Z3003: Data Mining at Scale Formula Sheet

Data Mining: Scaling to Large Data (Weeks 1–15) · Co-taught with Dr. Mahendran V · innocent@iitmz.ac.in

1. Bayesian Decision Theory & MLE

- **Posterior:** $P(\omega_j \mid \mathbf{x}) = \dfrac{p(\mathbf{x} \mid \omega_j)P(\omega_j)}{\sum_k p(\mathbf{x} \mid \omega_k)P(\omega_k)}$.
- **Expected Risk (Asymmetric Cost λ_{ij}):** Choose α_i minimizing $R(\alpha_i \mid \mathbf{x}) = \sum_j \lambda(\alpha_i \mid \omega_j)P(\omega_j \mid \mathbf{x})$.
- **Gaussian MLE:** $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n x_i$, $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (x_i - \hat{\mu})^2$.
- **Bernoulli / Naive Bayes:** $\hat{p} = \frac{k+\alpha}{n+\alpha|V|}$ (Laplace smoothing $\alpha = 1$).

2. Entropy, Information Gain, PCA & LDA

- **Shannon Entropy:** $H(Y) = -\sum_y p(y)\log_2 p(y)$.
- **Information Gain:** $IG(Y \mid X) = H(Y) - \sum_x p(x)\, H(Y \mid X = x)$.
- **Gini Impurity:** $G = 1 - \sum_k p_k^2$.
- **PCA:** Center $\tilde{X} = X - \bar{x}$; top-k eigenvectors of covariance $\Sigma = \frac{1}{n}\tilde{X}^T\tilde{X}$ maximize projected variance $\mathbf{w}^T\Sigma\mathbf{w}$.
- **Fisher LDA:** Maximize $J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}}$ via generalized eigenvectors of $S_W^{-1}S_B$.

3. Class Imbalance & Ensemble Learning

- **Metrics:** $P = \frac{TP}{TP+FP}$, $R = \frac{TP}{TP+FN}$, $F_1 = \frac{2PR}{P+R}$, MCC $= \frac{TP\cdot TN - FP\cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$.
- **SMOTE:** Synthetic sample $\mathbf{x}_{\text{new}} = \mathbf{x}_i + \lambda(\mathbf{x}_{nn} - \mathbf{x}_i)$, $\lambda \sim U(0,1)$.
- **AdaBoost:** Sample error $\epsilon_t = \sum_{y_i \neq h_t(x_i)} w_i$; learner weight $\alpha_t = \frac{1}{2}\ln\frac{1-\epsilon_t}{\epsilon_t}$; update $w_i \leftarrow w_i e^{-\alpha_t y_i h_t(x_i)}/Z_t$.
- **Gradient Boosting:** Fit $h_m(x)$ to negative gradient $r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F=F_{m-1}}$ (equals $y_i - \hat{y}_i$ for squared error).

4. MapReduce, MinHash & LSH

- **Jaccard Similarity:** $J(A,B) = \frac{|A\cap B|}{|A\cup B|}$; MinHash property $\Pr[h_{\min}(A) = h_{\min}(B)] = J(A,B)$.
- **LSH S-Curve (b bands, r rows, $n = br$):** Pr[candidate pair] $= 1 - (1 - s^r)^b$, threshold $t \approx (1/b)^{1/r}$.
- **PySpark Transformations (Lazy):** `map`, `filter`, `flatMap`, `reduceByKey` (map-side combine), `join`. **Actions:** `collect`, `count`, `take`.

5. Frequent Itemsets & Bloom Filters

- **Support & Confidence:** $\text{supp}(X) = \frac{|T_X|}{N}$, $\text{conf}(X \to Y) = \frac{\text{supp}(X\cup Y)}{\text{supp}(X)}$.
- **Lift:** $\frac{P(X\cup Y)}{P(X)P(Y)}$. **Null-Invariant Kulczynski:** $\frac{1}{2}(P(X|Y) + P(Y|X))$.
- **Imbalance Ratio:** $\text{IR} = \frac{|\text{supp}(X) - \text{supp}(Y)|}{\text{supp}(X)+\text{supp}(Y)-\text{supp}(X\cup Y)}$.
- **Apriori Principle:** Every subset of a frequent itemset is frequent (anti-monotonicity).
- **Bloom Filter (m bits, k hashes, n items):** False positive rate $p \approx \left(1 - e^{-kn/m}\right)^k$; optimal $k^* = \frac{m}{n}\ln 2 \approx 0.693\frac{m}{n}$; bits per item $m/n \approx 1.44\log_2(1/p)$.

6. Scalable Clustering (Part 1 Bridge & Part 2)

- **k-Means Objective (SSE):** $\sum_{j=1}^k \sum_{x\in C_j} \|x - \mu_j\|_2^2$.
- **Silhouette Score:** $s(i) = \frac{b(i)-a(i)}{\max(a(i),b(i))} \in [-1,1]$.
- **BIRCH CF-Tuple:** CF $= (N, \overrightarrow{LS}, SS)$ where $\overrightarrow{LS} = \sum x_i$, $SS = \sum \|x_i\|^2$ (additive across merged subclusters).
- **Spectral Clustering:** Unnormalized Laplacian $L = D - W$; bottom-k eigenvectors of L (or $L_{\text{sym}} = D^{-1/2}LD^{-1/2}$) feed k-means.

7. Recommender Systems & Matrix Factorization

- **Cosine / Centered Cosine (Pearson):** $\cos(\mathbf{u},\mathbf{v}) = \frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$.
- **Baseline + Latent Factor SVD:** $\hat{r}_{ui} = \mu + b_u + b_i + \mathbf{q}_i^T\mathbf{p}_u$.
- **Regularized SGD Objective:** $\min \sum_{(u,i)\in\mathcal{K}} (r_{ui} - \hat{r}_{ui})^2 + \lambda(\|\mathbf{p}_u\|^2 + \|\mathbf{q}_i\|^2 + b_u^2 + b_i^2)$.
- **RMSE:** $\sqrt{\frac{1}{|\mathcal{T}|}\sum_{(u,i)\in\mathcal{T}} (r_{ui} - \hat{r}_{ui})^2}$.

8. Web Graph & Link Analysis

- **PageRank with Teleportation ($\beta \approx 0.85$):** $\mathbf{r}^{(t+1)} = \beta M\mathbf{r}^{(t)} + \frac{1-\beta}{N}\mathbf{1}$, where $M_{ji} = 1/\text{outdeg}(i)$ if $i \to j$ (handle spider traps & dead ends via teleportation / column renormalization).
- **HITS (Hubs h, Authorities a):** $\mathbf{a} \leftarrow L^T\mathbf{h}$, $\mathbf{h} \leftarrow L\mathbf{a}$, normalize $\|\mathbf{a}\|_2 = \|\mathbf{h}\|_2 = 1$ at each iteration (converges to principal eigenvectors of L^TL and LL^T).

9. Streaming Sketches (FM, AMS, DGIM)

- **Flajolet-Martin (Distinct Elements F_0):** Track maximum trailing zeros R of $h(x)$; estimate $\hat{F}_0 = 2^R$ (or median of group averages of 2^{R_j} to reduce variance).
- **AMS Second Moment ($F_2 = \sum_a m_a^2$):** Variable $X = n(2c-1)$ for count c from random position; unbiased $\mathbb{E}[X] = F_2$.
- **DGIM Sliding-Window Bit Counting:** Store buckets of sizes $1, 2, 4, 8, \ldots$ with at most 2 buckets per power of 2; space $O(\log^2 N)$ bits for window N, max relative error $\leq 50\%$.
- **Reservoir Sampling (k items):** Keep item $t > k$ with probability k/t, replacing a uniformly chosen slot in $[0, k-1]$.

10. Worked Exam Templates & Scale Rules

- **LSH Tuning Rule:** Given Jaccard threshold t and signature length $n = br$, solve $b \approx t^{-r}$. Increasing r lowers false positives; increasing b lowers false negatives.

- **Bloom Bit-Budget:** For $n = 10^6$ keys and target $p = 1\%$, $m/n \approx 1.44 \log_2(100) \approx 9.58$ bits/key (≈ 1.2 MB total) using $k^* \approx 7$ hashes.
- **MapReduce Partition Rule:** Use combiners only for associative and commutative aggregations (sum, count, min, max); compute exact averages via ‘(sum, count)‘ pairs.
- **SVD Bias Sanity Check:** Always fit global mean μ, user bias b_u, item bias b_i, and factors $\mathbf{p}_u, \mathbf{q}_i$ strictly on training ratings to avoid test leakage.