KHAI PHÁ DỮ LIỆU – CHƯƠNG 3
HỒI QUY DỮ LIỆU (DATA REGRESSION)

Tổng hợp kiến thức, câu hỏi tự luận và trắc nghiệm

1. GIỚI THIỆU VỀ HỒI QUY

1.1. Định nghĩa Hồi Quy

- **J. Han et al.:** Hồi quy là cơ chế thống kê cho phép dự đoán các giá trị số thực và liên tục.
- **R.D. Snee (1977):** Hồi quy là cơ chế thống kê cho phép dự đoán, kiểm soát và học các quy tắc sinh ra dữ liệu từ thực nghiệm.
- **Wiki (2009):** Phân tích hồi quy là cơ chế thống kê ước lượng mối quan hệ giữa các biến độc lập.

Phân biệt quan trọng:

- **Hồi quy (Regression):** Dự đoán giá trị số thực (real-valued output) – ví dụ: giá nhà, giá cổ phiếu.
- **Phân lớp (Classification):** "Dự đoán" cho giá trị rời rạc (discrete values) – ví dụ: spam/not spam.

1.2. Mô hình hồi quy

Phương trình hồi quy tổng quát:

$$Y = f(X, \theta)$$

- X: tập biến độc lập/dự báo (predictors/independent variables) – mô tả các thay đổi của Y.
- Y: biến đáp ứng/phụ thuộc (response/dependent variable) – mô tả sự kiện quan tâm.
- θ: hệ số hồi quy (regression coefficients) – mô tả mức độ ảnh hưởng tương đối của X lên Y.

1.3. Phân loại mô hình hồi quy

- **Tuyến tính (Linear) vs. Phi tuyến (Non-linear):** Linear: quan hệ tuyến tính giữa các tham số ảnh hưởng lên Y; Non-linear: ngược lại.
- **Đơn biến (Univariate) vs. Đa biến (Multivariate):** $X = (X_1)$ vs. $X = (X_1, X_2, \ldots, X_k)$.
- **Tham số (Parametric), phi tham số (Nonparametric), bán tham số (Semiparametric):**
  - Parametric: $Y = \theta_0 + \theta_1 X$ (số tham số hữu hạn).
  - Nonparametric: $Y = \theta_0 + f(X)$ (số tham số vô hạn).
  - Semiparametric: $Y = \theta_0 + \theta_1 X_1 + f(X_2)$ (hữu hạn tham số quan tâm).
- **Đối xứng (Symmetric) vs. Bất đối xứng (Asymmetric):**
  - Symmetric: mô hình mô tả (log-linear models).
  - Asymmetric: mô hình dự đoán (generalized linear models).

2. HỒI QUY TUYẾN TÍNH ĐƠN BIẾN

2.1. Giả thuyết và hàm mục tiêu

Mô hình: $h_\theta(x) = \theta_0 + \theta_1 x$
Ký hiệu:

- N: số mẫu huấn luyện.
- $x^{(i)}, y^{(i)}$: giá trị đầu vào/đầu ra của mẫu thứ i.

Sai số dự đoán (residual): $e^{(i)} = h_\theta(x^{(i)}) - y^{(i)}$
Hàm chi phí (Cost function – MSE):

$$J(\theta_0, \theta_1) = \frac{1}{2N} \sum_{i=1}^{N} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

Mục tiêu: Tìm (θ_0^*, θ_1^*) để tối thiểu hóa $J(\theta_0, \theta_1)$.

2.2. Phương pháp Gradient Descent

Ý tưởng: Xuất phát từ điểm ngẫu nhiên (θ_0, θ_1), lặp lại cập nhật để giảm J:
Thuật toán (cập nhật đồng thời θ_0 và θ_1):

$$\theta_j := \theta_j - \alpha \frac{\partial J}{\partial \theta_j}$$

$$\frac{\partial J}{\partial \theta_0} = \frac{1}{N}\sum_{i=1}^{N}(h_\theta(x^{(i)}) - y^{(i)}), \quad \frac{\partial J}{\partial \theta_1} = \frac{1}{N}\sum_{i=1}^{N}(h_\theta(x^{(i)}) - y^{(i)}) \cdot x^{(i)}$$

Learning rate α:

- α quá nhỏ: học chậm, mất nhiều vòng lặp.
- α quá lớn: khó hội tụ, J có thể không giảm tại mỗi bước.
- Thực nghiệm: thử $\alpha \in \{0.001, 0.003, 0.01, 0.03, 0.1, 0.3, \ldots\}$.

2.3. Công thức trực tiếp (Closed-form)

Khi $\theta_0 = 0$: Dừng lại ở $\hat{\theta}_1 = \frac{\Sigma(x^{(i)} - \bar{x})(y^{(i)} - \bar{y})}{\Sigma(x^{(i)} - \bar{x})^2}$ và $\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x}$.

3. HỒI QUY TUYẾN TÍNH ĐA BIẾN

3.1. Mô hình đa biến

$$h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_n x_n = \theta^T x$$

(với $x_0 = 1$)

Hàm chi phí:

$$J(\theta) = \frac{1}{2N}\sum_{i=1}^{N}(h_\theta(x^{(i)}) - y^{(i)})^2$$

Gradient descent đa biến:

$$\theta_j := \theta_j - \alpha \frac{1}{N}\sum_{i=1}^{N}(h_\theta(x^{(i)}) - y^{(i)}) \cdot x_j^{(i)}, \quad j = 0, 1, \ldots, n$$

3.2. Feature Scaling (Chuẩn hóa đặc trưng)

Vấn đề: Các thuộc tính có thang đo khác nhau (ví dụ: kích thước nhà 0–2000 ft², số phòng 1–5) làm gradient descent hội tụ chậm do contour plot J bị kéo dài.

Giải pháp: Chuẩn hóa tất cả đặc trưng về cùng thang đo, ví dụ $x'_j \approx [-1, 1]$:

$$x'_j = \frac{x_j - \mu_j}{\text{range}_j} \quad \text{(mean normalization)}$$

3.3. Phương pháp Normal Equation

Thay vì dùng gradient descent lặp, có thể tính trực tiếp:

$$\theta = (X^T X)^{-1} X^T Y$$

trong đó X là ma trận $N \times (n+1)$ (thêm cột $x_0 = 1$), Y là vector $N \times 1$.

So sánh Gradient Descent và Normal Equation:

Gradient Descent	**Normal Equation**
Cần chọn α	Không cần chọn α
Cần nhiều vòng lặp	Không cần lặp
Hoạt động tốt khi n lớn ($n = 10^6$)	Phải tính $(X^T X)^{-1} - O(n^3)$
	Không hiệu quả khi $n > 10^4$

Vấn đề không khả nghịch (Non-invertible): Xảy ra khi: (1) các biến phụ thuộc tuyến tính (ví dụ: kích thước m và feet); (2) $n > N$ (nhiều biến hơn mẫu).

4. HỒI QUY PHI TUYẾN

Khi Y là hàm phi tuyến theo tham số θ:

$$Y = f(X, \theta) \quad \text{(ví dụ: hàm mũ, logarit, Gaussian)}$$

Xác định θ tối ưu bằng: thuật toán tối ưu cục bộ (local optimization) hoặc tối ưu toàn cục (global optimization) tối thiểu hóa tổng bình phương sai số.

5. ỨNG DỤNG VÀ VẤN ĐỀ TRONG HỒI QUY

5.1. Ứng dụng

- **Tiền xử lý DM:** Làm mịn dữ liệu, loại bỏ nhiễu.
- **Nhiệm vụ DM:** Dự đoán giá trị số, phân tích mô tả.
- **Các lĩnh vực:** Sinh học, nông nghiệp, kinh tế, kinh doanh, tài chính, bảo hiểm, e-commerce, khoa học, robot, điều khiển tự động.

5.2. Đánh giá mô hình hồi quy

Phương pháp tách dữ liệu:

- **Holdout method:** Chia ngẫu nhiên D thành tập huấn luyện (2/3) và tập kiểm tra (1/3).
- **K-fold cross-validation:** Chia D thành k phần bằng nhau. Lặp k lần: dùng phần k để kiểm tra, phần còn lại để huấn luyện. Tính trung bình độ chính xác.

Các độ đo đánh giá độ chính xác:

- **SSE (Sum of Squared Errors):** $\text{SSE} = \sum_{i=1}^{n}(\hat{y}_i - y_i)^2$ – đo lường tổng thể, nhỏ hơn tốt hơn.
- **MSE (Mean Squared Error):** $\text{MSE} = \frac{\text{SSE}}{n-m}$ (m: số hệ số hồi quy) – đo phương sai còn lại, nhỏ hơn tốt hơn.
- **S (Standard Error of Estimate):** $S = \sqrt{\text{MSE}}$ – sai số trung bình trong dự đoán, đo độ chính xác của dự đoán.

5.3. Vấn đề trong hồi quy

- Giả định về phân phối dữ liệu (quan hệ giữa biến độc lập và phụ thuộc).
- Tính độc lập của các biến dự báo.
- Biến phải liên tục (cả predictors và responses).
- Lượng dữ liệu không đủ lớn.
- Xác định dạng mô hình hồi quy phù hợp.
- **Kỹ thuật nâng cao:** ANN (Artificial Neural Network), SVM (Support Vector Machine).

Các yếu tố ảnh hưởng đến thành công của mô hình hồi quy:

1. Xác định bài toán đúng đắn.
2. Chọn biến quan trọng và dạng mô hình.
3. Tập dữ liệu tốt (cả về khối lượng và chất lượng).
4. Thuật toán ước lượng hệ số tốt (ví dụ: gradient descent).
5. Kỹ thuật kiểm định mô hình.

6. TÓM TẮT

- Hồi quy là kỹ thuật thống kê dự đoán giá trị số thực liên tục.
- Phân loại: Linear/Non-linear, Univariate/Multivariate, Parametric/Nonparametric/Semiparametric.
- Đơn biến: $h_\theta(x) = \theta_0 + \theta_1 x$; tối thiểu hóa J bằng gradient descent.
- Đa biến: $h_\theta(x) = \theta^T x$; cần feature scaling; có thể dùng normal equation.
- Đánh giá: SSE, MSE, S; holdout hoặc k-fold cross-validation.
- Đơn giản nhưng hữu ích; ứng dụng rộng rãi; là ví dụ điển hình về đóng góp của thống kê vào DM.

7. CÂU HỎI TỰ LUẬN

Câu 1. Hồi quy (Regression) là gì? Phân biệt hồi quy với phân lớp (Classification). Nêu ít nhất 3 ví dụ ứng dụng thực tiễn của hồi quy trong khai phá dữ liệu.

Câu 2. Giải thích phương trình hồi quy $Y = f(X, \theta)$. Các thành phần X, Y, θ biểu diễn điều gì? Ý nghĩa của dấu và độ lớn của hệ số θ_i là gì?

Câu 3. Trình bày đầy đủ các loại mô hình hồi quy: linear/nonlinear, univariate/multivariate, parametric/nonparametric/semiparametric. Cho ví dụ cụ thể cho mỗi loại.

Câu 4. Hàm chi phí (Cost function) $J(\theta_0, \theta_1)$ trong hồi quy tuyến tính đơn biến là gì? Tại sao sử dụng MSE thay vì MAE (Mean Absolute Error)?

Câu 5. Mô tả chi tiết thuật toán **Gradient Descent** cho hồi quy đơn biến. Tại sao cần cập nhật θ_0 và θ_1 **đồng thời**? Điều gì xảy ra nếu cập nhật tuần tự?

Câu 6. Learning rate α trong Gradient Descent ảnh hưởng đến quá trình học như thế nào? Trình bày các trường hợp α quá nhỏ, α quá lớn và α phù hợp. Cách chọn α thực tế?

Câu 7. Trình bày mô hình hồi quy tuyến tính đa biến. Sự khác biệt về thuật toán Gradient Descent giữa đơn biến và đa biến là gì?

Câu 8. **Feature Scaling** là gì? Tại sao Feature Scaling quan trọng trong hồi quy đa biến với Gradient Descent? Mô tả hiệu ứng của thiếu scaling lên contour plot của J.

Câu 9. So sánh **Gradient Descent** và **Normal Equation** để tối ưu hồi quy đa biến. Khi nào nên dùng mỗi phương pháp?

Câu 10. Giải thích vấn đề **ma trận không khả nghịch (Non-invertible)** trong Normal Equation. Nguyên nhân và cách khắc phục như thế nào?

Câu 11. Phương trình Normal Equation là gì? Dẫn xuất công thức $\theta = (X^T X)^{-1} X^T Y$ và giải thích ý nghĩa của từng ma trận.

Câu 12. Hồi quy phi tuyến (Non-linear Regression) khác hồi quy tuyến tính như thế nào? Cho 3 ví dụ về hàm phi tuyến và tình huống thực tế phù hợp.

Câu 13. Trình bày phương pháp **K-fold Cross-Validation** để đánh giá mô hình hồi quy. Ưu điểm của K-fold so với Holdout method là gì? Giá trị k nào phổ biến?

Câu 14. Giải thích SSE, MSE và S (Standard Error of Estimate). Mối quan hệ giữa 3 độ đo này là gì? Khi nào thì mô hình được coi là có độ chính xác tốt?

Câu 15. Trình bày ví dụ hồi quy đa biến với tập dữ liệu bán hàng: $y = \text{Quantity Sold} = 8536.21 - 835.72 \times \text{Price} + 0.592 \times \text{Advertising}$. Giải thích ý nghĩa của các hệ số.

Câu 16. Giải thích tại sao hồi quy tuyến tính đơn biến với dạng $h_\theta(x) = \theta_0 + \theta_1 x$ tạo ra **convex cost function** J và không có local minima?

Câu 17. Trình bày ứng dụng của hồi quy trong **tiền xử lý dữ liệu** (cụ thể trong Data Smoothing và Noise Removal). Hồi quy được dùng như thế nào để loại bỏ nhiễu?

Câu 18. Phân biệt **Symmetric regression** và **Asymmetric regression**. Loại nào phù hợp cho nhiệm vụ dự đoán trong DM? Cho ví dụ cụ thể.

Câu 19. Nêu 5 yếu tố ảnh hưởng đến sự thành công của mô hình hồi quy. Yếu tố nào bạn cho là quan trọng nhất và tại sao?

Câu 20. So sánh hồi quy tuyến tính truyền thống với **ANN (Artificial Neural Network)** và **SVM (Support Vector Machine)** trong bối cảnh hồi quy nâng cao. Khi nào cần dùng các phương pháp nâng cao?

8. CÂU HỎI TRẮC NGHIỆM

Câu 1. Hồi quy (Regression) trong DM dùng để dự đoán:

A. Giá trị rời rạc (nhãn lớp).

B. Giá trị số thực và liên tục.

C. Các tập phổ biến trong tập giao dịch.

D. Số lượng cụm trong tập dữ liệu.

Câu 2. Mô hình hồi quy $Y = \theta_0 + \theta_1 X$ thuộc loại:

A. Nonparametric univariate.

B. Parametric univariate linear.

C. Semiparametric multivariate.

D. Nonlinear parametric.

Câu 3. Hàm chi phí $J(\theta_0, \theta_1)$ trong hồi quy tuyến tính được tối:

A. Tối đa hóa.

B. Tối thiểu hóa.

C. Giữ cố định.

D. Tối đa hóa ở bước đầu, tối thiểu hóa ở bước sau.

Câu 4. Sai số dự đoán (residual) $e^{(i)}$ được định nghĩa là:

A. $y^{(i)} - \bar{y}$

B. $h_\theta(x^{(i)}) - y^{(i)}$

C. $x^{(i)} - \bar{x}$

D. $\theta_1 x^{(i)} - \theta_0$

Câu 5. Trong Gradient Descent, learning rate α quá lớn dẫn đến:

A. Hội tụ rất nhanh đến nghiệm tối ưu.

B. Hội tụ chậm, cần nhiều vòng lặp.

C. $J(\theta)$ có thể không giảm tại mỗi bước, thuật toán có thể không hội tụ.

D. Gradient descent luôn hội tụ bất kể α.

Câu 6. Trong Gradient Descent đơn biến, cập nhật θ_0 và θ_1 phải:

A. Cập nhật θ_0 trước, sau đó mới dùng θ_0 mới để cập nhật θ_1.

B. Cập nhật θ_1 trước, sau đó cập nhật θ_0.

C. Cập nhật đồng thời (simultaneously update).

D. Cập nhật luân phiên theo từng vòng lặp.

Câu 7. Feature Scaling trong hồi quy đa biến giúp:

A. Tăng độ chính xác của mô hình hồi quy.

B. Gradient Descent hội tụ nhanh hơn bằng cách cân bằng thang đo các đặc trưng.

C. Giảm overfitting.

D. Tự động chọn giá trị α tối ưu.

Câu 8. Normal Equation $\theta = (X^T X)^{-1} X^T Y$ có ưu điểm so với Gradient Descent là:

A. Hiệu quả hơn khi n rất lớn ($n > 10^6$).

B. Không cần chọn α và không cần vòng lặp.

C. Luôn tìm được nghiệm tối ưu toàn cục dù J không lồi.

D. Có thể áp dụng khi ma trận $(X^T X)$ không khả nghịch.

Câu 9. Ma trận $(X^T X)$ trong Normal Equation **không khả nghịch** khi:

A. Số mẫu N quá nhỏ và $n < N$.

B. Tồn tại biến phụ thuộc tuyến tính hoặc số biến n lớn hơn số mẫu N.

C. Tất cả biến đều độc lập tuyến tính.

D. α được chọn quá nhỏ.

Câu 10. Mô hình hồi quy $Y = \theta_0 + f(X)$ thuộc loại:

A. Parametric.

B. Nonparametric.

C. Semiparametric.

D. Linear.

Câu 11. Độ đo nào sau đây đo lường **tổng thể** sai số của mô hình hồi quy?

A. MSE.

B. S (Standard Error of Estimate).

C. SSE (Sum of Squared Errors).

D. R^2.

Câu 12. MSE (Mean Squared Error) được tính bằng:

A. $\text{SSE} \times (n-m)$

B. $\text{SSE}/(n-m)$

C. $\sqrt{\text{SSE}}$

D. SSE/n

Câu 13. Trong đánh giá mô hình hồi quy, **K-fold cross-validation** với $k = 10$ có nghĩa là:

A. Chia dữ liệu thành 10 phần, mỗi lần dùng 9 phần test và 1 phần train.

B. Chia dữ liệu thành 10 phần; mỗi lần dùng 1 phần test, 9 phần train; lặp 10 lần.

C. Lặp gradient descent 10 lần.

D. Chọn 10 giá trị α khác nhau và lấy trung bình.

Câu 14. Hồi quy tuyến tính đơn biến dạng $h_\theta(x) = \theta_0 + \theta_1 x^2$ thuộc loại:

A. Phi tuyến về tham số (nonlinear in parameters).

B. Tuyến tính về tham số (linear in parameters).

C. Phi tham số (nonparametric).

D. Bán tham số (semiparametric).

Câu 15. Dấu của hệ số θ_i trong mô hình hồi quy biểu diễn:

A. Độ lớn tuyệt đối của ảnh hưởng X_i lên Y.

B. Hướng ảnh hưởng (tương quan thuận/nghịch) của X_i lên Y.

C. Mức độ tương quan giữa X_i và các biến khác.

D. Xác suất biến X_i xuất hiện trong mô hình.

Câu 16. Standard Error S của mô hình hồi quy đo lường:

A. Tổng bình phương sai số.

B. Sai số trung bình trong quá trình dự đoán, đo độ chính xác của dự đoán.

C. Sai số khi ước lượng θ bằng gradient descent.

D. Số vòng lặp cần thiết để hội tụ.

Câu 17. Khi vẽ $J(\theta)$ theo số vòng lặp, thuật toán gradient descent **không hội tụ** khi:

A. $J(\theta)$ giảm đơn điệu qua mỗi vòng lặp.

B. $J(\theta)$ giảm nhanh lúc đầu rồi chậm dần.

C. $J(\theta)$ tăng hoặc dao động.

D. $J(\theta)$ bằng 0.

Câu 18. Với mô hình $y = 9323 - 823 \times \text{price}$ và $R^2 = 0.65$, R^2 cho biết:

A. 65% biến động của y được giải thích bởi mô hình.

B. 65% các dự đoán là chính xác.

C. SSE bằng 0.65.

D. Mô hình chính xác 65% trường hợp.

Câu 19. Phương pháp **Holdout** chia dữ liệu thành:

A. k phần bằng nhau.

B. Tập huấn luyện và tập kiểm tra (thường 2/3 và 1/3).

C. Tập huấn luyện, tập validation và tập kiểm tra.

D. 10 fold để cross-validation.

Câu 20. Hồi quy tuyến tính đơn biến với $J(\theta_0, \theta_1) = \text{MSE}$ có hình dạng của J là:

A. Hàm lõm (concave function).

B. Hàm lồi (convex function) – dạng bowl/paraboloid.

C. Hàm có nhiều điểm cực tiểu cục bộ.

D. Hàm tuyến tính phẳng.

Câu 21. Với hồi quy đa biến, khi $n = 10^6$ (triệu biến), nên dùng:

A. Normal equation vì không cần vòng lặp.

B. Gradient descent vì Normal equation tính $(X^T X)^{-1}$ rất tốn kém.

C. Cả hai đều hiệu quả như nhau.

D. Chỉ có thể dùng K-NN regression.

Câu 22. Kỹ thuật hồi quy nâng cao nào được đề cập trong slide C3?

A. K-Means và DBSCAN.

B. Logistic Regression và Decision Tree.

C. ANN (Artificial Neural Network) và SVM (Support Vector Machine).

D. Apriori và FP-Growth.

Câu 23. Mô hình $Y = \theta_0 + \theta_1 X_1 + f(X_2)$ thuộc loại:

A. Parametric.

B. Nonparametric.

C. Semiparametric.

D. Linear nonparametric.

Câu 24. Trong hồi quy, nếu $\theta_1 = -835.72$ (hệ số của biến Price), điều đó có nghĩa:

A. Giá tăng 1 đơn vị thì Quantity Sold tăng 835.72.

B. Giá tăng 1 đơn vị thì Quantity Sold giảm 835.72.

C. Quantity Sold luôn bằng $-835.72 \times$ Price.

D. Price và Quantity Sold không có quan hệ tuyến tính.

Câu 25. Hồi quy tuyến tính **đơn biến** có giả thuyết:

A. $h_\theta(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$

B. $h_\theta(x) = \theta_0 + \theta_1 x$

C. $h_\theta(x) = g(\theta^T x)$

D. $h_\theta(x) = \theta^T x + f(x)$

Câu 26. Trong Gradient Descent, điều kiện dừng (convergence) thường là:

A. Số vòng lặp đạt đúng n.

B. $J(\theta)$ không thay đổi đáng kể giữa các vòng lặp (hoặc thay đổi dưới ngưỡng).

C. $\theta_0 = \theta_1 = 0$.

D. $J(\theta) = 0$.

Câu 27. Ứng dụng của hồi quy trong **tiền xử lý dữ liệu** là:

A. Phân lớp điểm dữ liệu vào các nhóm.

B. Làm mịn và loại bỏ nhiễu dữ liệu.

C. Tìm các luật kết hợp.

D. Phân cụm dữ liệu.

Câu 28. Khi dùng Normal Equation với $n = 4$ biến và $N = 4$ mẫu, rủi ro lớn nhất là:

A. Gradient descent không hội tụ.

B. $(X^T X)$ có thể không khả nghịch do $n \approx N$ (quá ít mẫu so với biến).

C. α quá lớn.

D. Dữ liệu không có phân phối Gaussian.

Câu 29. Tổng bình phương sai số (SSE) trong đánh giá hồi quy được tính bằng:

A. $\sum_{i=1}^{n} |\hat{y}_i - y_i|$

B. $\sum_{i=1}^{n} (\hat{y}_i - y_i)^2$

C. $\frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$

D. $\sqrt{\frac{\text{SSE}}{n-m}}$

Câu 30. Mô hình $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$ biểu diễn:

A. Hồi quy phi tuyến về tham số.

B. Hồi quy tuyến tính về tham số (linear in parameters) với x^2 là một đặc trưng.

C. Hồi quy phi tham số.

D. Không thể xác định.

Câu 31. Trong cross-validation, mục đích của việc dùng dữ liệu không thấy trong quá trình huấn luyện (unseen test data) để đánh giá là:

A. Kiểm tra khả năng tổng quát hóa (generalization) của mô hình.

B. Tối ưu hóa thêm các tham số θ.

C. Giảm SSE của tập huấn luyện.

D. Tăng tốc độ huấn luyện.

ĐÁP ÁN

Câu hỏi tự luận – Hướng dẫn trả lời

Câu	Nội dung cần trình bày
1	Hồi quy: dự đoán giá trị số thực. Khác phân lớp ở output (liên tục vs. rời rạc). Ví dụ: dự đoán giá nhà, giá cổ phiếu, doanh số bán hàng.
2	X: biến độc lập/dự báo; Y: biến phụ thuộc/đáp ứng; θ: hệ số. Dấu θ_i: hướng ảnh hưởng (+: thuận, -: nghịch). Độ lớn $\lvert\theta_i\rvert$: mức độ ảnh hưởng.
3	Linear: $Y = \theta_0 + \theta_1 X$; Nonlinear: $Y = e^{\theta_1 X}$; Univariate: 1 biến đầu vào; Multivariate: nhiều biến; Parametric: hữu hạn tham số; Nonparametric: $Y = f(X)$; Semiparametric: $Y = \theta_1 X + f(X_2)$.
4	$J = \frac{1}{2N}\sum(h_\theta(x^{(i)}) - y^{(i)})^2$. MSE hơn MAE vì: khả vi liên tục (dễ lấy đạo hàm), phạt sai số lớn mạnh hơn, tối ưu hóa thuận lợi hơn.
5	Khởi tạo (θ_0, θ_1); lặp: tính gradient cho cả θ_0 và θ_1 bằng tham số hiện tại; cập nhật đồng thời. Nếu tuần tự: θ_1 cập nhật sau sẽ dùng θ_0 mới $\Rightarrow$ gradient tính sai, sai kết quả.
6	α nhỏ: J giảm chậm, nhiều vòng lặp. α lớn: J tăng/dao động, không hội tụ. Chọn thực tế: thử 0.001, 0.003, 0.01, ... theo cấp số nhân; vẽ J vs. số vòng lặp để kiểm tra.
7	Đa biến: $h_\theta(x) = \theta^T x = \theta_0 + \sum \theta_j x_j$. Gradient descent cập nhật $n+1$ tham số; công thức gradient: $\frac{\partial J}{\partial \theta_j} = \frac{1}{N}\sum(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$.
8	Feature Scaling: chuẩn hóa đặc trưng về $\approx [-1, 1]$. Thiếu scaling: contour plot của J bị kéo dài hình elip, gradient descent đi theo đường ziczac $\Rightarrow$ hội tụ chậm.
9	Gradient descent: cần chọn α, lặp nhiều; tốt với n lớn. Normal equation: không cần α, không lặp; phải tính $(X^TX)^{-1}$ $O(n^3)$; không dùng khi $n > 10^4$.
10	Non-invertible khi: (1) biến phụ thuộc tuyến tính (ví dụ: kích thước m và feet^2) $\rightarrow$ xóa biến trùng; (2) $n > N \rightarrow$ giảm số biến hoặc thu thập thêm dữ liệu.
11	Normal equation: đặt gradient = 0: $\frac{\partial J}{\partial \theta} = 0 \Rightarrow X^TX\theta = X^TY \Rightarrow \theta = (X^TX)^{-1}X^TY$. X: ma trận đặc trưng $N \times (n+1)$; Y: vector nhãn $N \times 1$; $(X^TX)^{-1}$: giả nghịch đảo.
12	Phi tuyến: Y là hàm phi tuyến theo θ. Ví dụ: (1) $Y = \theta_0 e^{\theta_1 X}$ – tăng trưởng; (2) $Y = \theta_0 \log(\theta_1 X)$ – suy giảm logarit; (3) $Y = \frac{1}{1+e^{-\theta^T X}}$ – sigmoid.

Câu	Nội dung cần trình bày
13	K-fold: chia thành k phần; lặp k lần, mỗi lần dùng 1 phần test; tính trung bình độ chính xác. Ưu điểm so với holdout: ít phụ thuộc cách chia dữ liệu, sử dụng toàn bộ dữ liệu để train/test. $k = 10$ phổ biến nhất.
14	SSE: tổng bình phương sai số (overall measure). MSE = SSE/(n-m): phương sai còn lại (nhỏ = ít phương sai chưa giải thích). $S = \sqrt{\text{MSE}}$: sai số trung bình dự đoán (đơn vị giống Y). Mô hình tốt khi SSE, MSE, S nhỏ.
15	$\theta_0 = 8536.21$: giá trị cơ sở khi Price=0 và Advertising=0 (không có ý nghĩa thực tế ở đây). $\theta_1 = -835.72$: Giá tăng 1\$ thì QS giảm 835.72 (quan hệ nghịch). $\theta_2 = 0.592$: Quảng cáo tăng 1\$ thì QS tăng 0.592.
16	$J(\theta_0, \theta_1)$ là hàm toàn phương (quadratic) – hình bowl (paraboloid lồi). Hàm lồi không có điểm cực tiểu cục bộ, chỉ có 1 cực tiểu toàn cục $\Rightarrow$ gradient descent luôn tìm được nghiệm tối ưu.
17	Hồi quy trong preprocessing: (1) Data smoothing – dùng đường hồi quy để làm mịn dữ liệu (thay giá trị thực bằng giá trị trên đường hồi quy); (2) Noise removal – xác định điểm xa đường hồi quy là nhiễu và loại bỏ.
18	Symmetric regression: mô hình mô tả – không phân biệt biến phụ thuộc/độc lập (ví dụ: log-linear). Asymmetric: mô hình dự đoán – phân biệt rõ X và Y (ví dụ: generalized linear). Trong DM dự đoán, dùng asymmetric.
19	5 yếu tố: (1) Xác định bài toán đúng đắn; (2) Chọn biến và dạng mô hình; (3) Dữ liệu tốt; (4) Thuật toán ước lượng tốt; (5) Kỹ thuật kiểm định. Quan trọng nhất: dữ liệu tốt – garbage in, garbage out.
20	Hồi quy tuyến tính: đơn giản, giải thích được, hoạt động tốt khi quan hệ tuyến tính. ANN: xử lý phi tuyến phức tạp, nhiều tham số, ít giải thích. SVM: hiệu quả với chiều cao, robust với outlier. Dùng nâng cao khi: quan hệ phi tuyến mạnh, dữ liệu phức tạp, yêu cầu độ chính xác cao.

Câu hỏi trắc nghiệm – Đáp án

Câu	ĐA	Câu	ĐA	Câu	ĐA	Câu	ĐA
1	B	11	C	21	B	31	B
2	B	12	B	22	C	32	B
3	B	13	B	23	C	33	B
4	B	14	B	24	B	34	B
5	C	15	B	25	B	35	B
6	C	16	B	26	B	36	B

Câu	ĐA	Câu	ĐA	Câu	ĐA	Câu	ĐA
7	B	17	C	27	B	37	B
8	B	18	A	28	C	38	A
9	B	19	B	29	B	39	B
10	B	20	B	30	B	40	A