What is Machine Learning

Research Skills: Machine Learning

Grzegorz Chrupała
g.chrupala @ uvt.nl

Instructors

- Grzegorz Chrupała (main instructor)
  - g.chrupala@uvt.nl
  - Room D340
  - Computational linguistics, applied machine learning
- Ákos Kádár (teaching assistant)
  - PhD student – language and vision
- Afra Alishahi (guest lecturer)
  - Computational models of language acquisition
  - Guest lecture on March 11

Practical Matters

Pre-requisites

- You need basic programming skills in Python, for example:
  - Data processing (Use)
  - Seminar data processing
- If you know **no** programming, this course will be hard (probably **too** hard)

Attendance

- Attendance is not strictly enforced, BUT
- Each session will start with a mini-quiz
  - You need to get a pass in at least 50% of them
- Slides are not meant to be self-contained
  - Take notes!

Individual assignments: programming exercises

- Two **individual** assignments
- Assignment 1
  - 20% course grade
  - Due **February 23**
- Assignment 2
  - 30% course grade
  - Due **March 16**
- YOUR own work – no collaboration

Group assignment:
Machine Learning Project

- Sign up for groups (up to 3 people) on BB
- Assignment 3 – project proposal
  - 10% final grade
  - due **March 1**
- Assignment 3 – final submission
  - 40% final grade
  - due **April 1**
- Collaborative work
  - You will need to describe work division and contribution of each student

ML project

- Define problem to solve
- Obtain dataset
- Run appropriate computational learning experiments
- Report findings

Course forum

- Subscribe to the course forum on BlackBoard
- Ask any question regarding course content and organization
- Try to answer fellow students' questions
- Ákós and me will be monitoring the forum
- Most active participants can gain a **+0.5 bonus** to add to their course grade

Resources

Textbooks

- A course in Machine Learning. Hal Daumé III. http://ciml.info/
  - This book is a work in progress but the parts which are finished give a good introductory overview of ML.
- Data Smart: Using Data Science to Transform Information into Insight. John W. Foreman. http://amzn.com/111866146X
  - Doing data science in a spreadsheet.

Tutorials and courses

- Scikit-learn tutorial materials.
  - May be hard to follow for the less advanced students. http://scikit-learn.org/stable/tutorial/index.html
- The Coursera course on Machine Learning with Andrew Ng
  - Matlab/Octave rather than Python. https://www.coursera.org/course/ml

Datasets

- See `Course Information | Datasets` on BlackBoard
- If you know of, or come across, an interesting dataset, let us know.

How can we automate problem solving?

For example: flagging spam in your e-mail

Your $5,000,000 released
Lottery Winner
get Exclusive aMedzOnline
Get 15% discount on all promotions
Free Lotto winning Notification
No love failure risk
Do you want to impress your wife
Please helpme,writing from Sick
Your overdue payment

Thesis evaluation
Link to the data
Call for papers
NAACL 2015 – author response
Invitation to program committee
changes in allotment of tasks staff
Thank you from Wikimedia
MPI-INF Distinguished Lecture
LangSci Colloquium

Rules

If (A or B or C) and not D, then SPAM

Learning from examples

- Find examples of SPAM and non-SPAM
- Come up with a **learning algorithm**
- A learning algorithm infers rules from examples
- These rules can then be applied to new data (emails)

Learning algorithms

- See several different learning algorithms
- Implement simple 2-3 simple ones from scratch in Python
- Learn about Python **libraries** for ML (**scikit-learn**)
- How to apply them to real-world problems

Machine learning – examples

- Recognize handwritten numbers and letters
- Recognize faces in photos
- Determine whether text expresses positive, negative or no opinion
- Guess person's age based on a sample of writing

Machine learning – examples

- Flag suspicious credit-card transactions
- Recommend books and movies to users based on their own and others' purchase history
- Recognize and label mentions of people's or organization names in text

Types of learning problems

Regression

- Response: a (real) number
  - Predict person's age
  - Predict price of a stock
  - Predict student's score on exam

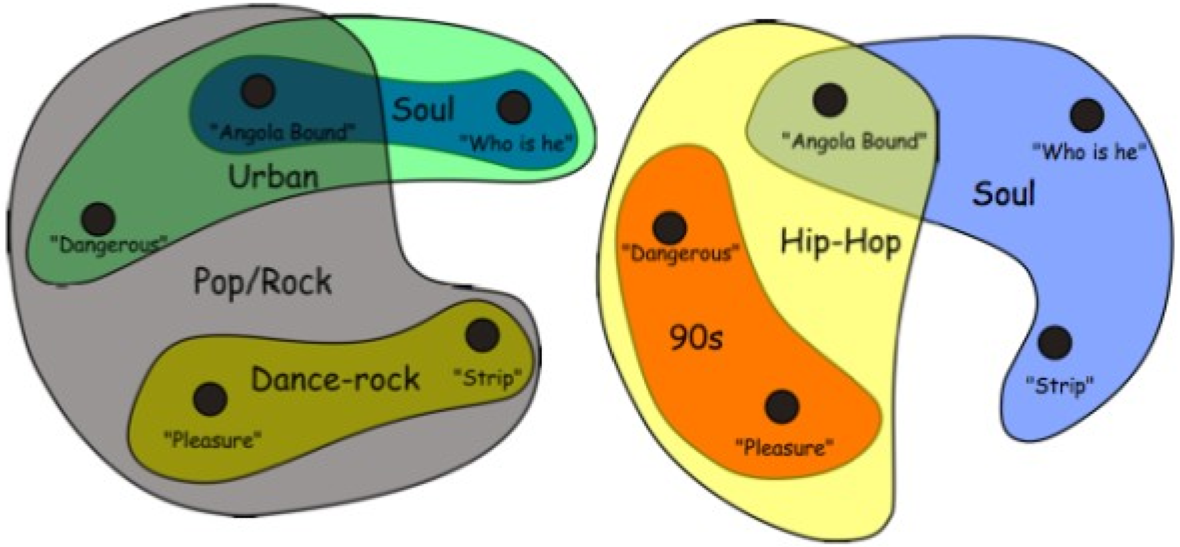
Binary classification

- Response: Yes/No answer
  - Detect SPAM
  - Predict polarity of product review: positive vs negative
  - Predict gender (simplify problem to male/female)

Source:
http://upload.wikimedia.org/wikipedia/commons/d/d2/Bird_Diversity_2011.png

Multiclass classification

- Response: one of a finite set of options
    - Classify newspaper article as
        - politics, sports, science, technology, health, finance
    - Detect species based on photo
        - Passer domesticus, Calidris alba, Streptopelia decaocto, Corvus corax, ...

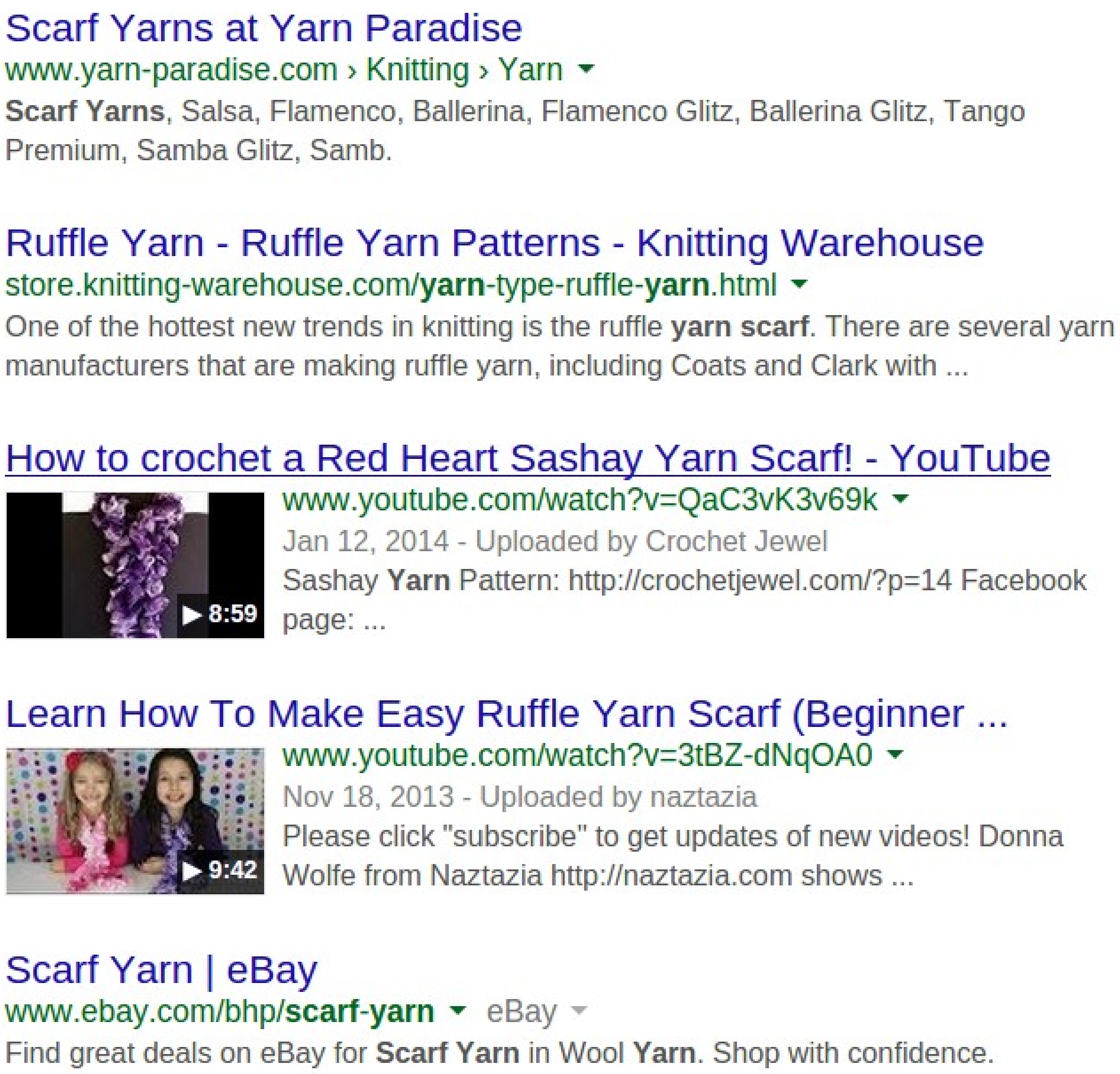


Source: https://musicmachinery.files.wordpress.com/2009/10/ismir2009-proceedings-pdf-page-372-of-775.png?w=620

Multilabel classification

- Response: a finite set of Yes/No answers
  - Assign songs to one or more genres
    - {rock, pop, metal}
    - {hip-hop, rap}
    - {jazz, blues}
    - {rock, punk}

yarn for scraves



1. **Scarf Yarns at Yarn Paradise**
www.yarn-paradise.com › Knitting › Yarn ▾
Scarf Yarns, Salsa, Flamenco, Ballerina, Flamenco Glitz, Ballerina Glitz, Tango Premium, Samba Glitz, Samb.

2. **Ruffle Yarn - Ruffle Yarn Patterns - Knitting Warehouse**
store.knitting-warehouse.com/**yarn**-type-ruffle-**yarn**.html ▾
One of the hottest new trends in knitting is the ruffle **yarn scarf**. There are several yarn manufacturers that are making ruffle yarn, including Coats and Clark with ...

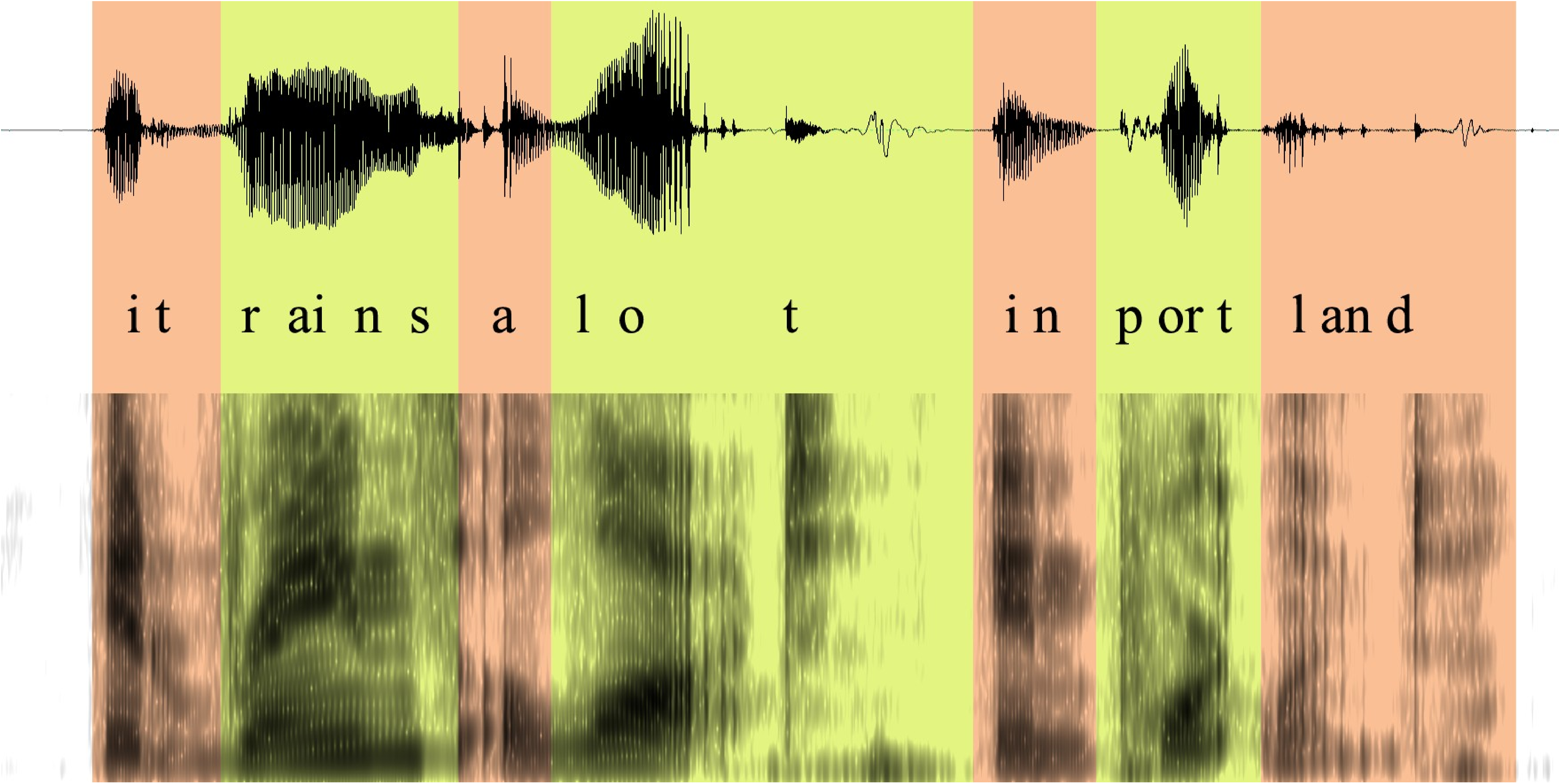
3. **How to crochet a Red Heart Sashay Yarn Scarf! - YouTube**
▶ 8:59
www.youtube.com/watch?v=QaC3vK3v69k ▾
Jan 12, 2014 - Uploaded by Crochet Jewel
Sashay **Yarn** Pattern: http://crochetjewel.com/?p=14 Facebook page: ...

4. **Learn How To Make Easy Ruffle Yarn Scarf (Beginner ...**
▶ 9:42
www.youtube.com/watch?v=3tBZ-dNqOA0 ▾
Nov 18, 2013 - Uploaded by naztazia
Please click "subscribe" to get updates of new videos! Donna Wolfe from Naztazia http://naztazia.com shows ...

5. **Scarf Yarn | eBay**
www.ebay.com/bhp/**scarf-yarn** ▾ eBay ▾
Find great deals on eBay for **Scarf Yarn** in Wool **Yarn**. Shop with confidence.

Ranking

- Order object according to relevance
  - rank web pages in response to user query
  - predict student's preference for courses in a program



Source: http://upload.wikimedia.org/wikipedia/commons/f/f1/Spectrogram_-_It_Rains_a_Lot_in_Portland.png

Sequence labeling

- Input: a sequence of elements (e.g. words)
- Response: a corresponding sequence of labels
  - Label words in a sentence with their syntactic category
    - Determiner Noun Adverb Verb Prep Noun
  - Label frames in speech signal with corresponding phonemes
    - w ɛ ð ɚ

Source: https://www.flickr.com/photos/samchurchill/8024126247/

Autonomous behavior

- Input: measurements from sensors – camera, microphone, radar, accelerometer, …
- Response: instructions for actuators – steering, accelerator, brake, ...

How well is the algorithm learning?

Evaluation

How to evaluate

- Predicting age
- Predicting gender
- Flagging spam
- ...

Predicting age – Regression

- **Mean absolute error** – the average (absolute) difference between true value and predicted value

$$MAE = \frac{1}{N}\sum_{n=1}^{N} abs(y_n - \hat{y}_n)$$

- **Mean squared error** – the average square of the difference between true value and predicted value

Predicting gender – classification

Error rate: proportion of mistakes

Flagging SPAM – classification

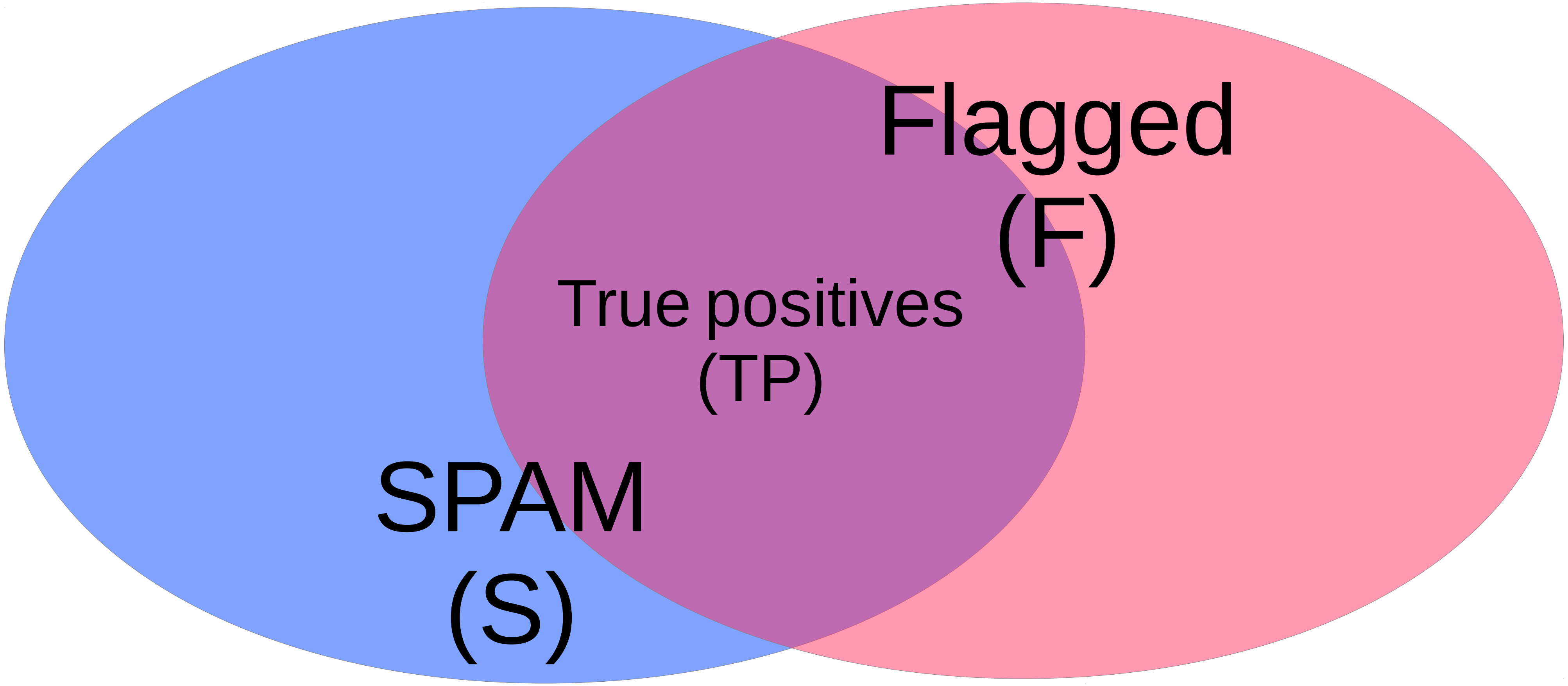
- We can use error rate
- Is there any disadvantage?

Kinds of mistakes

- False positive – Flagged as SPAM, but not non-SPAM
- False negative – Not flagged, but is SPAM
- False positives are a bigger problem!

Precision and Recall

- Metrics which focus on one kind of mistake
- Precision – what fraction of flagged emails were real SPAMs?
- Recall – what fraction of real SPAMs were flagged?



Flagged
(F)

True positives
(TP)

SPAM
(S)

Precision and Recall

- Precision

$$P = \frac{|TP|}{|F|}$$

- Recall

$$R = \frac{|TP|}{|S|}$$

F-score

- Harmonic mean between precision and recall
  - a kind of average

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R}$$

  - aka F-measure

F_β

Parameter β quantifies how much more we care about recall than precision.

$$F_\beta = (1+\beta^2) \cdot \frac{P \cdot R}{\beta^2 \cdot P + R}$$

For example $F_{0.5}$ is the metric to use if we care half as much about recall as about precision.

Exercise 1

- Define F_1 as a function of P and R in Python
- Check what scores you get for different values of P and R
- Remember that P and R are proportions – between 0.0 and 1.0

Exercise 2

Define precision and recall as Python functions

- Each functions should accept two sets as arguments
- In Python you can create a set thus:
  - `a = set(['x','y','z'])`
- You can compute intersection of sets `a` and `b` thus:
  - `c = a.intersection(b)`
- You can compute the size of the set `a` thus:
  - `len(a)`

Using examples

Imagine you're studying for a very competitive exam – how do you use learning material?

Example sets

- Training set:
  - Observe patterns, infer rules
- Development set:
  - Monitor performance, choose best learning options
- Test set:
  - REAL EXAM
  - Not accessible in advance

Summary

- Machine learning studies algorithms which can learn to solve problems from examples
- Several canonical problem types
- First step: decide on evaluation metric
- Separate training, development and test examples