Introduction to Machine Learning NPFL 054

http://ufal.mff.cuni.cz/course/npfl054

Barbora Hladká **Martin Holub**

{Hladka | Holub}@ufal.mff.cuni.cz

Charles University,
Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics

Inter-annotator agreement (IAA) — data 2014

CRY – confusion matrix (50 instances, 33 agreements = 66 %)

		B				
		1	**4**	**7**	**u**	**x**
	1	24	3	1	3	0
	4	3	3	0	1	1
A	**7**	0	2	4	0	1
	u	1	0	0	0	0
	x	0	1	0	0	2

ENLARGE – confusion matrix (50 instances, 31 agreements = 62 %)

		B				
		1	**2**	**3**	**4**	**u**
	1	18	2	0	2	0
	2	4	7	1	4	0
A	**3**	0	0	0	0	0
	4	2	1	2	5	0
	u	0	0	0	1	1

What agreement would be reached by chance?

Example 1

Assume two annotators (A_1, A_2), two classes (t_1, t_2), and the following distribution:

	t_1	t_2
A_1	50 %	50 %
A_2	50 %	50 %

Then

- the best possible agreement is

What agreement would be reached by chance?

Example 1

Assume two annotators (A_1, A_2), two classes (t_1, t_2), and the following distribution:

	t_1	t_2
A_1	50 %	50 %
A_2	50 %	50 %

Then

- the best possible agreement is 100 %
- the worst possible agreement is

What agreement would be reached by chance?

Example 1

Assume two annotators (A_1, A_2), two classes (t_1, t_2), and the following distribution:

	t_1	t_2
A_1	50 %	50 %
A_2	50 %	50 %

Then

- the best possible agreement is 100 %
- the worst possible agreement is 0 %
- the "agreement-by-chance" *would be*

What agreement would be reached by chance?

Example 1

Assume two annotators (A_1, A_2), two classes (t_1, t_2), and the following distribution:

	t_1	t_2
A_1	50 %	50 %
A_2	50 %	50 %

Then

- the best possible agreement is 100 %
- the worst possible agreement is 0 %
- the "agreement-by-chance" *would be* 50 %

What agreement would be reached by chance?

Example 2

Assume two annotators (A_1, A_2), two classes (t_1, t_2), and the following distribution:

	t_1	t_2
A_1	90 %	10 %
A_2	90 %	10 %

Then

- the best possible agreement is

What agreement would be reached by chance?

Example 2

Assume two annotators (A_1, A_2), two classes (t_1, t_2), and the following distribution:

	t_1	t_2
A_1	90 %	10 %
A_2	90 %	10 %

Then

- the best possible agreement is 100 %
- the worst possible agreement is

What agreement would be reached by chance?

Example 2

Assume two annotators (A_1, A_2), two classes (t_1, t_2), and the following distribution:

	t_1	t_2
A_1	90 %	10 %
A_2	90 %	10 %

Then

- the best possible agreement is 100 %
- the worst possible agreement is 80 %
- the "agreement-by-chance" *would be*

What agreement would be reached by chance?

Example 2

Assume two annotators (A_1, A_2), two classes (t_1, t_2), and the following distribution:

	t_1	t_2
A_1	90 %	10 %
A_2	90 %	10 %

Then

- the best possible agreement is 100 %
- the worst possible agreement is 80 %
- the "agreement-by-chance" *would be* 82 %

What agreement would be reached by chance?

Example 3

Assume two annotators (A_1, A_2), two classes (t_1, t_2), and the following distribution:

	t_1	t_2
A_1	90 %	10 %
A_2	80 %	20 %

Then

- the best possible agreement is

What agreement would be reached by chance?

Example 3

Assume two annotators (A_1, A_2), two classes (t_1, t_2), and the following distribution:

	t_1	t_2
A_1	90 %	10 %
A_2	80 %	20 %

Then

- the best possible agreement is 90 %
- the worst possible agreement is

What agreement would be reached by chance?

Example 3

Assume two annotators (A_1, A_2), two classes (t_1, t_2), and the following distribution:

	t_1	t_2
A_1	90 %	10 %
A_2	80 %	20 %

Then

- the best possible agreement is 90 %
- the worst possible agreement is 70 %
- the “agreement-by-chance” *would be*

What agreement would be reached by chance?

Example 3

Assume two annotators (A_1, A_2), two classes (t_1, t_2), and the following distribution:

	t_1	t_2
A_1	90 %	10 %
A_2	80 %	20 %

Then

- the best possible agreement is 90 %
- the worst possible agreement is 70 %
- the "agreement-by-chance" *would be* 74 %

Example in R

The situation from Example 3 can be simulated in R

```
# N will be the sample size
> N = 10^6

# two annotators will annotate randomly
> A1 = sample(c(rep(1, 0.9*N), rep(0, 0.1*N)))
> A2 = sample(c(rep(1, 0.8*N), rep(0, 0.2*N)))

# percentage of their observed agreement
> mean(A1 == A2)
[1] 0.740112

# exact calculation -- just for comparison
> 0.9*0.8 + 0.1*0.2
[1] 0.74
```

Cohen's kappa

Cohen's kappa was introduced by Jacob Cohen in 1960.

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

- $Pr(a)$ is the relative observed agreement among annotators
  $=$ percentage of agreements in the sample
- $Pr(e)$ is the hypothetical probability of chance agreement
  $=$ probability of their agreement if they annotated randomly
- $\kappa > 0$ if the observed agreement is better than what would be expected by chance

Limitations

- Cohen's kappa measures agreement between two annotators only
- for more annotators you should use the more general Fleiss' kappa
  – see `http://en.wikipedia.org/wiki/Fleiss'_kappa`

Inter-annotator agreement (2014)

CRY
Number of agreements: 33 (66 %)
Number of disagreements: 17 (34 %)
Cohen's kappa: 0.437
Fleiss's kappa: 0.434

ENLARGE
Number of agreements: 31 (62 %)
Number of disagreements: 19 (38 %)
Cohen's kappa: 0.438
Fleiss's kappa: 0.433

Inter-annotator agreement (2015)

CRY – **Cohen's kappa**

	A	B	C	D
A	–	0.36	0.28	0.41
B	–	–	0.37	0.41
C	–	–	–	0.33
D	–	–	–	–

ENLARGE – **Cohen's kappa**

	A	B	C	D
A	–	0.31	0.41	0.30
B	–	–	0.22	0.32
C	–	–	–	0.37
D	–	–	–	–

CRY – **Fleiss's kappa** 0.35
ENLARGE – **Fleiss's kappa** 0.32

Automatic classifier – training error analysis ENLARGE (2014)

		GS 1	2	3	4	u
	1	224	1	1	12	2
	2	2	17	3	0	0
C	**3**	1	2	15	0	0
	4	3	1	0	14	1
	u	0	0	1	0	0

	GS 1	2	3	4	u
1	0.97	0.05	0.05	0.46	0.67
2	0.01	0.81	0.15	0.00	0.00
3	0.00	0.10	0.75	0.00	0.00
4	0.01	0.05	0.00	0.54	0.33
u	0.00	0.00	0.05	0.00	0.00

Number of agreements: 270 (90 %)
Number of disagreements: 30 (10 %)

A + B error analysis – ENLARGE (2014)

		GS				
		1	**2**	**3**	**4**	**u**
	1	46	0	0	0	0
	2	11	14	0	1	0
A+B	**3**	3	0	0	0	0
	4	12	0	0	10	0
	u	0	0	0	1	2

	GS				
	1	**2**	**3**	**4**	**u**
1	0.64	0.00	0.00	0.00	0.00
2	0.15	1.00	0.00	0.08	0.00
3	0.04	0.00	0.00	0.00	0.00
4	0.17	0.00	0.00	0.83	0.00
u	0.00	0.00	0.00	0.08	1.00

Number of agreements: 72 (72 %)
Number of disagreements: 28 (28 %)

Summary of manual annotation data analysis + Examination Requirements

You should be able to practically compute and understand/use

- categorical data distribution
- confusion matrices
- classifier accuracy
- inter-annotator agreement
  - simple percentage
  - Cohen's kappa
- probability (both conditional and unconditional) of errors of different types

Practical exercises in R

- Download two files with annotated data cry-A.csv and cry-C.csv.
  - `https://ufal.mff.cuni.cz/courses/npfl054/demo`
- Run R and read the data using `read.csv()`.
  - Hint: see the posted Tutorial, Part I.
  - ... and create objects `cry.A` and `cry.C`.
- Make the confusion matrix between groups A and C.
  - Hint: use `table(cry.A$class, cry.C$class)`
- Compute simple agreement (in percentage) between A and C.
  - Hint: use `diag()` and `sum()`
- compute the Cohen's kappa value between groups A and C.
  - For hints see Part III of the Tutorial.

Homework

- Go through all details in the Tutorial (Parts I, II, and III)
- Get familiar with the `data.table` package
  – just to understand Part II
- Do all exercises in Part III