Linear Regression

10-701/15-781 Machine Learning - Recitation

January 26th 2006

Plan for today

- Linear regression
  - What is regression?
  - LR – derivation
  - LR – example
- Test set / training set error – example
- Ovefitting example

What is regression?

- Given some data (x_j, t_j)
  - E.g. x = {age, weight}, t={time to run a mile}
- t(x) is a random variable
- Want to predict the mean: $\hat{t}(x)$

What is regression?

- Hypothesis space
  - Linear regression: $\hat{t}(x) = \sum_i w_i f_i(x)$
    - Linear in w, not in x!
    - This is linear: $\hat{t}(x) = \sum_i w_i x^i$
    - This is also linear: $\hat{t}(x) = \sum_i w_i \sin(i^2 x^7)$
  - Nonlinear regression, e.g. $\hat{t}(x) = \sum_i e^{w_i x}$
- Minimize the loss function, e.g.

$$\sum_j (\hat{t}(x_j) - t_j)^2$$

Why linear regression?

- MLE if the noise is independent Gaussian
- Easy to compute – closed-form solution

Linear regression - derivation

- Hypothesis: $\hat{t}(x) = w_0 + \sum_i w_i f_i(x)$
- Want to minimize:

$$\sum_j (\hat{t}(x_j) - t_j)^2 = \sum_j ((w_0 + \sum_i w_i f_i(x)) - t_j)^2$$

Linear regression - derivation

$$\hat{t}(x) = w_0 + \sum_i w_i f_i(x)$$

w_0 stands out – put it inside the sum too

$$f_0(x) \equiv 1 \qquad \hat{t}(x) = \sum_{i=0}^{m} w_i f_i(x)$$

Vector notation:

$$\hat{t}(x) = \begin{pmatrix} 1 & f_1(x) & \dots & f_k(x) \end{pmatrix} \begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_k \end{pmatrix} = \vec{f}^{\,T}(x) w$$

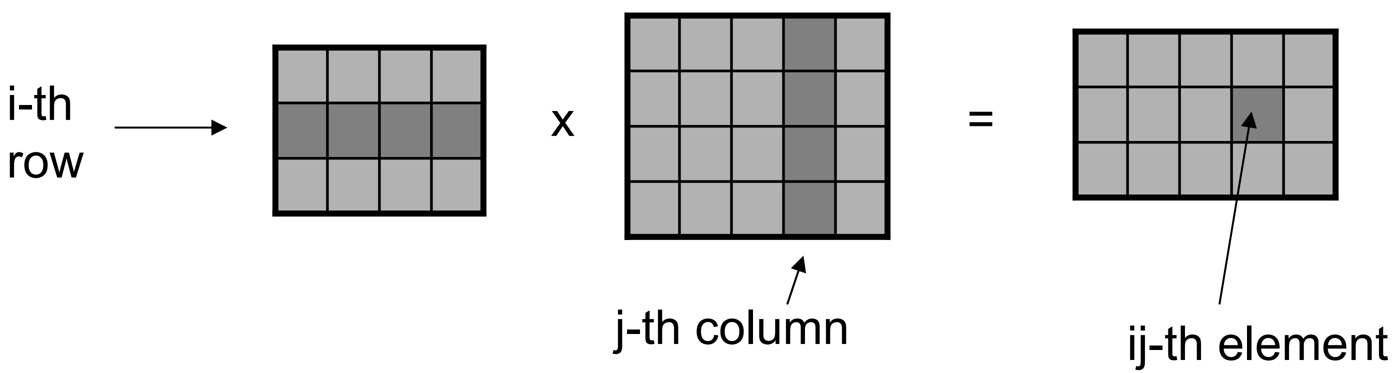
Matrices basics

- Matrix A → 2-dimensional array of numbers (n rows x m columns)
- a_{ij} → number on i-th row and j-th column
- Vector → (n x 1) matrix
- $C = A+B : c_{ij} = a_{ij} + b_{ij}$
- A^T – transpose – ‘rotated around diagonal’
  - $B = A^T \leftrightarrow b_{ij} = a_{ij}$
  - i.e. i-th row is now i-th column

Matrices basics

- Multiplication
  - (n by k) x (k by m) → (n by m)
  - C = AB ↔ $c_{ij} = \sum_k a_{ik} b_{kj}$
  - (AB)C = A(BC), (A+B)C = AC + BC
  - AI = IA = A, I – identity matrix (of the right size)
  - AB ≠ BA (even when BA is defined!)
  - A^{-1} – inverse: A x A^{-1} = A^{-1} x A = I
    - Not always exists!

$$I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix}$$



Linear regression - derivation

$$\sum_j (\hat{t}(x_j) - t_j)^2 = \begin{pmatrix} \hat{t}(x_1) - t_1 & \dots & \hat{t}(x_n) - t_n \end{pmatrix} \begin{pmatrix} \hat{t}(x_1) - t_1 \\ \vdots \\ \hat{t}(x_n) - t_n \end{pmatrix} = (\hat{t} - t)^T (\hat{t} - t)$$

$$\hat{t} = \begin{pmatrix} \hat{t}(x_1) \\ \vdots \\ \hat{t}(x_n) \end{pmatrix} = \begin{pmatrix} f_0(x_1) & \dots & f_k(x_1) \\ \vdots & \ddots & \vdots \\ f_0(x_n) & \dots & f_k(x_n) \end{pmatrix} \begin{pmatrix} w_0 \\ \vdots \\ w_k \end{pmatrix} = Fw$$

$$\sum_j (\hat{t}(x_j) - t_j)^2 = (Fw - t)^T (Fw - t)$$

Linear regression - derivation

- To minimize, take derivative w.r.t w (remember, w is a vector! → the derivative is a vector)

$$\frac{\partial}{\partial w}(Fw-t)^T(Fw-t)=\cdots$$

- Properties: $\frac{\partial}{\partial X}X^T X = 2X$ $\frac{\partial}{\partial X}AX = A^T$

- Therefore… $\frac{\partial}{\partial w}(Fw-t)^T(Fw-t)=\frac{\partial}{\partial w}(w^T F^T Fw-w^T F^T t-t^T Fw+t^T t)=$

$$=F^T Fw-2F^T t$$

Linear regression - derivation

$$F^T F w - F^T t = 0$$

under mild conditions $F^T F$ is invertible, so

$$w = (F^T F)^{-1} F^T t$$

We're done!