# Z3003: Weeks 1–6 Part 1 Revision & Mock Quiz with Solutions

Data Mining: Scaling to Large Data (BS Core) · Dr. Innocent Nyalala · innocent@iitmz.ac.in

**Part 1 Scope:** Covers Dr. Nyalala's Weeks 1–6 curriculum: Bayesian ML, MLE, Feature Selection, PCA/LDA, Imbalance & Ensembles, OLAP/MapReduce/PySpark, MinHash/LSH, Apriori, and Bloom Filters. Solutions appear on Page 2.

## Weeks 1–6 Concept Summary & Mock Quiz [60 Marks]

**Q1.** [**Bayesian Risk, 10 marks**] A crop disease screen has prior $P(\text{Infected}) = 0.10$. False negative cost is $\lambda_{01} = 15$ and false positive cost is $\lambda_{10} = 1$. Derive the optimal posterior threshold $P(\text{Infected} \mid x)$ above which the classifier must flag infection.

**Q2.** [**Information Gain, 10 marks**] A node of 8 farms has 4 High-yield and 4 Low-yield samples. Splitting on `Fertilizer=Yes` sends (3 High, 1 Low) left and (1 High, 3 Low) right. Compute parent entropy $H(Y)$ and exact Information Gain $IG(Y \mid \text{Fertilizer})$.

**Q3.** [**PCA vs. LDA & Leakage, 10 marks**] State the mathematical difference between PCA and Fisher LDA, and explain why running `StandardScaler` before `train_test_split` invalidates held-out metrics.

**Q4.** [**AdaBoost Update, 10 marks**] Weak classifier $h_1$ achieves weighted error $\epsilon_1 = 0.25$ on 4 equally weighted points ($w_i = 0.25$). Compute $\alpha_1$ and the normalized weight assigned to the single misclassified point for round 2.

**Q5.** [**LSH S-Curve, 10 marks**] A 120-row MinHash signature is split into $b = 30$ bands of $r = 4$ rows. Compute the exact probability that two documents with Jaccard similarity $s = 0.6$ become a candidate pair, and state the approximate threshold $t \approx (1/b)^{1/r}$.

**Q6.** [**Apriori & Bloom Filters, 10 marks**] In 1,000 agro-dealer baskets, `hoe` appears in 600, `seed` in 200, and both in 110. Compute Support, Confidence, and Lift for seed $\rightarrow$ hoe. Next, for a Bloom filter with $m = 1{,}000$ bits and $n = 100$ items, compute the optimal hash count $k^*$.

## Worked Step-by-Step Mock Quiz Solutions

**Q1. Bayesian Decision Rule:** Let $p = P(\text{Infected} \mid x)$. Expected risk of predicting Infected ($R_1$) vs. Healthy ($R_0$) is:

$$R_1 = \lambda_{10}(1-p) = 1 \cdot (1-p), \qquad R_0 = \lambda_{01} p = 15p.$$

Flag Infected when $R_1 < R_0 \iff 1 - p < 15p \iff 16p > 1 \iff \mathbf{p > 1/16 = 0.0625}$. (Notice the threshold drops below the prior 0.10 because missing a diseased plot is 15× costlier than a false alarm.)

**Q2. Entropy & Information Gain:** Parent has $(4,4)$, so $H(Y) = -0.5\log_2(0.5) - 0.5\log_2(0.5) = \mathbf{1.0}$ bit. Each child has $(3,1)$ or $(1,3)$: $H(\text{child}) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 2 - \frac{3}{4}\log_2 3 \approx 2 - 1.1887 = \mathbf{0.8113}$ bits. Weighted child entropy $= \frac{4}{8}(0.8113) + \frac{4}{8}(0.8113) = 0.8113$. Thus $IG = 1.0 - 0.8113 = \mathbf{0.1887}$ bits.

**Q3. PCA vs. LDA & Leakage:** PCA maximizes total unlabeled variance $\mathbf{w}^T\Sigma\mathbf{w}$ via top eigenvectors of $\Sigma$, whereas Fisher LDA uses class labels to maximize between-class separation over within-class scatter $\frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}$ via $S_W^{-1} S_B$. Scaling before splitting leaks test-set mean and standard deviation into training features; fit scalers, imputers, and PCA solely on training folds inside a `Pipeline`.

**Q4. AdaBoost Update:** $\alpha_1 = \frac{1}{2}\ln\frac{1-\epsilon_1}{\epsilon_1} = \frac{1}{2}\ln\frac{0.75}{0.25} = \frac{1}{2}\ln 3 \approx \mathbf{0.5493}$. Before normalization, misclassified weight multiplies by $e^{+\alpha_1} = \sqrt{3}$ $(0.25\sqrt{3})$, whereas each of the 3 correct points multiplies by $e^{-\alpha_1} = 1/\sqrt{3}$ $(0.25/\sqrt{3})$. Normalization sum $Z_1 = 0.25(\sqrt{3} + 3/\sqrt{3}) = 0.25(2\sqrt{3}) = \frac{\sqrt{3}}{2}$. Thus the misclassified point's normalized weight is $\frac{0.25\sqrt{3}}{0.25\cdot 2\sqrt{3}} = \mathbf{0.50}$ (matching $\frac{1}{2}$ total error mass).

**Q5. MinHash LSH Probability:** Probability all $r = 4$ rows match in one band is $s^r = 0.6^4 = 0.1296$. Probability of colliding in at least one of $b = 30$ bands is:

$$\Pr[\text{candidate}] = 1 - (1 - 0.1296)^{30} = 1 - (0.8704)^{30} \approx 1 - 0.0155 = \mathbf{0.9845}\ (\mathbf{98.45}\%).$$

Approximate S-curve inflection threshold is $t \approx (1/30)^{1/4} \approx \mathbf{0.427}$.

**Q6. Apriori Interest & Bloom Filter:** $\text{supp}(\text{seed} \cup \text{hoe}) = 110/1000 = \mathbf{0.11}$; $\text{conf}(\text{seed} \to \text{hoe}) = 110/200 = \mathbf{0.55}$. Expected random co-occurrence is $P(\text{hoe}) = 0.60$, so Lift $= \frac{0.55}{0.60} = \mathbf{0.917 < 1}$ (negative correlation; `hoe` is a decoy item!). For the Bloom filter ($m = 1000, n = 100$), optimal hash count is $k^* = \frac{m}{n}\ln 2 = 10 \times 0.6931 \approx \mathbf{7}$ hash functions.

**Scaling Bridge to Part 2:** When dataset size grows 1,000× (e.g., 500,000 farms), replace exact in-memory PCA with randomized/incremental PCA, run distributed aggregations via PySpark `reduceByKey` (map-side combining), and apply LSH or Bloom filters to bound memory to $O(m)$ bits.