UNIVERSITY COMPUTER SCIENCE • EXERCISE SHEET

13.4 Regression and classification

Name: _____ Class: _____ Date: _____ **Total: 24 marks**

KEY WORDS linear regression • decision boundary

Objective

Fit by least squares, read a boundary, and price complexity in held-out error.

1 Key knowledge and worked examples

■ **The fit**

Course points $(1,2),(2,3),(3,5),(4,4),(5,6)$: the line $\hat{y} = 0.9x + 1.3$ with SSE 1.9 ✓.

■ **Worked example: the overfit**

The 4th-degree polynomial scores SSE 0 on those five points but ≈ 42.7 on the held-out five; the line scores 0.55 ✓.

2 Practice

2.1 What does the logistic boundary $2x_1 - x_2 - 1 = 0$ predict at $(2,2)$? [3]

2.2 Where does the margin's doubt live? [3]

2.3 Training error or held-out error – which is the product? [3]

3 Exam-style questions

3.1 Fit the least-squares line to $(1,3),(2,3),(3,4),(4,6),(5,6)$.

Compute $\bar{x}$, $\bar{y}$, the slope numerator and denominator, and w, b. Give the residuals and the SSE. Predict at $x = 6$ and state one reason the prediction could mislead. [8]

3.2 A logistic model has $w_1 = 1, w_2 = 2, b = -4$ over features (distance in km, traffic index).

Write the boundary equation and solve it for the traffic axis. Classify the points $(2,1)$ and $(2,1.5)$ with their scores. State which point the model is more confident about and why. Explain the design consequence of an unscaled traffic index on the fit. [7]

Solutions

2.1

- Score $2(2) - 2 - 1 = 1 > 0$: predicted late ✓.

2.2

- Points nearest the boundary, where scores are small ✓.

2.3

- Held-out error: training error is an advertisement ✓.

3.1

- $\bar{x} = 3$, $\bar{y} = 4.4$ ✓.
- Deviations from the means give numerator 9.0 and denominator 10 ✓.
- Slope $w = 9/10 = 0.9$; intercept $b = 1.7$ ✓.
- Residuals $0.4, -0.5, -0.4, 0.7, -0.2$; SSE 1.1 ✓.
- Prediction at 6: $0.9 \times 6 + 1.7 = 7.1$ – an extrapolation beyond the data range ✓.

3.2

- Boundary: $x_1 + 2x_2 - 4 = 0$, i.e. $x_2 = (4 - x_1)/2$ ✓.
- $(2, 1)$: score $2 + 2 - 4 = 0$ – exactly on the boundary, maximally doubtful ✓.
- $(2, 1.5)$: score $2 + 3 - 4 = +1 > 0$: positive class, with margin 1 ✓.
- More confident at $(2, 1.5)$: its score magnitude is larger – distance from the boundary ✓.
- Unscaled index: its weight absorbs raw magnitude, the loss surface ravines (unit 13.5) – standardise before fitting ✓.

exercises.lol exercises.com.cn