UNIVERSITY AT ALBANY
STATE UNIVERSITY OF NEW YORK

CSI 436/536 (Fall 2024)

Machine Learning

Lecture 5: Elements of Machine Learning

Chong Liu

Assistant Professor of Computer Science

Sep 12, 2024

Announcement

- Homework 1 has been released this Tuesday. Start working on it with your group!

- Group member registration due today!
  - If you are on the Waitlist, you will be randomly assigned to a group tomorrow.
  - Group course project registration due next Thursday. One week to discuss your project choice.

Recap: review of probability and statistics

- Probability
  - Basic concepts
  - Probability properties
  - Random variable and distribution
  - Expectation and variance
  - Independence
  - Bernoulli distribution and Gaussian distribution
- Statistics
  - Maximum likelihood estimation

In-class exercise: Linear regression

- $P(y|x)$ is modeled by "Linear Gaussian model"

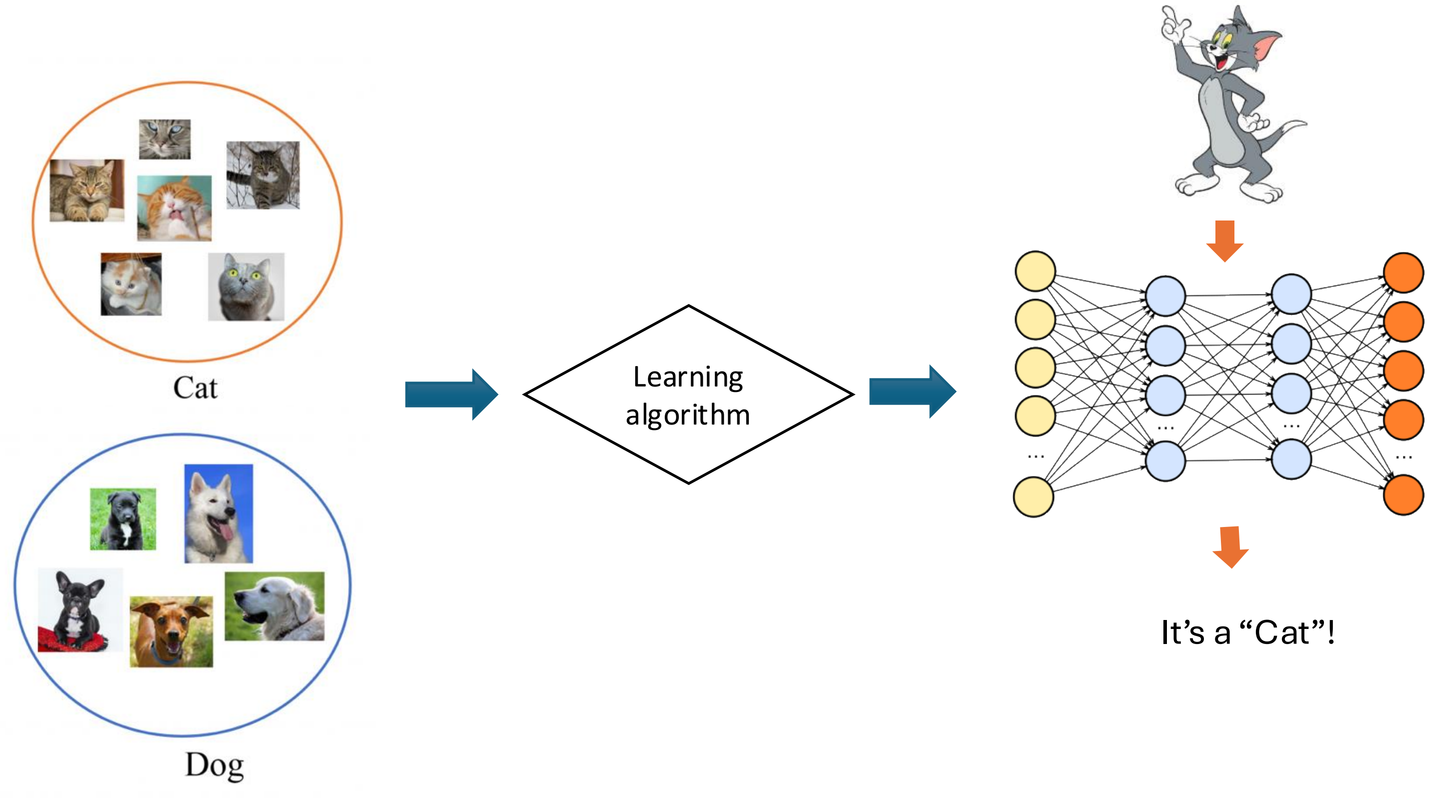
$$y_i = x_i^T \theta^* + \epsilon_i \quad \text{where } \epsilon_i \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.}$$

- Data: $(x_1, y_1), ..., (x_n, y_n)$
- Work out the optimization problem to solve for the MLE for θ^*.

Today

- Machine learning overview
- Supervised learning: Binary classification
- Feature design and feature extraction
- Example of classifier: Decision Trees

Recap: Machine learning studies "*computer programs that automatically improve (its **performance** on a task) with **experience**.*"



Discussion: In this example

- What's the performance?
- What's the task?
- What's the experience?

Discussion: How do we learn?

- Learning from …

- Learning by …

- What does it mean to have learned something?

Different tasks / problems in Machine Learning

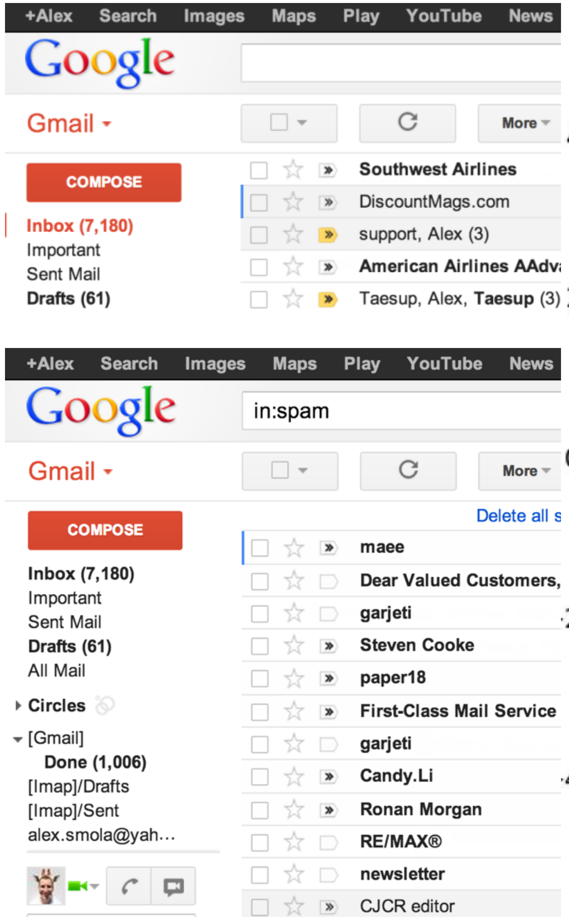
- Supervised Learning — Spam Filter.
- Unsupervised Learning — Topics of a text corpus
- Reinforcement Learning — Atari Games. Serve Ads.
- Structured Prediction — Machine translation.

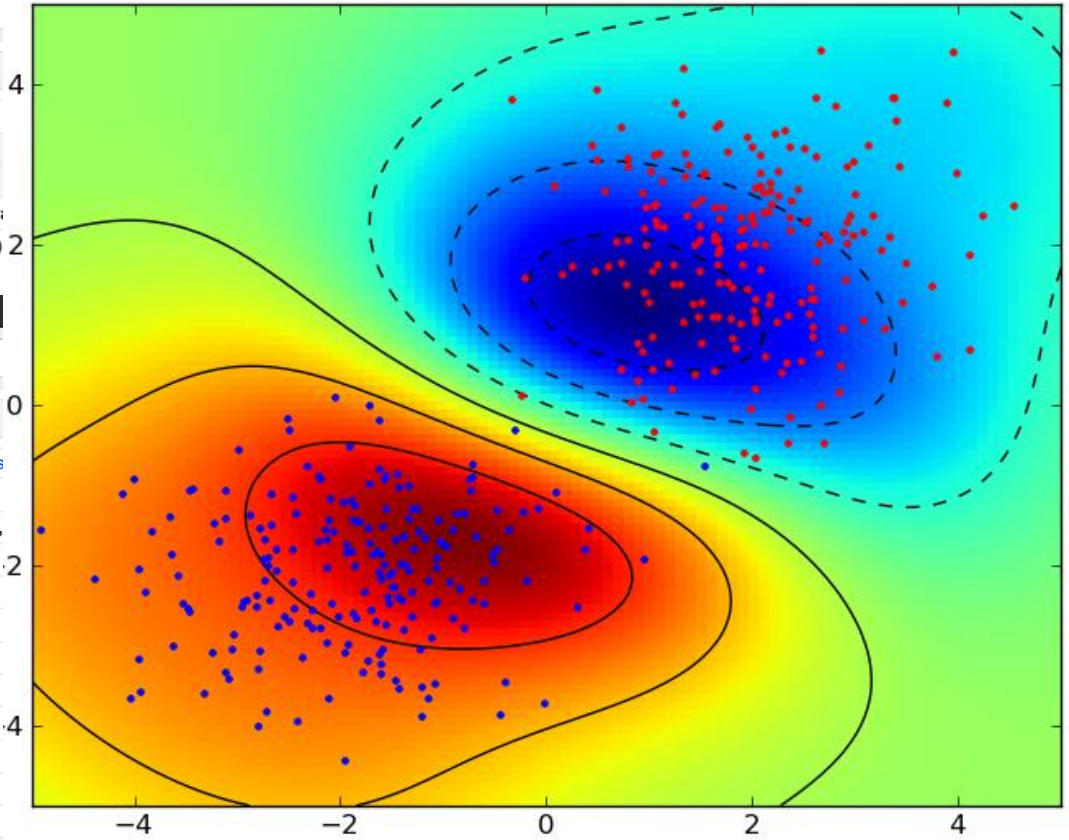
Semi-supervised learning, active learning,

ranking /search / recommendation

self-supervised learning and many more!

Supervised learning is about predicting label y using feature x by learning from labeled examples.





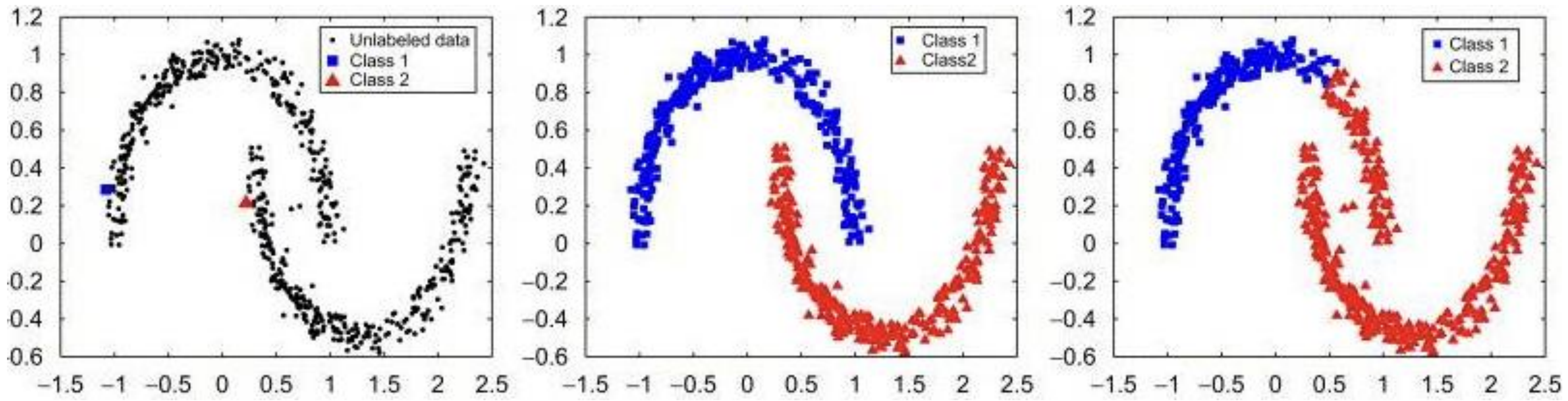
Unsupervised Learning is about finding structures in an unlabeled dataset.

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

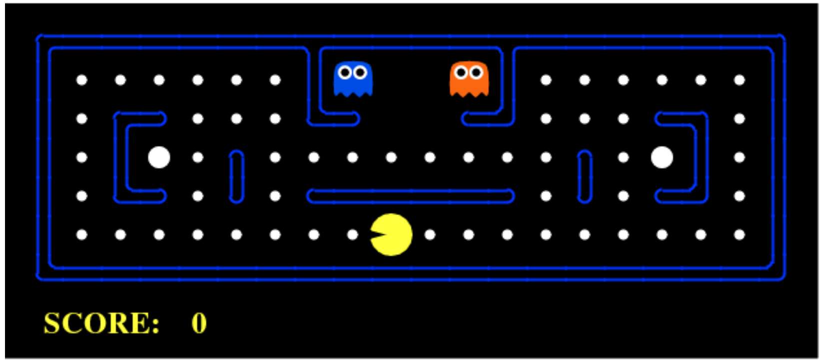
Example: Topic model. "Latent Dirichlet Allocation"

Semi-supervised Learning using both labeled and unlabeled data.



Zhu and Ghahramani (2022) “Learning from labeled and unlabeled data with label propagation”

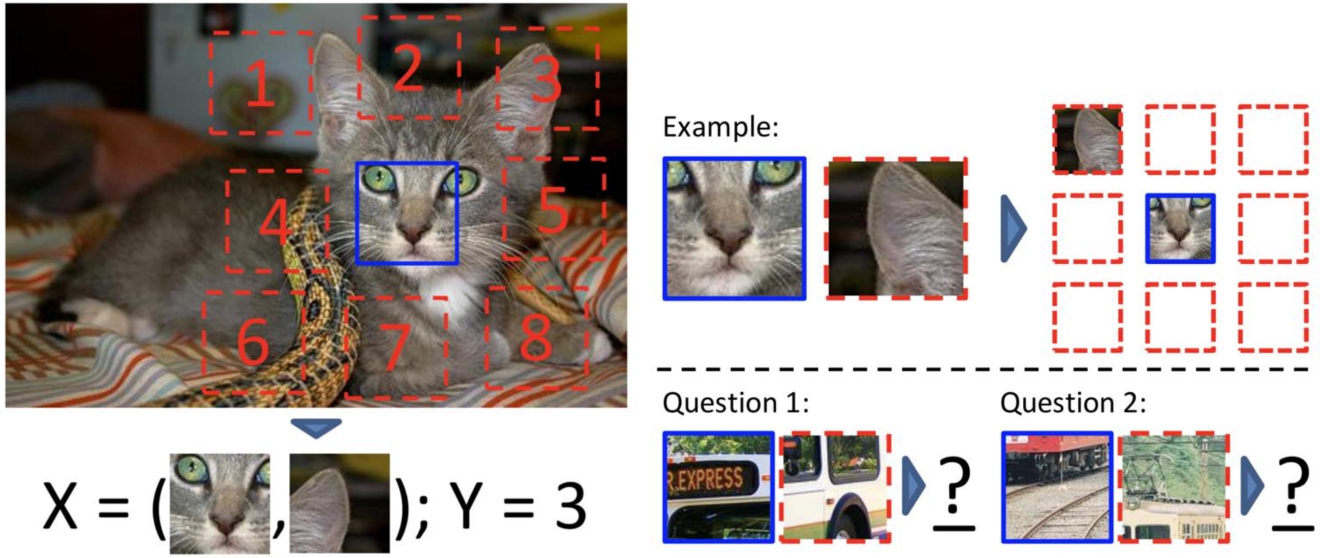
Reinforcement learning learns to make decisions for long-term rewards by trials-and-errors.



Recommendations

buy or not buy

Self-supervised learning learns to predict parts of x using other parts of x.



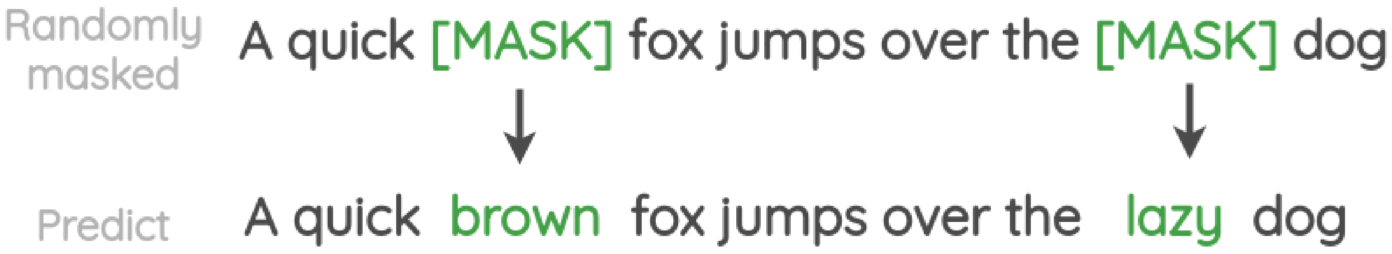


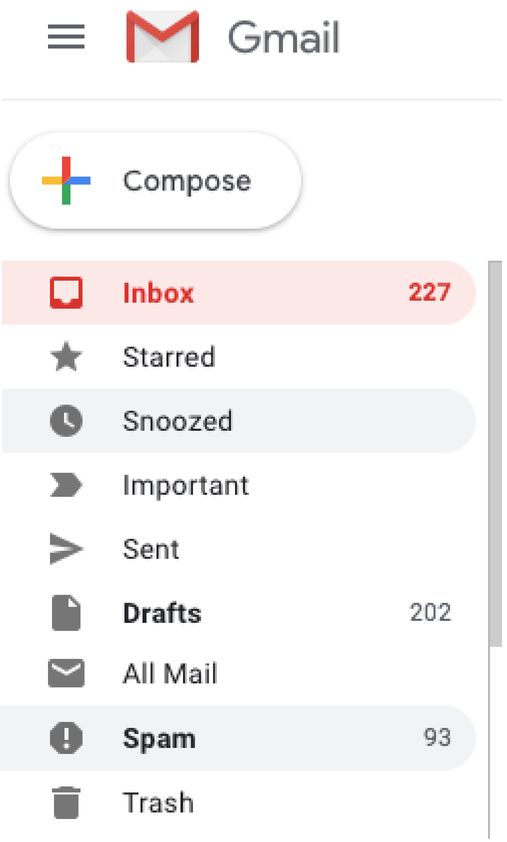
Image example from (Doersch et al, 2015) , text example from Amit Chaudhary

Summary of different ML problems

ML problem	Input	Output	What do we learn?	Applications	This course
Supervised learning	$[(x_1, y_1), \dots, (x_n, y_n)]$	$\hat{y}$, given new x	Mapping $g\colon\ X \to Y$	Price prediction	13 lectures
Unsupervised learning	$[x_1, \dots, x_m]$	Task dependent	Structural information of X	Biotech (dimension reduction)	3 lectures
Semi-supervised learning	$[(x_1, y_1), \dots, (x_n, y_n)]$ and $[x_1, \dots, x_m]$	$\hat{y}$, given new x	Mapping $g\colon\ X \to Y$	Large-scale ML	N/A
Reinforcement learning	An open environment where the learner can select x	A sequence of selected $[x_1, \dots, x_n]$ and their associated rewards $[y_1, \dots, y_n]$	Good policy to make decisions in selecting x	Materia/drug discovery	2 lectures
Self-supervised learning	An incomplete sequence x	A complete sequence $\hat{x}$	Good policy to fill the unknown part	Natural language (email auto filling)	N/A

The focus of today's lecture is "Supervised Learning"

- Actually, just "binary classification".
- Typical Example: Spam filtering
  - Design an "agent" to look at my email
  - And predict whether it is "Spam" or "Ham"



Example of SPAM emails

Mail thinks this message is Junk Mail. Move to Inbox

MICROWORLD CORPORATIO... December 20, 2019 at 2:38 AM MC

CLAIMS.

To: undisclosed-recipients:;,

Reply-To: microworld219@gmail.com

MICROWORLD CORPORATIONS:
CUSTOMER SERVICE:
FRIEDRICHSTRAΗE 10,BERLIN ALEMANHA
REFERENCE NUMBER: MBB-009-D54-DE
BATCH NUMBER: MGC-2019- SM-009
TICKET NUMBERS: 2,6,13,21,26,32

OFFICIAL WINNING NOTIFICATION.
We are pleased to inform you of the released results of Microworld Promotion...
This is a promotional program organized by Microworld Corporations, in conjunction with the Foundation for the promotion of software products, and use of email addresses. Held on Thursday 19th, December 2019, in Berlin, Alemanha.
Your email address won a cash award of Four hundred and eighty eight thousand two hundred and fifty euros (488,250.00 Euros)..
Contact Our Foreign Transfer Manager for claims with your winning details and your contact information.
Mrs. Helena Bosch.
Email: micropromo19@yahoo.com
Congratulations!!
Sincerely,
Rosa Van Beek.

Mail thinks this message is Junk Mail. Move to Inbox

☆ **MARK ZUCKERBERG** Junk - Google August 24, 2018 at 10:48 AM MZ

WINNING AMOUNT

Reply-To: MARK ZUCKERBERG

WINNING AMOUNT

My name is Mark Zuckerberg,A philanthropist the founder and CEO of the social-networking website Facebook,as well as one of the world's youngest billionaires and Chairman of the Mark Zuckerberg Charitable Foundation, One of the largest private foundations in the world. I believe strongly in'giving while living' I had one idea that never changed in my mind - that you should use your wealth to help people and i have decided to secretly give {$1,500,000.00} to randomly selected individuals worldwide. On receipt of this email, you should count yourself as the lucky individual. Your email address was chosen online while searching at random.Kindly get back to me at your earliest convenience,so I know your email address is valid.(mzuckerberg2444@gmail.com) Email me Visit the web page to know more about me: https://en.wikipedia.org/wiki/ Mark_Zuckerberg/ or you can google me (Mark Zuckerberg)

Regards,
MARK ZUCKERBERG

Example of a HAM (non-spam) email

Dear Professor Foo,

I am a student in your machine learning class.

I have a question about the second term project and I was not able to find the answer on the syllabus. Should our project be only about the topics listed on the second part of the syllabus, or can I incorporate topics from the whole course, as long as it fits with the subject of the class?

I look forward to hearing from you.
Best regards,
Bar

Discussion: What are the features that we can use to describe an email?

- What are characteristics of spam and ham emails?

- What are the information that we can extract from text, and hyper-texts to describe an email?

- What are typical characteristic of a spam email?

Possible features

- Number of special characters: $, %
- Mentioning of: Award, cash, free
- Greetings: generic, or specific
- Bad grammars and misspelled words: e.g. m0ney, c1ick here.
- Excessive excitement: Many “!”, “!!!”, “?!”, words in CAPITAL LETTERS.

- Whether the senders on the contact list
- Length of an email
- Whether the receiver has responded to sender before

Example of a feature vector of dimension 4

Contains hyperlinks	Whether on the contact list	Proportion of misspelled words	Length of the message
1	**0**	**0.0375**	**80**

Step 1 in Modelling
Feature extractor:
Converting the object of interest
to a vector of numerical values.

Mathematically defining a classifier

- Feature space: $\mathcal{X} = \mathbb{R}^d$
- Label space: $\mathcal{Y} = \{0, 1\} = \{\text{non-spam}, \text{spam}\}$
- A classifier (hypothesis): $h : \mathcal{X} \rightarrow \mathcal{Y}$

Math notation for "function definition", e.g., function "add"

How do you write in python?

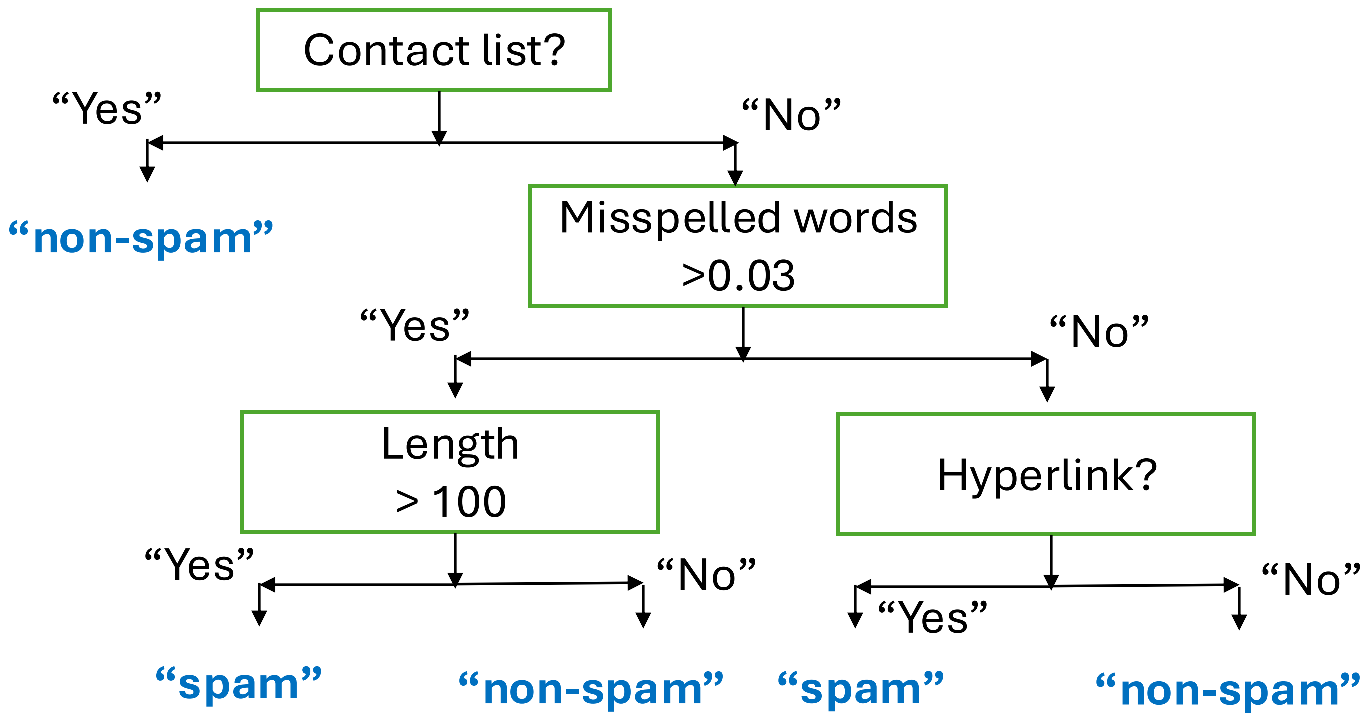
How do we make use of this feature vector? What is a reasonable “classifier” based on this feature representation?

Contains hyperlinks	Whether the contact list	Proportion of misspelled words	Length of the message
1	**0**	**0.0375**	**80**

- Feature space: $\{0, 1\} \times \{0, 1\} \times \mathbb{R} \times \mathbb{N}$
- Label space: $\mathcal{Y} = \{0, 1\} = \{\text{non-spam}, \text{spam}\}$

Discussion: How are we going to use these features as a human?

Decision trees



- **Question:** How is each decision tree determined? What are its parameters?

How is a decision tree specified?

- Parameters (built-in parameters of a model)
  - Which feature(s) to use when branching?
  - How to branch? Thresholding? Where to put the threshold?
  - Which label to assign at leaf nodes?
- Hyperparameters (parameters that you can set)
  - Max height of a decision tree?
  - Number of features the tree can use in each branch?

- **In-class exercise:** Consider a problem with **4 binary features**.
  - How many decision trees of **3 layers** are there, if each decision uses only one feature? (you may repeat features)
  - How many possible feature vectors are there?