



Supervised Learning

Supervised learning
Linear regression
Geometric intuition
Computational aspects



Machine Learning

- **Supervised learning:** generate a function from training data ((input, output) pairs).
 - regression: predict a continuous value as a function of the input
 - classification: predict the class label of the input object
- **Unsupervised learning:** fit a model to a observations without a priori output.



Example

- Suppose we would like to predict the exchange rate of US dollars based on the rates of some other currencies, as well as the US dollar recent rate history.
- Formally: estimate $y(t) \in \mathbb{R}$ given $x(t) \in \mathbb{R}^n$
- For example

$$x(t) = \begin{pmatrix} \text{Euro} \\ \text{Yen} \\ \$US \text{ (yesterday)} \\ \$US \text{ (2 days ago)} \\ \text{Price of gold} \end{pmatrix} \in \mathbb{R}^5$$



Linear Regression

- Obtain the solution as a linear combination:

$$y(t) = \sum_{i=1}^n w_i x_i(t) = w^T x(t)$$

- How do we find the weights w ?
- Use a “training set” $\{(x(t), y(t))\}_{t=1}^m$ to “learn” w , then use w to obtain $y(t)$ from new (previously unseen) vectors $x(t)$.



Linear Regression

- Assumptions:
 - Linear model captures relationship between $x(t)$ and $y(t)$
 - Achieving a small error in the training phase will lead to good prediction performance.
- Let
$$J(w) = \sum_{t=1}^m (w^T x(t) - y(t))^2$$
- Solve the optimization problem (“least squares”)
$$w^* = \arg \min_w J(w)$$



Linear Regression

- Observation: if m (the size of the training set) is equal to $n = \dim(x(t))$, and the vectors $x(t)$ are linearly independent, there exists a vector of weights w , such that $J(w) = 0$
- Proof:

$$w^T x(t) = x(t)^T w = y(t) \quad \text{for } t = 1, \dots, m$$

$$\begin{pmatrix} x_1(1) & x_2(1) & \cdots & x_n(1) \\ x_1(2) & x_2(2) & \cdots & x_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(m) & x_2(m) & \cdots & x_n(m) \end{pmatrix} w_{(n \times 1)} = y_{(m \times 1)}$$



Normal Equations

- More interesting/typical case: size of the training set is much larger, $m \gg n$
- Theorem: Let $w^* = \arg \min_w J(w)$
then w^* is a solution of the Normal Equations:

$$\begin{aligned} (X^T X) w &= X^T y \\ (n \times n) \cdot (n \times 1) &= (n \times m) \cdot (m \times 1) \end{aligned}$$

(X is the (m x n) matrix from the previous slide)

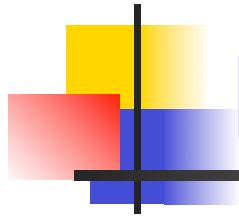


Proof

$$J(w) = \sum_{t=1}^m (w^T x(t) - y(t))^2 = \sum_{t=1}^m (x(t)^T w - y(t))^2$$

$$\begin{aligned} \frac{\partial J(w)}{\partial w} &= \sum_t \frac{\partial}{\partial w} (x(t)^T w - y(t))^2 \\ &= 2 \sum_t (w^T x(t) - y(t)) x(t)^T = \vec{0} \end{aligned}$$

$$\sum_t w^T x(t) x(t)^T = \sum_t y(t) x(t)^T$$

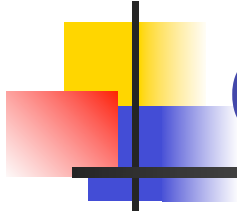


Proof (continued)

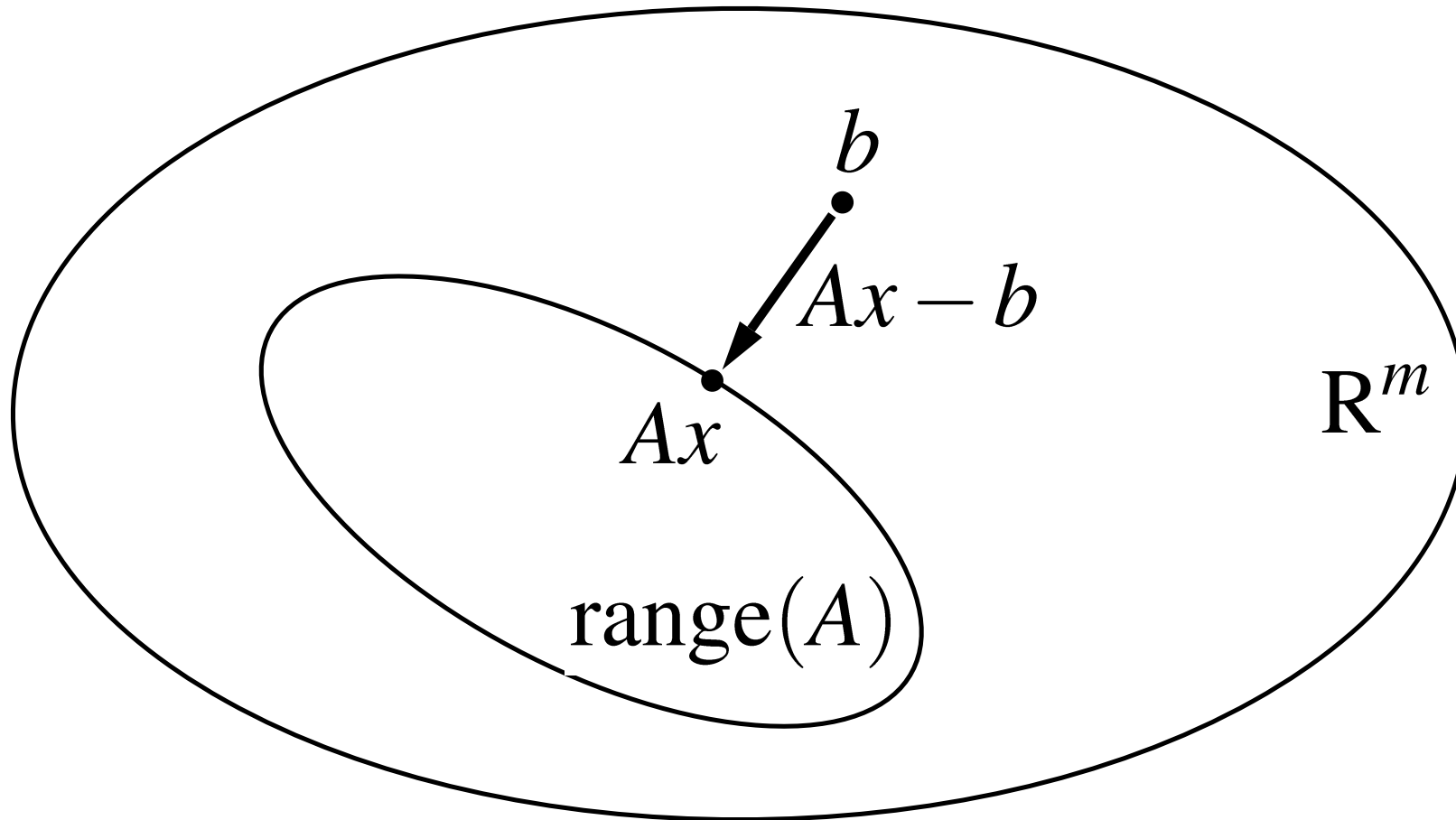
$$\sum_t w^T x(t) x(t)^T = \sum_t y(t) x(t)^T$$

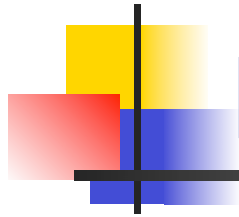
$$(w^T X^T) X = y^T X$$

$$(X^T X) w = X^T y$$



Geometric Interpretation





Normal Equations

overdetermined system:

$$Ax = b$$

define the "residual":

$$r = Ax - b$$

make it orthogonal to columns of A:

$$A_i^T r = 0$$

rewrite in matrix notation:

$$A^T r = 0$$

substitute definition of r:

$$A^T (Ax - b) = 0$$

$$A^T Ax = A^T b$$

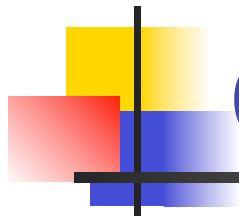


Solution

- If the matrix A has full column rank ($\text{rank}(A) = n$), then $A^T A$ is invertible, and

$$x = \underbrace{(A^T A)^{-1} A^T}_{\text{pseudoinverse}(A)} b$$

- But forming the pseudoinverse explicitly is rarely the best strategy.

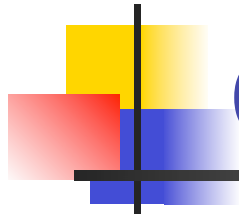


Cholesky Decomposition

- Let M be a symmetric positive definite matrix. Then there exists a decomposition

$$M = LL^T$$

- where L is an lower triangular matrix.
- Computational cost is $n^3/3$ (twice as fast as the generic LU decomposition).



Cholesky Decomposition

given the equation $A^T A x = A^T b$

compute Cholesky $LL^T = A^T A$

solve for y $Ly = A^T b$

solve for x $L^T x = y$



QR Decomposition

- Another alternative for least squares (slower, but with wider applicability)
- Compute $A = QR$, where Q is orthogonal, and R is upper triangular.

$$A^T Ax = A^T b$$

$$(QR)^T QRx = (QR)^T b$$

$$R^T Q^T QRx = R^T Q^T b$$

$$Rx = Q^T b$$



SVD Decomposition

- Let $A = U \Sigma V^T$ be the SVD decomposition of A , then the pseudoinverse of A is given by: $V \Sigma^+ U^T$
- Σ^+ is the transpose of Σ with every non-zero entry replaced by its reciprocal
- The solution to the least squares problem is given by
$$x = V \Sigma^+ U^T b$$
- Slowest, but most robust.