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0.5 \\ 1 & 0.5 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 5 \\ 0 & 5 & 0 \end{bmatrix}$$

0	1	0	1
1	0	5	6
0	5	0	5
1	6	5	**12**

Expected disagreement:

	Box	E-1	E-2	
Box	9	15	6	30
E-1	15	25	10	50
E-2	6	10	4	20
	30	50	20	100

$$\bullet \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0.5 \\ 1 & 0.5 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 15 & 6 \\ 15 & 0 & 5 \\ 6 & 5 & 0 \end{bmatrix}$$

0	15	6	21
15	0	5	20
6	5	0	11
21	20	11	**52**

$$\kappa_w = 1 - \frac{0.12}{0.52} \approx 0.77 \qquad K = \frac{.78 - .38}{1 - .38} \approx 0.65$$

Krippendorff's α: a generalized weighted coefficient

Krippendorff's α:

- Generalization of K with various distance metrics
  - Allows multiple coders
- Similar to K when categories are nominal
- Allows numerical category labels
  - Related to ANOVA (analysis of variance)

Analysis of variance

Numerical judgments (e.g. magnitude estimation)

- Single-variable ANOVA, each item = separate level

$$F = \frac{\text{between-level variance}}{\text{error variance}}$$

F = 1: Levels non-distinct; random

F > 1: Levels distinct to some extent; effect exists

$$\frac{\text{error variance}}{\text{total variance}}$$

0: No error; perfect agreement

1: Random; no distinction

2: Maximal value

$$\alpha = 1 - \frac{\textit{error variance}}{\textit{total variance}}$$

Example of α

Item	C-1	C-2	C-3	C-4	C-5	Mean	Variance
(a)	7	7	7	7	7	7.0	**0.0**
(b)	5	4	5	6	5	5.0	**0.5**
(c)	5	5	5	6	4	5.0	**0.5**
(d)	7	8	6	7	7	7.0	**0.5**
(e)	4	2	3	3	2	2.8	**0.7**
(f)	6	7	6	6	6	6.2	**0.2**
(g)	6	6	6	5	6	5.8	**0.2**
(h)	7	6	9	6	9	7.4	**2.3**
(i)	5	5	5	4	5	4.8	**0.2**
(j)	4	5	2	4	6	4.2	**2.2**
(k)	3	5	2	4	4	3.6	**1.3**
(l)	5	5	6	6	5	5.4	**0.3**
(m)	3	4	2	3	3	3.0	**0.5**
(n)	2	3	4	3	4	3.2	**0.7**
(o)	7	7	6	7	7	6.8	**0.2**
(p)	7	8	7	8	7	7.4	**0.3**
(q)	3	3	3	1	3	2.6	**0.8**
(r)	4	2	4	2	4	3.2	**1.2**
(s)	3	2	3	3	3	2.8	**0.2**
(t)	4	4	2	4	4	3.6	**0.8**
(u)	5	6	4	5	6	5.2	**0.7**
(v)	4	3	4	3	1	3.0	**1.5**
(w)	6	6	7	5	7	6.2	**0.7**
(x)	4	5	2	4	3	3.6	**1.3**
(y)	4	5	5	6	5	5.0	**0.5**

Mean variance per item: **0.732**

Overall: 25 items, 125 judgments.

'1' **2**	'2' **11**	'3' **19**	'4' **24**	'5' **23**
'6' **22**	'7' **19**	'8' **3**	'9' **2**	

Mean: 4.792, Variance: **3.085**

$$\alpha = 1 - \frac{0.732}{3.085} = 0.763$$

$$F(24, 100) = \frac{12.891}{0.732} = 17.611, p < 1^{-15}$$

α with different distance metrics

General formula for α

$$\alpha = 1 - \frac{\textit{error variance}}{\textit{total variance}} = 1 - \frac{\textit{mean item distance}}{\textit{mean overall distance}} = 1 - \frac{D_o}{D_e}$$

Observed and expected disagreements computed with various **distance metrics**

Distance metrics for α

Interval α (numeric values)

$$\mathbf{d}_{ab} = (a - b)^2$$

Nominal α (all disagreements equal)

$$\mathbf{d}_{ab} = \begin{cases} 0 \ \textit{if } a = b \\ 1 \ \textit{if } a \neq b \end{cases}$$

Nominal $\alpha \approx K$

Computing α: observed disagreement

Number of coders: $\mathbf{c}$
Number of items: $\mathbf{i}$
Distance of a single pair of labels q_a, q_b*:* $\mathbf{d_{q_a q_b}}$

Observed disagreement

Number of judgment pairs per item: $\mathbf{c}(\mathbf{c}-1)$

Mean distance within item i*:* $$\frac{1}{\mathbf{c}(\mathbf{c}-1)} \sum_{q_a} \sum_{q_b} \mathbf{n}_{iq_a} \mathbf{n}_{iq_b} \mathbf{d}_{q_a q_b}$$

Mean distance within items: $$\mathrm{D_o} = \frac{1}{\mathbf{ic}(\mathbf{c}-1)} \sum_{i} \sum_{q_a} \sum_{q_b} \mathbf{n}_{iq_a} \mathbf{n}_{iq_b} \mathbf{d}_{q_a q_b}$$

Computing α: expected disagreement

Number of coders: $\mathbf{c}$
Number of items: $\mathbf{i}$
Distance of a single pair of labels q_a, q_b*:* $\mathbf{d_{q_a q_b}}$

Expected disagreement:

Total number of judgment pairs: $\mathbf{ic}(\mathbf{ic} - 1)$

Overall mean distance:

$$D_e = \frac{1}{\mathbf{ic}(\mathbf{ic}-1)} \sum_{q_a} \sum_{q_b} \mathbf{n}_{q_a} \mathbf{n}_{q_b} \mathbf{d}_{q_a q_b}$$

Summary

For nominal agree/disagree distinctions, $K \approx \alpha$

- Use either coefficient

For grades of agreement, use α

- Take care with choosing the distance metric

Interpreting agreement

Agreement measures
are not
hypothesis tests

- Evaluating magnitude, not existence/lack of effect
- Not comparing two hypotheses
- No clear probabilistic interpretation

Agreement values (historical note)

Krippendorff 1980, page 147:

> *In a study by Brouwer et al. (1969) we adopted the policy of reporting on variables only if their reliability was above .8 and admitted variables with reliability between .67 and .8 only for drawing highly tentative and cautious conclusions. These standards have been continued in work on cultural indicators (Gerbner et al., 1979) and might serve as a guideline elsewhere.*

Carletta 1996, page 252:

> *[Krippendorff] says that content analysis researchers generally think of $K > .8$ as good reliability, with $.67 < K < .8$ allowing tentative conclusions to be drawn.*

Agreement and error

Agreement metrics are difficult to understand.

Can we relate the amount of agreement to an error rate?

- Assumes existence of "correct" annotation
- Requires explicit model of annotator error

Model I: concentrated error

Error model assumptions (inspired by but different from Aickin):

- Items are either **easy** or **hard**
- Coders always agree on easy items
- Coders classify hard items at random

a: *proportion of* **easy** *items*

$$A_o = \mathbf{a} + (1 - \mathbf{a})A_e^{hard}$$

$$\mathbf{a} = \frac{A_o - A_e^{hard}}{1 - A_e^{hard}}$$

Model I: concentrated error

$$\mathbf{a} = \frac{A_o - A_e^{hard}}{1 - A_e^{hard}}$$

Additional assumption:

- $A_e = A_e^{hard}$

Interpretation: Dist. of hard judgments = dist. of easy items

Then:

$$\mathbf{a} = K \textit{ or } \alpha$$

Interpretation: K or α = proportion of principled judgments

Model II: evenly spread error

Error model assumptions:

- Fixed probability **p** of non-random judgment
- Dist. of random judgments = dist. of principled judgments

Category labels: $q_1, \ldots, q_n$

True distribution: $\mathrm{P}(q_1), \ldots, \mathrm{P}(q_n)$

Expected agreement on an item of (true) category q

$$(p + (1-p)\mathrm{P}(q))^2 + \sum_{q' \neq q} ((1-p)\mathrm{P}(q'))^2$$

Model II: evenly spread error

$$E(\mathrm{A_o}) = \sum_{q\in Q} \mathrm{P}(q)\Big[(p+(1-p)\mathrm{P}(q))^2 + \sum_{q'\neq q}((1-p)\mathrm{P}(q'))^2\Big]$$

$$= p^2 + (1-p^2)\Big[\sum_{q\in Q}(\mathrm{P}(q))^2\Big]$$

$$E(\mathrm{A_e}) \approx \sum_{q\in Q}(\mathrm{P}(q))^2$$

$$\mathbf{E(K)} \approx \frac{\left[p^2+(1-p^2)E(\mathrm{A_e})\right]-E(\mathrm{A_e})}{1-E(\mathrm{A_e})} = \mathbf{p^2}$$

Comparing the two error models

Random judgments concentrated in specific items:

$$\textit{proportion of principled judgments} = K$$

Random judgments uniformly spread among items:

$$\textit{proportion of principled judgments} = \sqrt{K}$$

The single number problem

One category prevalent: K sensitive to rare categories

	A	B	C	Total
A	92	1	1	94
B	1	0	2	3
C	1	2	0	3
Total	94	3	3	100

$$A_o = 0.92$$
$$A_e = 0.8854$$
$$K = \frac{0.92 - 0.8854}{1 - 0.8854} \approx 0.30$$

Two categories prevalent: K **ignores** rare category

	A	B	C	Total
A	46	2	1	49
B	2	46	1	49
C	1	1	0	2
Total	49	49	2	100

$$A_o = 0.92$$
$$A_e = 0.4806$$
$$K = \frac{0.92 - 0.4806}{1 - 0.4806} \approx 0.85$$

Latent Class Analysis

Model:

- Unknown number of underlying **classes**
- Each class has unique distribution for emitting category labels
- Estimate underlying probabilities from the observed labels

Allows analysis in terms of **diagnostic accuracy**:

- Probability of class given a label (or set of labels)
- Probability of labels given an underlying class