INF367 25H: Selected Topics in Artificial Intelligence

Diamonds and Rust in the AI Treasure Chest

Dario Garigliotti - UiB - 29.09.2025

Plan for today

- Admin
  - Course plan
    - Lectures and additional lecture(s)
    - Assignments
- Recap from last lecture
- K nearest neighbours
  - Activity: density estimation

Admin

Some admin

Course plan

- Updated course plan
  - Ongoing topics
  - Additional lecture(s)
  - Assignments timeline

Some admin

Assignment 1

- Assignment 1
  - Released: today after lecture (on MittUiB, to be also announced)
  - Deadline: Friday October 17, 10:59 AM
  - Mandatory
  - Individual
  - Written delivery of an investigation report

Some admin

Assignment 1

- Assignment 1
  - Investigate one among 8 possible topics (“concepts”)
    - Each student receives a concept
      - not all students receive the same
  - Investigation reports will be used to build the presentation that I will give about these concepts in “Selected topics I and II” lectures in October

Some admin

Assignment 2

- Assignment 2
  - Released: Tuesday October 14 (on MittUiB, to be also announced)
  - Deadline: depends on when you present it
  - Mandatory
  - Team of k students, 2 <= k <= 2 (ja, team of 2)
  - Oral presentation of research article(s)
    - Every team studies and presents a different paper

Some admin

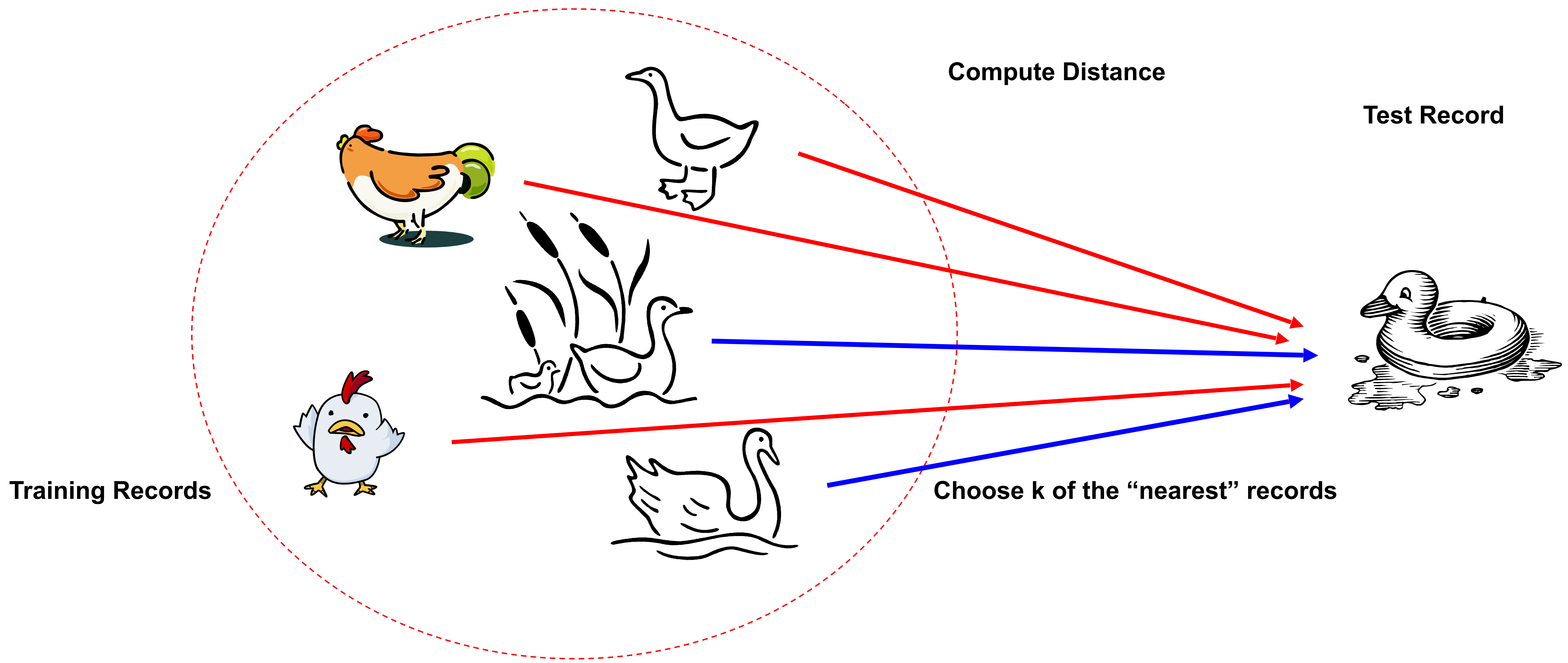
Assignment 2

- Assignment 2
  - Team of 2 students due to several reasons
  - Part 0: before it even begins: suggest your team
    - Deadline: Tuesday October 14, 13:59
      - Send a message on MittUiB Inbox/Outbox with the two names
      - Or communicate your situation on MittUiB Inbox/Outbox:
        - Withdrawn from course? -> Ok, no issues
        - Single without team partner? -> Ok, no issues: I will assign you a partner/team

k nearest neighbours

k nearest neighbours

- Basic idea:
  - "If it walks like a duck, quacks like a duck, then it's probably a duck"

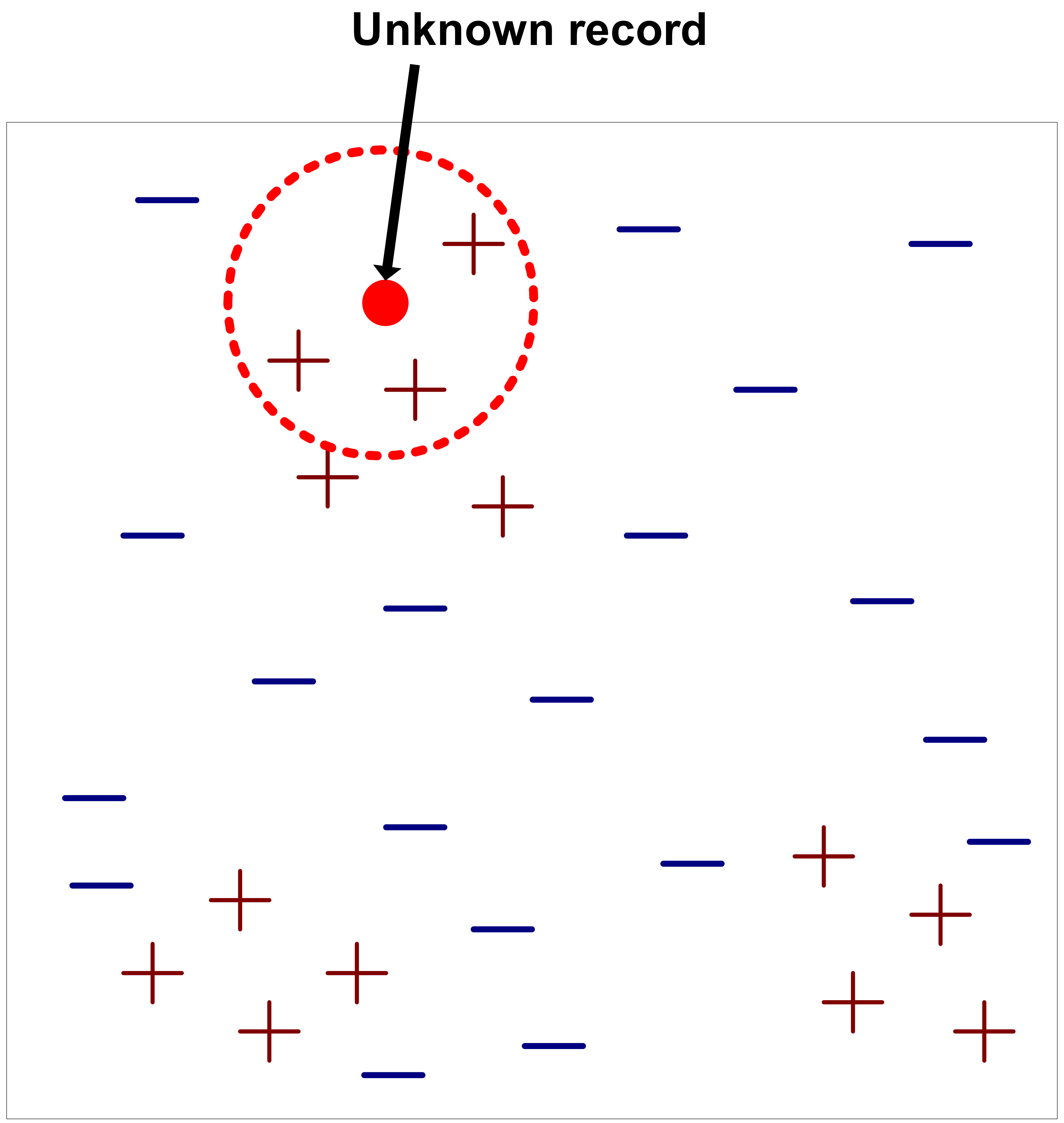


k nearest neighbours

- Requires three things
  - The set of stored instances
  - Distance Metric to compute distance between instances
  - The value of k, the number of nearest neighbors to check

k nearest neighbours

- To classify an unknown instance
  - Compute distance to other training instances
  - Identify k-nearest neighbours
  - Use class labels of nearest neighbours to determine the class label of unknown instance (e.g., by taking majority vote)



k nearest neighbours

Activity

- What are possible issues with this technique?
- Remember: it requires three things
  - Distance Metric to compute distance between instances
  - The value of k, the number of nearest neighbours to check
  - The set of stored instances

k nearest neighbours

- What are possible issues with this technique?
  - Distance Metric to compute distance between instances
    - Some metrics could better capture contributions per dimension
  - The value of k, the number of nearest neighbours to check
    - Too small? Too large?
  - The set of stored instances
    - Possibly costly in size, in time

Activity 2

Activity

k nearest neighbours? Do you mean density estimation?

- *N* data points drawn from *p(x)*, *K* of them are in a region *R* with volume *V*
- *m* (or tiny *k* subscript below) classes {*C_i* : i = 1,…, *m*}, *N_m* have class *C_m*
- *K_m* points among the *K* have class *C_m*
- **How would we formulate k-nn from this?**

$$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}$$

- <u>Hint:</u> density *p(x)* can be estimated like this:

$$p(\mathbf{x}) = \frac{K}{NV}.$$

Activity

k nearest neighbours? Do you mean density estimation?

- **Why does that estimate *p(x)*?** It's (a bit) technical:
  - We think on the probability *P* that a point is in *R*:

$$P = \int_{\mathcal{R}} p(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

  - The total K points in *R* distribute according to Binomial distribution Bin*(K|N,P)*
  - Calculating mean and var. of Bin*(K|N,P)*, and assuming a couple of properties
    - R sufficiently large w.r.t. *p(x)*, hence Bin*(K|N,P)* sharply peaked around mean
    - R sufficiently small, hence *p(x)* is approx. constant
  - We get *K ~approx= N*P*, and *P ~approx= p(x)*V*, hence *p(x) = K / N*V*

Activity

k nearest neighbours? Do you mean density estimation?

- **How do we arrive to k-nn** from optimal Bayes? $$p(\mathcal{C}_k|\mathbf{x}) = \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{p(\mathbf{x})}$$
  - From the previous result, *p(x) = K / N*V*,
  - We replace each of the parts on the right-side of Bayes expression using
    - *p(x|C_m) = K_m / N_m*V*
    - *p(C_m) = N_m / N*
    - And we obtain our proportion used in k-nn: *p(C_m|x) = K_m / K*
    - k nearest neighbours is just finding the class *m* that maximizes that. OK.

Activity

Density estimation

- The results from this activity present k-nn in a different way:
  - As a particular case of a more general technique, density estimation
    - We want to use the true, yet ignored, density distribution $p(x)$
    - From this density estimate $p(x) = K / N*V$, …
    - *…we* decided to fix K (the K of K-nn) and find an appropriate (region of volume) V around it by the nearest points to the x of interest

Activity

Density estimation

- It also shows that the previous developments about probability come handy:
  - The whole approach is **probabilistic** in the sense of that region of interest where K points out of the whole dataset are in
  - The process of finding a good, i.e. larger, K, $1 <= K < N$ is a **smoothing** of the boundaries for robustness
  - Just similarly to what was done for naive Bayes, also here we take a lot of **assumptions**, e.g. to estimate p(x) and then find k-nn via optimal Bayes expression