🏘️ Câu chuyện về Bé Na và những căn nhà thần kỳ

Ngày xửa ngày xưa, ở một làng nhỏ, có một cô bé tên là Na rất thích đoán giá nhà. Mỗi khi đi qua một căn nhà, Na lại tự hỏi:

> "Nhà này rộng bao nhiêu, có mấy phòng ngủ, cách xa trung tâm thì giá sẽ ra sao nhỉ?"

Một hôm, bác thầy giáo dạy toán đến và nói với Na:

> "Có một phép toán kỳ diệu, tên là **Hồi Quy Tuyến Tính** (Linear Regression), giúp con dự đoán giá nhà dựa trên các con số!"

Đầu tiên, bác thầy giáo bảo Na hãy nghĩ:

- Diện tích nhà càng lớn thì giá càng cao
- Nhiều phòng ngủ hơn thì giá càng cao
- Xa trung tâm thì giá giảm đi một chút

Mô hình toán của bác là như vầy:

$$y \approx f(x) = w_1 x_1 + w_2 x_2 + w_3 x_3 + w_0$$

Trong đó:

- (x_1): diện tích nhà (mét vuông)
- (x_2): số phòng ngủ
- (x_3): khoảng cách tới trung tâm (km)
- (y): giá nhà (triệu đồng)
- (w_1, w_2, w_3, w_0): các con số bí ẩn cần tìm!

Bác thầy giáo tiếp tục:

> "Na ơi, con hãy thử đoán giá một căn nhà rộng 100m², có 3 phòng ngủ, cách trung tâm 2 km nhé!"

Giả sử bác thầy giáo đã tìm được các số bí ẩn là:

- (w_1 = 0.5) (mỗi mét vuông thêm 0.5 triệu đồng)
- (w_2 = 10) (mỗi phòng ngủ thêm 10 triệu đồng)
- (w_3 = -2) (mỗi km xa trung tâm giảm 2 triệu đồng)
- (w_0 = 20) (giá khởi điểm 20 triệu đồng)

Cô bé Na sẽ tính:

$$y = 0.5 \times 100 + 10 \times 3 + (-2) \times 2 + 20$$

Na nhẩm tính:

- (0.5 \times 100 = 50)
- (10 \times 3 = 30)
- (-2 \times 2 = -4)
- (20) (giá khởi điểm)

Cộng lại:

$$y = 50 + 30 - 4 + 20 = 96$$

Vậy, cô bé Na đoán căn nhà này giá **96 triệu đồng**!

Nhưng Na thắc mắc:

> "Tại sao có lúc máy tính đoán chưa đúng nhỉ?"

Bác thầy giáo giải thích:

> "Vì mỗi căn nhà thực tế có thể khác một chút so với dự đoán, nên ta có một khái niệm tên là **sai số** (error):"

$$e = y - \hat{y}$$

Trong đó:

- (y): giá thực tế
- (\hat{y}): giá dự đoán

Để cho các lỗi nhỏ đi, ta tính tổng tất cả lỗi của nhiều căn nhà:

$$L(w) = \frac{1}{2} \sum_{i=1}^{N} (y_i - \bar{x}_i w)^2$$

Na hiểu rằng, càng làm cho giá trị này nhỏ, mô hình của mình càng dự đoán đúng.

Bác thầy giáo lại nói:

> "Muốn tìm các số bí ẩn (w) tốt nhất, ta phải giải một bài toán, tìm (w) sao cho hàm mất mát này nhỏ nhất!"

Và công thức tuyệt vời là:

$$w^* = \arg\min_w L(w)$$

Nghĩa là: tìm (w) sao cho tổng sai số nhỏ nhất!

Nếu có nhiều căn nhà, bác thầy giáo sẽ dùng đại số tuyến tính để tính nhanh hơn. Đặt:

- (\mathbf{y}): vector giá thực tế
- (\mathbf{\bar{X}}): ma trận thông tin các căn nhà
- (\mathbf{w}): vector các số bí ẩn

Công thức tổng quát:

$$L(\mathbf{w}) = \frac{1}{2}\|\mathbf{y} - \bar{\mathbf{X}}\mathbf{w}\|_2^2$$

Muốn tìm nghiệm tốt nhất, bác thầy giáo giải phương trình đạo hàm bằng 0:

$$\bar{\mathbf{X}}^T \bar{\mathbf{X}} \mathbf{w} = \bar{\mathbf{X}}^T \mathbf{y}$$

Nếu ma trận (\mathbf{\bar{X}}^T \mathbf{\bar{X}}) có thể nghịch đảo:

$$\mathbf{w} = (\bar{\mathbf{X}}^T \bar{\mathbf{X}})^{-1} \bar{\mathbf{X}}^T \mathbf{y}$$

Nếu không, ta dùng **giả nghịch đảo** (pseudo-inverse):

$$\mathbf{w} = (\bar{\mathbf{X}}^T \bar{\mathbf{X}})^\dagger \bar{\mathbf{X}}^T \mathbf{y}$$

Nhờ vậy, cô bé Na luôn tìm ra các số bí ẩn để đoán giá nhà, dù dữ liệu phức tạp thế nào.

Bác thầy giáo còn nhắc nhở:

> "Nếu có điểm ngoại lai (outlier), như một căn nhà nhỏ mà giá cực cao, mô hình sẽ bị sai lệch. Vậy nên phải lọc dữ liệu kỹ trước khi học!"

Và thế là, nhờ toán học đẹp đẽ và các công thức kỳ diệu, bé Na đã trở thành cô phù thủy nhỏ đoán giá nhà tài ba nhất làng!

💡 Tóm tắt:

- Mỗi yếu tố ảnh hưởng giá được gán một con số
- Dùng công thức tuyến tính để dự đoán giá
- Tìm các con số tối ưu bằng cách giảm sai số dự đoán
- Áp dụng đại số tuyến tính để tổng quát cho nhiều dữ liệu
- Luôn kiểm tra và xử lý dữ liệu thật tốt!

Câu chuyện kết thúc rồi, bé Na đã hiểu được sức mạnh của toán học trong dự đoán tương lai!