UT-ECE Data Science Final Assignment
Complete Solution Manual

Teaching Assistant Team

Spring 2025

Q1. Advanced Data Engineering & SQL

Q1A. Window-function solution

```
WITH citation_velocity AS (
    SELECT
        UserID,
        Country_Origin,
        Year,
        Research_Citations,
        AVG(Research_Citations) OVER (
            PARTITION BY Country_Origin
            ORDER BY Year
            ROWS BETWEEN 2 PRECEDING AND CURRENT ROW
        ) AS moving_avg_citations
    FROM Professionals_Data
)
SELECT
    UserID,
    Country_Origin,
    Year,
    Research_Citations,
    moving_avg_citations,
    DENSE_RANK() OVER (
        PARTITION BY Country_Origin
        ORDER BY moving_avg_citations DESC
    ) AS country_rank
FROM citation_velocity;
```

Q1B. Leakage diagnosis

Direct leakage: `Visa_Approval_Date` (post-outcome variable).

Potential temporal leakage: `Last_Login_Region` and `Passport_Renewal_Status`, if logged after migration decision.

Usually safe: `Years_Since_Degree`, provided degree date is known before inference.

Q2. Statistical Inference & Linear Models

Q2A. Elastic Net gradient

Given

$$J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2 + \lambda_1 \sum_{j=1}^{n} |\theta_j| + \frac{\lambda_2}{2} \sum_{j=1}^{n} \theta_j^2,$$

for coordinate θ_j:

$$\nabla_{\theta_j} J(\theta) = \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_j^{(i)} + \lambda_1 \, \partial|\theta_j| + \lambda_2 \theta_j.$$

Subgradient of absolute value:

$$\partial|\theta_j| = \begin{cases} +1 & \theta_j > 0 \\ -1 & \theta_j < 0 \\ [-1, 1] & \theta_j = 0 \end{cases}$$

Thus coefficients may remain exactly zero under coordinate-descent optimization.

Q2B. Interpretation

With coefficient 0.52, p-value 0.003, and 95% CI $[0.18, 0.86]$:

- p-value $< 0.05 \Rightarrow$ reject $H_0 : \beta = 0$.
- CI excludes zero, confirming statistical significance.
- Entire CI is positive, indicating a positive association with migration propensity.

Q3. Optimization & Gradient Descent

Ravine geometry: one direction has high curvature and another low curvature. Vanilla SGD oscillates in steep direction and advances slowly.

Momentum update:

$$v_t = \beta v_{t-1} + \eta \nabla J(\theta_t), \quad \theta_{t+1} = \theta_t - v_t.$$

Opposing gradients across steep walls cancel over iterations, while consistent shallow-direction gradients accumulate velocity.

Adam:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t, \quad s_t = \beta_2 s_{t-1} + (1 - \beta_2) g_t^2,$$

with bias correction and coordinate-wise scaling. Adam typically handles anisotropic curvature and mixed feature scales better.

Q4. Non-Linear Models & Kernels

Q4A. RBF overfitting control

Kernel is $K(x, x') = \exp(-\gamma \|x - x'\|^2)$.

If overfitting occurs, **decrease** γ. Larger γ means narrow influence radius and highly wiggly boundaries; smaller γ broadens influence and smooths decision boundary.

Q4B. Cost-complexity pruning

$$R_\alpha(T) = R(T) + \alpha|T|.$$

α penalizes leaf count:

- small α: larger trees (low bias, high variance)
- large α: smaller trees (higher bias, lower variance)

Optimal α is selected by validation/CV.

Q5. Unsupervised Learning

Q5A. PCA explained variance

Given eigenvalues $\lambda_1, \lambda_2, \lambda_3$ of covariance matrix Σ:

$$\mathrm{EVR}(PC_k) = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \lambda_3}.$$

Eigenvalue λ_k equals variance captured along principal component k.

Q5B. Elbow argument for K-Means

WCSS decreases monotonically with K because each added centroid can only reduce minimum point-to-centroid squared distance.

The elbow is the approximate point of maximum curvature where marginal gain

$$\Delta_K = \mathrm{WCSS}(K-1) - \mathrm{WCSS}(K)$$

starts diminishing substantially, giving a practical complexity-vs-fit compromise.

Q6. Capstone Explainability

For SHAP local explanation:

- `base_value`: expected model output over reference/background data.
- `output_value`: model output for a specific candidate.

Their difference is the sum of per-feature SHAP contributions for that candidate.

If a high-citation candidate is predicted `No Migration`, positive citation contribution can be outweighed by stronger negative contributions from other features (e.g., region/policy interaction, experience profile).