

Quiz 1 Review

DS701 Session 9 — Mon Oct 5, 2026 · quiz is Wed Oct 7

Quiz 1 — Wednesday

Logistics

- **Wed Oct 7, in class.** targeted 60 minutes but you can use entire 105 minutes of class, written for ~40 minutes of work. You're dismissed when you hand it in.
- **Paper and pencil.** Closed notes, no devices.
- You get a **printed copy of the reference sheet** — the same one posted last Wednesday: [quiz1-reference.pdf](#). It is frozen; what you see is what you'll have.
- Read it **before** Wednesday. Knowing *where* a formula is on the sheet is worth more than remembering it.
- Bring a pencil and your BU ID. No calculator needed — numbers are chosen to work by hand, and you may leave answers as $\sqrt{7}$, e^{-2} , fractions.

What's in scope — and what isn't

Lectures 01–07

- 01 What is Data Science
- 02 Distances and similarity
- 03 Clustering I: k -means
- 04 Clustering II: hierarchical
- 05 Probabilistic modeling, MLE, CI
- 06 Clustering III: GMM and EM
- 07 Generalization

Not on Quiz 1

- Today's lecture (trees and forests)
- Anything in a module (M1–M13)
- Python syntax or `sklearn` calls
- Derivations from the appendices (05A, 06A–06E)

Calibration: the problems are at the level of **HW1–HW4** plus the bias–variance problem coming in HW5. Fresh scenarios, fresh numbers.

Shape of the quiz

Five problems, 20 points each:

1. **Mechanical** — compute distances/similarities and interpret the result
2. **Mechanical** — run *one step* of an algorithm by hand (k -means, agglomerative merge, or EM)
3. **Conceptual** — explain or evaluate claims; short derivation (MLE)
4. **Conceptual** — reason about generalization, validation, hard vs soft assignment
5. **Transfer** — a new scenario: pick the method, justify it, state its assumptions and what could go wrong

Partial credit is real: a correct setup with an arithmetic slip earns most of the points. A bare number with no work earns few.

Quick-fire check

Distances (L02)

Q1. Why is $1 - \cos(\mathbf{x}, \mathbf{y})$ **not** a metric? Which axiom fails?

→ Triangle inequality — and $d(\mathbf{x}, 2\mathbf{x}) = 0$, so “ $d = 0 \iff$ equal” fails too.

Q2. As $p \rightarrow \infty$, which coordinate dominates $\|\mathbf{x} - \mathbf{y}\|_p$?

→ The single largest $|x_i - y_i|$ (Chebyshev).

Q3. When would you prefer Jaccard over Hamming for two bit vectors?

→ When the sets differ a lot in size — Jaccard normalizes by the union.

Q4. One feature is in dollars, another in years. What do you do before computing Euclidean distance, and why?

→ Standardize; otherwise the large-range feature dominates.

Clustering (L03, L04, L06)

Q5. Across Lloyd iterations, can the k -means objective ever go **up**?

→ No. Each step (assign, update) can only lower WCSS — but it stops at a *local* minimum, hence k -means++ and multiple restarts.

Q6. Single vs complete linkage: which one chains, and which prefers compact clusters?

→ Single chains (min distance); complete prefers compact (max distance).

Q7. In the E-step, what does $\sum_k \gamma_{ik}$ equal, and why?

→ 1 — responsibilities are posterior probabilities $P(C_k \mid x_i)$ via Bayes' rule.

Q8. Name the one thing EM guarantees and the one thing it doesn't.

→ Log-likelihood never decreases; no guarantee of the *global* maximum.

Probability and generalization (L05, L07)

Q9. The MLE of σ^2 divides by N or $N - 1$?

→ N . The sample variance s^2 divides by $N - 1$. Know which one a problem asks for.

Q10. Doubling n does what to the width of a 95% CI for the mean?

→ Shrinks it by $1/\sqrt{2}$ — width $\propto 1/\sqrt{n}$.

Q11. k in a degree- k polynomial fit: parameter or hyperparameter? Who sets it?

→ Hyperparameter; chosen by cross-validation, **never** by training error.

Q12. Training error falls to zero as complexity grows. Is that bias or variance you're seeing?

→ Variance: a different training sample would give a very different fit.

How to prepare

Between now and Wednesday

- **Work the sheet, don't read it.** For each algorithm box (Lloyd's, agglomerative, EM), run one iteration by hand on 5–6 made-up numbers. That is exactly what the mechanical problems ask.
- **Redo one problem from each of HW1–HW4** without looking at your solution. The quiz is calibrated to them.
- **Re-answer the knowledge-check questions** for lectures 01–07 from memory, then check the recap decks.
- For the transfer problem, practice the **techniques map** on the sheet: given a scenario, which question type is it (Leek & Peng), which method, what assumption would break it?
- Office hours and Piazza are open through Tuesday — ask early.

On the day

- Skim all five problems first; start with the one you're surest of.
- **Show work.** Write the formula from the sheet, substitute, then compute.
- State assumptions you make; label units; box your final answers.
- If a number looks wrong, say so and keep going — flagging it earns credit.
- Stuck on arithmetic? Leave it symbolic and move on.

 After the quiz: the **team midterm project** is released Wednesday and HW5 goes out Thursday. Enjoy the rest of today's activity — trees are *not* on the quiz.