Recherche Zen
Session 3 : Experiments

Carlos Ramisch and Manon Scholivet
Partly based on the course by Adeline Paiement

March 29 2023

Plan

Research question $\rightarrow$ Experiment

Data creation

Data science experiments

Evaluation metrics

Why do we need experiments ?

- A research question and its sub-questions

  $\longrightarrow$ Precise, concise, feasible, interesting
- Hypotheses related to each sub-question
- They are anchored in the litterature and justified

Why do we need experiments ?

- A research question and its sub-questions

  → Precise, concise, feasible, interesting
- Hypotheses related to each sub-question
- They are anchored in the litterature and justified

Experiment goals

1. To build further evidence that will eventually lead to accepting or rejecting the hypothesis
2. Lead to new interesting research questions

Designing an experiment

1. Identify the target hypothesis

   → Prioritise hypotheses wrt. impact and constraints

2. Identify the needs of the experiment

   → Data, datasets, evaluation metrics

3. Instantiate under-specified aspects of the question/hypotheses

   → The devil is in the details

4. If the result is X, I will be able to conclude Y

   → Reformulate hypotheses in terms of experiment outcomes

Experiment design : example

Hypothesis

It is possible to learn a model for language L (with no annotations available) from a set of languages L' (with available annotations)

Refining the hypothesis :

- A model for which task ? Question answering ? Parsing ?
- A supervised or unsupervised model ?
- What exact set of languages ?
- What configurations will be tested ?
  - $\rightarrow$ L' contains 1 language, 5 languages. . .
  - $\rightarrow$ L is similar to a language in L' or not ?
- How to assess if the model for L is good ?
  - $\rightarrow$ Evaluation metrics

Scope

- Experiments in computer science
- Expriments using data
- $\Longrightarrow$ Experiments in data science

Scope

- Experiments in computer science
- Expriments using data
- $\Longrightarrow$ Experiments in data science

Data science

Is data science a science ?

Scope

- Experiments in computer science
- Expriments using data
- $\implies$ Experiments in data science

Data science

Is data science a science ?

Disclaimer : This is not a machine learning course

Experimental protocol

- Step-by-step description of the experiment
- "Algorithm" of the experiment

How formal is your protocol ?

- Depends on the discipline
- A good protocol description can speed up paper writing
- In any case, to be defined before launching experiments

Making choices

- Beware of the combinatorial explosion

  $\rightarrow$ # datasets $\times$ # configs $\times$ # models $\times$ # metrics

  $\rightarrow$ Grid search = experiments run forever

- Choices must be justified

  $\rightarrow$ An arbitrary justification is better than none

  $\rightarrow$ E.g. *the parameter was chosen after trial and error*

- Favour more promising aspects

  $\rightarrow$ E.g. Metrics are more or less equivalent $\implies$ choose one

  Datasets are heterogeneous $\implies$ test all of them

  $\rightarrow$ Small pilot experiments $\implies$ trends $\implies$ choices

Plan

Research question $\rightarrow$ Experiment

Data creation

Data science experiments

Evaluation metrics

Where does data come from ?

- Supervised methods require :
  - input x + associated gold/reference prediction y

- Machine learning / NLP courses :

```
from sklearn.datasets import load_digits
digits = load_digits()
print(digits.target[:20]) # magic !
```

- Real life :
  - Here's some data (x), apply some learning on it !

Where does data come from ?

- Supervised methods require :
  - input x + associated gold/reference prediction y
- Machine learning / NLP courses :

```
from sklearn.datasets import load_digits
digits = load_digits()
print(digits.target[:20]) # magic !
```

- Real life :
  - Here's some data (x), apply some learning on it !

Question

- How to obtain gold predictions y ?
  - supervision to learn models
  - reference to evaluate models

Data annotation recipe

1. Select material to annotate
   - licence, biases, representativity, diversity
2. Write annotation guidelines
   - domain expertise, pilot annotation
3. Develop or adapt an annotation platform
   - adaptable, easy to use
4. Train annotators
   - hard cases, speed, biases
5. Evaluate quality
   - inter-rater agreement
6. Combine annotations
   - adjudication, averaging
7. Export and release
   - stable website, format, documentation, articles

Data selection for annotation

- Similarity with target application data
- Trade-off between realistic vs. artificial

  $\rightarrow$ E.g. newspaper vs. tweets
- Raw data is noisy $\implies$ harder to annotate/exploit

  $\rightarrow$ E.g. dialects, typos, code switching, slang

Example : Text crawling

- Dedicated web-based corpus tools : BootCat, Sketch
    - parallelisation, `robots.txt`, priority queue, loops
- Start from pre-downloaded web dumps : CommonCrawl
- Pre-processing and cleaning
    1. Language identification
        - Document, paragraph, sentence level
    2. Deduplication
        - N-gram hashing : Onion
    3. Text extraction
        - HTML $\rightarrow$ text : Beautiful Soup
        - Boilerplate removal : jusText
    4. Content filtering
        - Length, stopword ratio, dictionary
    5. Paragraph/sentence segmentation, tokenisation

Personal data

- Anonymisation :
  - Remove all information which allows identifying individuals
  - Aggregate, shuffle
- Pseudo-anonymisation/De-identification
  - Remove identity-related information (name, phone, email)
  - Analysis/crossing could recover individuals identities
- **In practice** : complete anonymisation is barely impossible

  Example : DECODA corpus (RATP call center transcriptions)

```
et ma carte vitale et tout
tout tout tout quoi c’ est c’ est à quel nom s’il vous plaît
NNAAMMEE ça s’ écrit NNAAMMEE
ouais
NNAAMMEE
ah c’ est ça NNAAMMEE
voilà
```

Indirect annotation

- Clever way to select data
- Europarl : text + translation
  - translations provided by EU
- Open Subtitles : text + translation
  - provided for free by series/movie fans
- CNN/Daily Mail : text + summary
  - News header as its summary
- Amazon products : text + polarity (positive/negative)
  - Reviews associated with 5-star rating
- Flickr30k : image + description
  - Captions provided by users on Flickr

Annotation guidelines

- Detailed definition of the task :
  - Summarise a text
  - Identify epidemiology events in news
  - Underline named entities

Annotation guidelines

- Detailed definition of the task :
  - Summarise a text

    → how many words/sentences, style, target public, entities
  - Identify epidemiology events in news

    → date, place, pathology agent, events per document
  - Underline named entities

    → categories, span, nesting, metonymy
- As objective as possible :

  → Definitions, notation conventions

  → Yes/no tests, decision trees, flowcharts

Annotation guidelines

- Detailed definition of the task :
  - Summarise a text

    → how many words/sentences, style, target public, entities
  - Identify epidemiology events in news

    → date, place, pathology agent, events per document
  - Underline named entities

    → categories, span, nesting, metonymy
- As objective as possible :

  → Definitions, notation conventions

  → Yes/no tests, decision trees, flowcharts
- Borderline cases

  → Discard input x

  → Arbitrary but consistent decision
- Many examples !
- Several pilot annotation campaigns

Annotation guidelines example : PARSEME

- ↳ Apply test S.1 - [**1HEAD**: Unique verb as functional syntactic head of the whole?]
  - ↳ **NO** ⇒ Apply the VID-specific tests ⇒ *VID tests positive?*
    - ↳ **YES** ⇒ Annotate as a VMWE of category **VID**
    - ↳ **NO** ⇒ It is not a VMWE, **exit**
  - ↳ **YES** ⇒ Apply test S.2 - [**1DEP**: *Verb v has exactly one lexicalized dependent d?*]
    - ↳ **NO** ⇒ Apply the VID-specific tests ⇒ *VID tests positive?*
      - ↳ **YES** ⇒ Annotate as a VMWE of category **VID**
      - ↳ **NO** ⇒ It is not a VMWE, **exit**
    - ↳ **YES** ⇒ Apply test S.3 - [**LEX-SUBJ**: *Lexicalized subject?*]
      - ↳ **YES** ⇒ Apply the VID-specific tests ⇒ *VID tests positive?*
        - ↳ **YES** ⇒ Annotate as a VMWE of category **VID**
        - ↳ **NO** ⇒ It is not a VMWE, **exit**
      - ↳ **NO** ⇒ Apply test S.4 - [**CATEG**: *What is the morphosyntactic category of d?*]
        - ↳ **Reflexive clitic** ⇒ Apply IRV-specific tests ⇒ *IRV tests positive?*
          - ↳ **YES** ⇒ Annotate as a VMWE of category **IRV**
          - ↳ **NO** ⇒ It is not a VMWE, **exit**
        - ↳ **Particle** ⇒ Apply VPC-specific tests ⇒ *VPC tests positive?*
          - ↳ **YES** ⇒ Annotate as a VMWE of category **VPC.full** or **VPC.semi**
          - ↳ **NO** ⇒ It is not a VMWE, **exit**

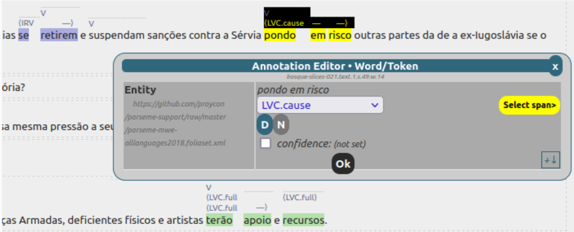
Source: `https://parsemefr.lis-lab.fr/parseme-st-guidelines/`

Double annotation

- Two (expert/trained) annotators :
  - same trainig, same annotation guidelines
  - annotate the same data
    - no communication while annotating
- Results should be (almost) identical
  - Inter-annotator agreement
  - Adjudication
- High agreement : guide OK, training OK, data quality OK
- Low agreement : restart until high agreement is reached
- "Low" and "High" $\rightarrow$ Numerical agreement score

Annotation interface

- Generic tools for text
  - Inception, webAnno, brat, FLAT, Arborator
  - Require configuration and administration
- Task-specific interfaces
  - Web forms



Inter-annotator agreement (IAA) : framework

Items, categories and coders :

- Set of *items* : $\{i|i \in I\}$ and is of cardinality i
- Set of *categories* : $\{k|k \in K\}$ and is of cardinality k
- Set of *coders* (annotators) : $\{c|c \in C\}$ is of cardinality c

Counting annotations :

- n_{ik} number of coders who assigned item i to category k
- n_{ck} number of items assigned by coder c to category k
- n_k total number of items assigned by all coders to category k

Source: Artstein and Poesio, 2005

Inter-annotator agreement (IAA)

- Simple case : two raters c_1 and c_2
- Observed agreement : proportion of identically annotated items

$$A_O = \frac{1}{\mathrm{i}} \sum_{k \in K} \delta(n_{1k}, n_{2k})$$

Item	Annot1	Annot2
1	Green	Blue
2	Blue	Blue
3	Blue	Green
4	Green	Green
5	Blue	Blue
6	Blue	Blue
	...	...

Contingency table

	Green	Blue	Total
Green	41	3	44
Blue	9	47	56
Total	50	50	100

$$A_O = \frac{41 + 47}{100} = 0.88$$

Adapted from Ron Artstein's slides :

http://ron.artstein.org/publications/2012-artstein-agreement-slides.pdf

Chance-corrected agreement

Task : diagnose whether patients are ill

	Healthy	Ill	Total
Healthy	990	5	995
Ill	5	0	5
Total	995	5	1000

$$A_O = \frac{990}{1000} = 0.99$$

- Most patients are not ill
  - No agreement in ill" category
- High expected agreement A_E
  - How to estimate A_E ?

Kappa inter-annotator agreement

- Proportion of agreement above chance

$$\kappa = \frac{A_O - A_E}{1 - A_E}$$

- Assume each annotator has their distribution

$$A_E^{\kappa} = \frac{1}{\text{¡}^2} \sum_{k \in K} n_{c_1 k} n_{c_2 k}$$

- i annotated items in total,
- K possible values per item,
- $n_{c_j k}$ items annotated as k by rater c_j

Adapted from Ron Artstein's slides :

http://ron.artstein.org/publications/2012-artstein-agreement-slides.pdf

Exercise : calculate kappa

	Healthy	Ill	Total
Healthy	990	5	995
Ill	5	0	5
Total	995	5	1000

- i = 1000 annotated items in total,
- $n_{c_j k}$ items annotated as k by rater c_j

$$A_O = \frac{990}{1000} = 0.99 \qquad \kappa = \frac{A_O - A_E}{1 - A_E} \qquad A_E^{\kappa} = \frac{1}{\mathrm{i}^2} \sum_{k \in K} n_{c_1 k} n_{c_2 k}$$

1. Calculate the kappa chance-corrected IAA score

Exercise : calculate kappa

	Healthy	Ill	Total
Healthy	990	5	995
Ill	5	0	5
Total	995	5	1000

- i = 1000 annotated items in total,
- $n_{c_j k}$ items annotated as k by rater c_j

$$A_O = \frac{990}{1000} = 0.99 \qquad \kappa = \frac{A_O - A_E}{1 - A_E} \qquad A_E^\kappa = \frac{1}{\mathrm{i}^2} \sum_{k \in K} n_{c_1 k} n_{c_2 k}$$

1. Calculate the kappa chance-corrected IAA score

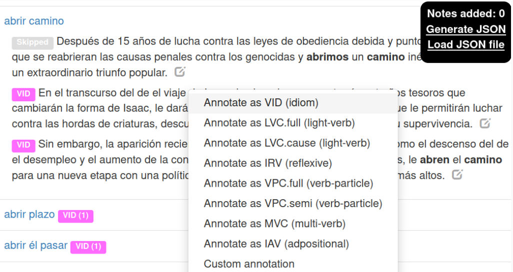
$$A_E = \frac{995^2 + 5^2}{1000^2} = 0.995^2 + 0.005^2 = 0.99005 \qquad A_O = 0.99 \qquad \kappa = -0.005$$

More complex cases

- More than 2 raters
    - Consider pairs of agreeing annotators
- Sporadic annotations
    - F-score between raters

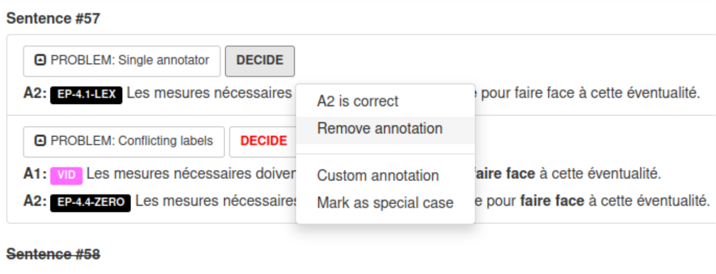
Consistency checks

- Vertical data visualisation
    - Aggregate similar units (e.g. by lemma, POS n-gram, etc)
- Adjudicator of expert annotator corrects mistakes



Adjudication

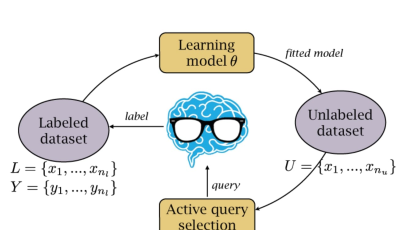
- Carried out by another expert (not an annotator)
- Dedicated interface
- Documented conflict resolution strategies



- Creation of final (adjudicated) dataset

Automatic pre-annotation

- Pre-annotation
  1. Annotate a small dataset and train predictive model
  2. Predict on the remaining unlabelled data
  3. Correct the predictions
- **Active learning**
  1. Annotate a given instance
  2. Append to training data and train predictive model
  3. Next instance to annotate chosen automatically
     - Maximise diversity of phenomena
     - Maximise the utility for the model



Crowdsourcing

- Compensate for subjectivity = averave over many annotators
  - Amazon Mechanical Turk, Crowdflower, ...
- Make the task simpler - accessible for non experts
  - Remuneration per HIT - Human Intelligence Task
- Data quality
  - Qualification pre-task, spammer filtering
- Ethical aspects : unfair remuneration, hard work

Gamification

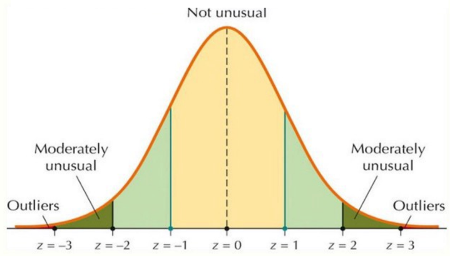
- Games with a purpose
  - Fun, visually attractive, competition
  - Background : free annotation
- Examples
  - Jeux de mots `https://www.jeuxdemots.org/`
  - ZombiLingo `http://gwap.grew.fr/`

Data cleaning

- Some annotations are outliers
- Cleaning must occur before experiments

Z-score filtering

Remove annotations that are more than z standard deviations away from the mean



Source: Further reading : `https://aclanthology.org/W16-1804/`

Data splitting i

- Evaluation must be carried out on held out data
  - → Test dataset
- Development must be carried out on held out data
  - → Development or validation dataset
  - → Attention : it is extremely easy to accidentally tune on test data
- Paramters must be learned from data
  - → Training dataset

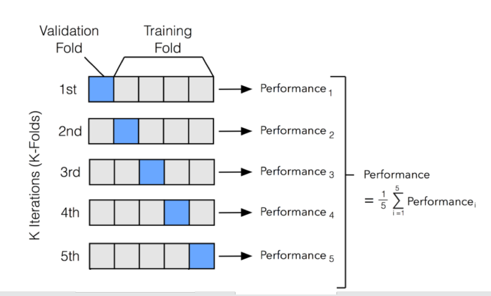
Data splitting ii

Fixed split

- Randomly pick 10% for test, 10% for dev, 80% for train
- Comparable across experiments, papers

Data splitting iii

k-fold cross validation



- Expensive : requires training k models instead of 1

Data splitting iv

Biased split

- Fixed split, but not random
- The test set has controlled characteristics

  $\longrightarrow$ E.g. test instances are unseen in training data

Data splitting v

Discussion

- *We need to talk about standard splits*

  $\longrightarrow$ `https://aclanthology.org/P19-1267/`
- *We need to talk about random splits*

  $\longrightarrow$ `https://aclanthology.org/2021.eacl-main.156/`
- ...

Understand the data

- Open your files !

  $\rightarrow$ Otherwise someone may troll you :

  https://medium.com/@yoav.goldberg/

  an-adversarial-review-of-adversarial-generation-of-natural-language-409ac
- Don't try to get blood from a turnip

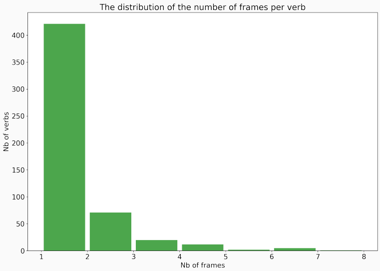
  $\rightarrow$ Maybe your prediction task is unrealistic

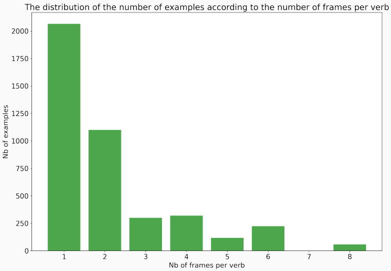
  $\rightarrow$ Maybe you need external resources

  $\rightarrow$ ...

Data analysis

- Distribution of classes, input characteristics
- Useful tool : histogram (e.g. `matplotlib.pyplot.hist`)
  - $\longrightarrow$ Use bins to discretise real-valued attributes





Source: Author : Anna Mosolova

Annotation beyond dataset creation

- Annotating = understanding your problem
  - → Hard for humans ? $\Longrightarrow$ maybe hard for models
  - → Low agreement $\Longrightarrow$ maybe ill-defined problem
  - → Annotation guidelines $\Longrightarrow$ inspiration for features

Plan

Experimental conditions

- Supervised, unsupervised, semi-supervised
- Generalisation and amount of supervision

  → Zero-shot, one-shot, few-shot
- Model's (hyper-)parameters

  → E.g. Neural network architecture, dimensions, ...

  → E.g. Clustering linking criterion, threshold

Baseline and topline i

- A model is never good or bad per se
- Situate the model performance wrt. a simpler model
  - $\rightarrow$ Baseline – simple model for the task
- Examples of baseline
  - $\rightarrow$ Random prediction
  - $\rightarrow$ Majoritary class
  - $\rightarrow$ A good model 5 years ago
  - $\rightarrow$ An interpretable model (rules, thresholds)
  - $\rightarrow$ State-of-the-art model published last month

Baseline and topline ii

- Situate the model performance wrt. a better model

  $\rightarrow$ Topline – upper bound for the performance

- Examples of topline

  $\rightarrow$ State-of-the-art model published last month

  $\rightarrow$ Large model released by big tech company

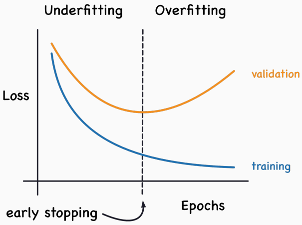
  $\rightarrow$ Human annotator performance/agreement

  $\rightarrow$ Same experiment in unrealistic (easy) condition

Overfitting

- The model "overfits" if it memorises the training set
- Tools to prevent overfitting
  - Rule of thumb of pre-neural models :

    $\rightarrow$ Less features than data items
  - Learning curves on dev set
  - Early stopping based in dev set performance



Hyperparameter search

- Some important hyperparameters
  - learning rate
  - epochs/early stopping patience
  - batch size
  - dropout ratios
  - model capacity (hidden layer dimensions)
  - number of stacked layers, attention heads
  - embedding size
- Tuning strategies
  - Grid search
  - Bayesian search
  - Random search
  - ...
- Unavoidable but usually not very intersting

Model instability

- Same hyperparameters, different random seeds
  - weight initialisaiton in fine-tuning layers
  - order of inputs/batches
- Substantially different results
  - Some data orders/initializations consistently better than others
  - Early stopping is effective
- **Report averages, error bars, confidence intervals**
  - Re-run training several times with different orders/random initialisation seeds

Source: Further reading : `https://arxiv.org/abs/2002.06305`

Experiment management

- Logbook
  - experimental conditions for each result
  - raw results and links to results
  - write down ideas, hypotheses, etc.
- Experiment management platform
  - Tensorboard, RayTune, MLFlow, Lightning
- Git : branches, merge requests, CI for testing
- Overleaf : collaborative LaTeX article writing

Reproducibility vs. replicability

- Results are **reproducible**
  - Data available under open licences
  - Model/code shared under open licences
  - Parameters and hyperparameters described
  - Computational requirements reasonable
- Results are **replicable**
  - Robust to other datasets
  - Robust to different experimental conditions
  - Robust across conditions

<u>Source</u>: `https://acl-reproducibility-tutorial.github.io/`

Plan

Research question → Experiment

Data creation

Data science experiments

Evaluation metrics

Disclaimer : all metrics are incomplete

- Ideally : measure a hidden variable or phenomenon
- In practice : measure what we can observe

  $\longrightarrow$ Formulation is simple enough to be interpretable
- Metrics are partial views of the results

Accuracy

Binary detection / classification

$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn}$$

- Percentage of well classified points
- Incomplete description of the method's performance
- Be careful ! Problem if class sizes are very unequal

[Image : Devin Soni, towardsdatascience.com]

Accuracy is an average

- Data items seen as n i.i.d. Bernoulli variables $Y_1 \dots Y_n$
  - $\rightarrow$ $Y_i = 0$ if prediction is wrong
  - $\rightarrow$ $Y_i = 1$ if prediction is correct
- Expected value of such variables is p (success probability)
- Expected value can be estimated by the mean :

$$E[Y_i] \approx \frac{1}{n} \sum_{i=1}^{n} Y_i$$

- This is precisely the definition of accuracy !
  - $\rightarrow$ Accuracy is normally distributed (CLT)

Precision, recall, F-score

- Binary detection / classification
- Precision/recall : Complementary measures, report both !
  - Precision
    $\rightarrow tp/(tp + fn)$
  - Recall = Sensitivity
    $\rightarrow tp/(tp + fn)$
  - Specificity :
    $\rightarrow tn/(tn + fp)$
- F-score : Harmonic mean of precision and recall

$$F = 2.\frac{precision.recall}{precision + recall}$$

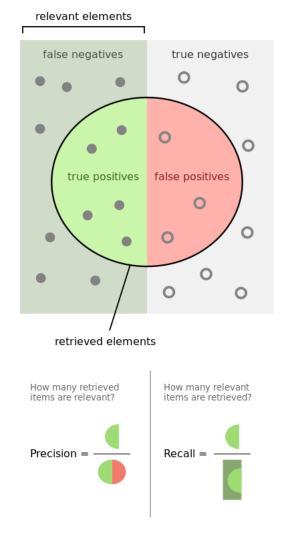


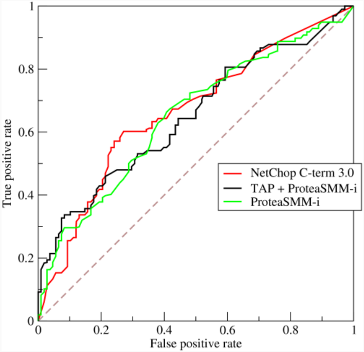
Image from Wikipedia

Precision is not an average

- Recall can be seen as an average like accuracy
- Precision cannot be seen as an average
  - $\rightarrow$ The denominator depends on the model
  - $\rightarrow$ Models class distribution is unpredictable
- $\implies$ F-score cannot be assumed to be normally distributed

ROC curve

ROC curves (*Receiver Operating Characteristic*) are very useful to chose a threshold.

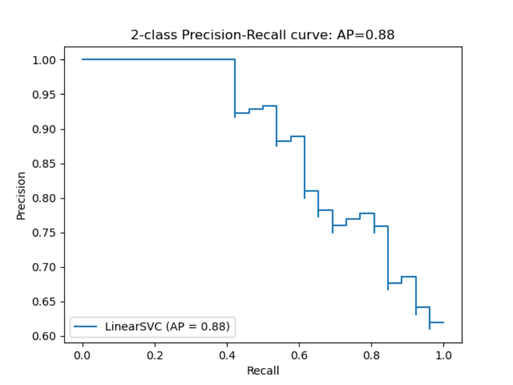


The AUC (*Area Under ROC*) is often used to estimate the model skill.

Image from Wikipedia

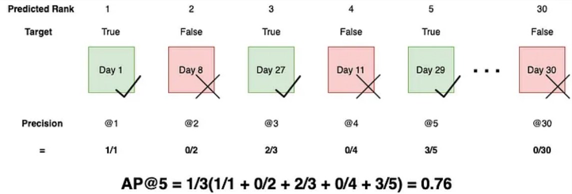
Precision-recall curve

Another way to do this is to use the Precision and the Recall instead of using the True positive and the False positive rates.



Mean average precision

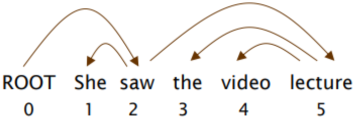
- Model predicts a numerical score
- Gold class is binary or discrete
- Evaluate without setting a fixed threshold



Structured prediction

- How to compare structured objects ?
  - $\longrightarrow$ Sub-sequences
  - $\longrightarrow$ Clusters
  - $\longrightarrow$ Syntax trees
  - $\longrightarrow$ Graphs

Structured prediction example : LAS/UAS



$$\text{Acc} = \frac{\#\text{ correct deps}}{\#\text{ of deps}}$$

"unlabelled attachment score"

UAS = 4 / 5 = 80%

LAS = 2 / 5 = 40%

"labelled AS"

Gold

1	2	She	nsubj
2	0	saw	root
3	5	the	det
4	5	video	nn
5	2	lecture	obj

Parsed

1	2	She	nsubj
2	0	saw	root
3	4	the	det
4	5	video	nsubj
5	2	lecture	ccomp

Source: `https://x-wei.github.io/xcs224n-lecture5.html`

Goodhart's law

"When a measure becomes a target, it ceases to be a good measure"

- Risk : optimise evaluation metric at any expense
  - $\rightarrow$ Overfitting, low generalisation
  - $\rightarrow$ Forgetting the research question
  - $\rightarrow$ Frustration with unrealistic goals
  - $\rightarrow$ ...

Source: Thanks to François Hamonic for this slide.

Sources

- Cours d'Adeline Paiement
- Wikipedia
- Google images