

Linear Regression

Aarti Singh

Machine Learning 10-315

Feb 7, 2022



MACHINE LEARNING DEPARTMENT



Supervised Learning Tasks

Classification

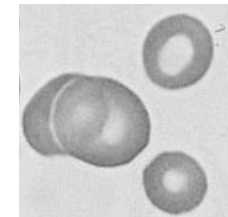
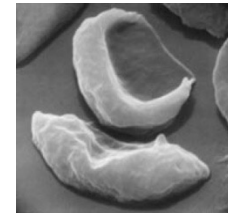


X = Document



Sports
Science
News

Y = Topic



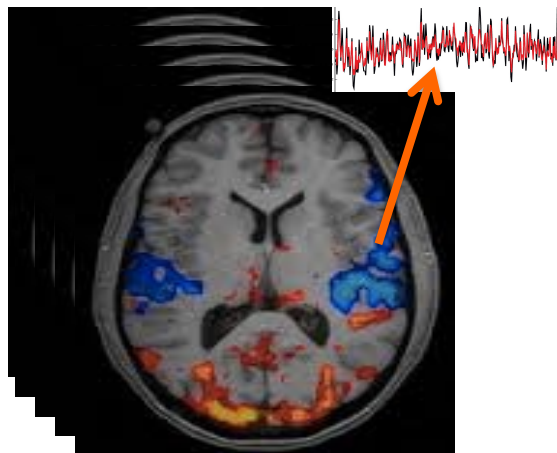
X = Cell Image



Anemic cell
Healthy cell

Y = Diagnosis

Regression



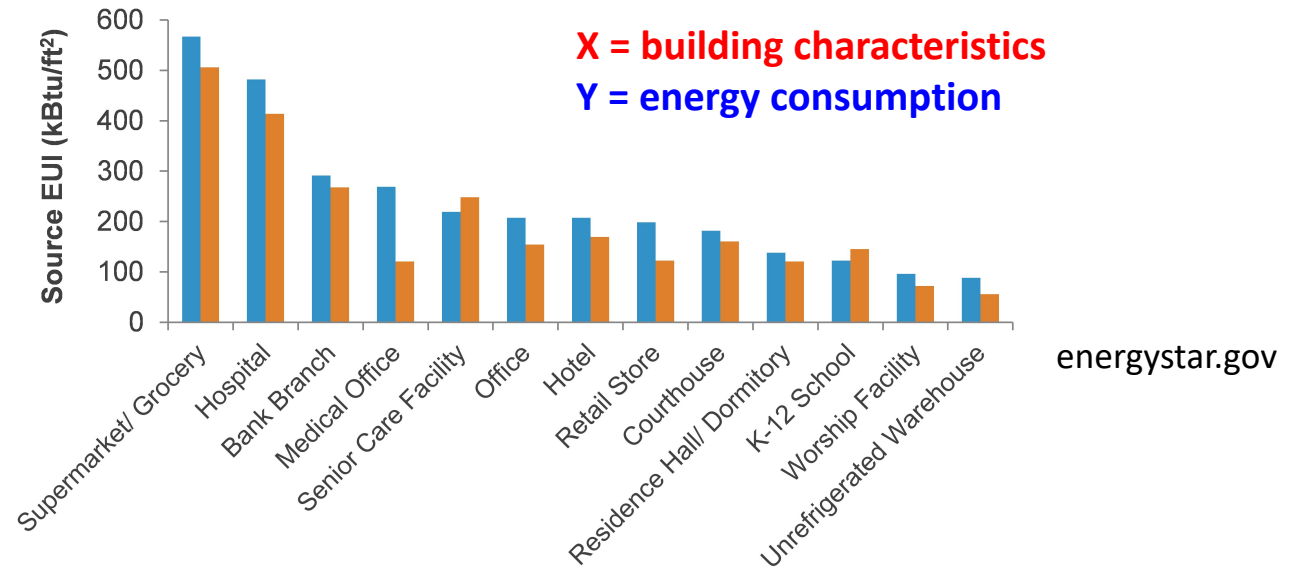
X = Brain Scan



Y = Age of a subject

Regression Tasks

Estimating
Energy Usage



Estimating
Contamination



Mean Squared Error (MSE) Minimization

Optimal predictor: $f^* = \arg \min_f \mathbb{E}[(f(X) - Y)^2]$

Empirical Minimizer: $\hat{f}_n = \arg \min_{f \in \mathcal{F}} \underbrace{\left(\frac{1}{n} \sum_{i=1}^n (f(X_i) - Y_i)^2 \right)}_{\text{Empirical mean}}$

Law of Large Numbers:

$$\frac{1}{n} \sum_{i=1}^n [\text{loss}(Y_i, f(X_i))] \xrightarrow{n \rightarrow \infty} \mathbb{E}_{XY} [\text{loss}(Y, f(X))]$$

Restrict class of predictors

Optimal predictor: $f^* = \arg \min_f \mathbb{E}[(f(X) - Y)^2]$

Empirical Minimizer: $\hat{f}_n = \arg \min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (f(X_i) - Y_i)^2$

Class of predictors

➤ Why?

- $\mathcal{F}$ - Class of Linear functions
- Class of Polynomial functions
- Class of nonlinear functions

Linear Regression

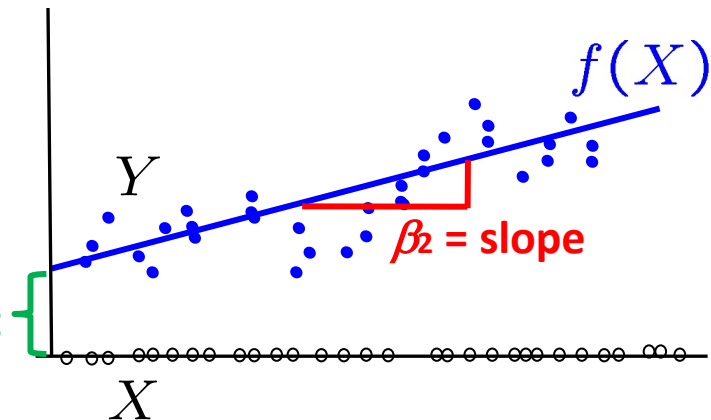
$$\hat{f}_n^L = \arg \min_{f \in \mathcal{F}_L} \frac{1}{n} \sum_{i=1}^n (f(X_i) - Y_i)^2 \quad \text{Least Squares Estimator}$$

$\mathcal{F}_L$ - Class of Linear functions

Uni-variate case:

$$f(X) = \beta_1 + \beta_2 X$$

β_1 - intercept



Multi-variate case:

$$f(X) = f(X^{(1)}, \dots, X^{(p)}) = \beta_1 X^{(1)} + \beta_2 X^{(2)} + \dots + \beta_p X^{(p)}$$

$$= X\beta \quad \text{where} \quad X = [X^{(1)} \dots X^{(p)}], \quad \beta = [\beta_1 \dots \beta_p]^T$$

Linear Regression (Matrix-vector form)

$$\hat{f}_n^L = \arg \min_{f \in \mathcal{F}_L} \frac{1}{n} \sum_{i=1}^n (f(X_i) - Y_i)^2 \quad f(X_i) = X_i \beta$$



$$\hat{\beta} = \arg \min_{\beta} \frac{1}{n} \sum_{i=1}^n (X_i \beta - Y_i)^2 \quad \hat{f}_n^L(X) = X \hat{\beta}$$

$$= \arg \min_{\beta} \frac{1}{n} (\mathbf{A} \beta - \mathbf{Y})^T (\mathbf{A} \beta - \mathbf{Y})$$

$$\mathbf{A} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} = \begin{bmatrix} X_1^{(1)} & \dots & X_1^{(p)} \\ \vdots & \ddots & \vdots \\ X_n^{(1)} & \dots & X_n^{(p)} \end{bmatrix} \quad \mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$$

Linear Regression

$$\hat{\beta} = \arg \min_{\beta} \frac{1}{n} (\mathbf{A}\beta - \mathbf{Y})^T (\mathbf{A}\beta - \mathbf{Y}) = \arg \min_{\beta} J(\beta)$$

$$J(\beta) = (\mathbf{A}\beta - \mathbf{Y})^T (\mathbf{A}\beta - \mathbf{Y})$$

➤ Poll

Is the objective convex in β ?

- A) Convex, quadratic in β
- B) Non-convex, $\mathbf{A}$ may not be positive semi definite
- C) Depends on conditioning (ratio of max:min eigenvalues) of $\mathbf{A}^T\mathbf{A}$
- D) Convex, $\mathbf{A}^T\mathbf{A}$ is positive semi definite

Linear Regression

$$\hat{\beta} = \arg \min_{\beta} \frac{1}{n} (\mathbf{A}\beta - \mathbf{Y})^T (\mathbf{A}\beta - \mathbf{Y}) = \arg \min_{\beta} J(\beta)$$

$$J(\beta) = (\mathbf{A}\beta - \mathbf{Y})^T (\mathbf{A}\beta - \mathbf{Y})$$

$$\left. \frac{\partial J(\beta)}{\partial \beta} \right|_{\hat{\beta}} = 0$$

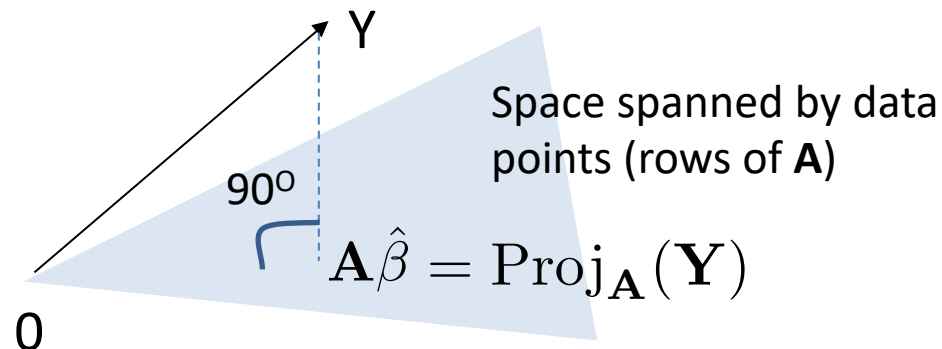
Linear regression solution satisfies Normal Equations

$$\underbrace{(\mathbf{A}^T \mathbf{A})}_{p \times p} \underbrace{\hat{\boldsymbol{\beta}}}_{p \times 1} = \underbrace{\mathbf{A}^T \mathbf{Y}}_{p \times 1}$$

If $(\mathbf{A}^T \mathbf{A})$ is invertible,

$$\hat{\boldsymbol{\beta}} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{Y} \qquad \hat{f}_n^L(X) = X \hat{\boldsymbol{\beta}}$$

Predicted labels for training points $\mathbf{A} \hat{\boldsymbol{\beta}} = \text{Proj}_{\mathbf{A}}(\mathbf{Y})$



Linear regression solution satisfies Normal Equations

$$\underbrace{(\mathbf{A}^T \mathbf{A})}_{p \times p} \underbrace{\hat{\beta}}_{p \times 1} = \underbrace{\mathbf{A}^T \mathbf{Y}}_{p \times 1}$$

If $(\mathbf{A}^T \mathbf{A})$ is invertible,

$$\hat{\beta} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{Y} \qquad \hat{f}_n^L(X) = X \hat{\beta}$$

Later: When is $(\mathbf{A}^T \mathbf{A})$ invertible ?

Now: What if $(\mathbf{A}^T \mathbf{A})$ is invertible but expensive (p very large)?

Gradient Descent

Even when $(\mathbf{A}^T \mathbf{A})$ is invertible, might be computationally expensive if $\mathbf{A}$ is huge.

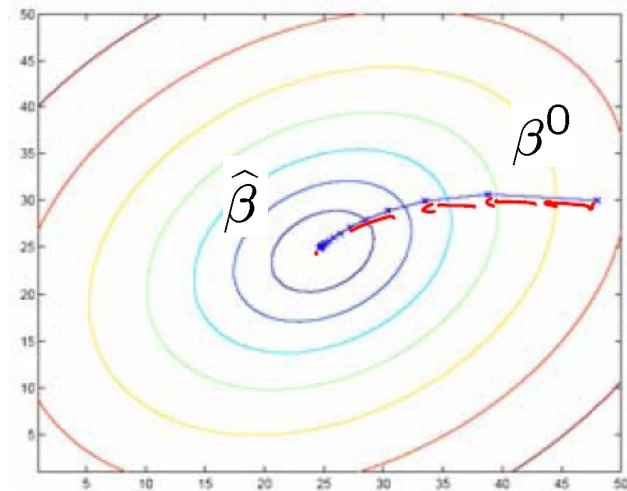
$$\hat{\beta} = \arg \min_{\beta} \frac{1}{n} (\mathbf{A}\beta - \mathbf{Y})^T (\mathbf{A}\beta - \mathbf{Y}) = \arg \min_{\beta} J(\beta)$$

Since $J(\beta)$ is convex, move along negative of gradient

Initialize: β^0

step size

Update:
$$\beta^{t+1} = \beta^t - \frac{\alpha}{2} \frac{\partial J(\beta)}{\partial \beta} \bigg|_t$$
$$= \beta^t - \alpha \underbrace{\mathbf{A}^T (\mathbf{A}\beta^t - \mathbf{Y})}_{0 \text{ if } \hat{\beta} = \beta^t}$$



Stop: when some criterion met e.g. fixed # iterations, or $\left. \frac{\partial J(\beta)}{\partial \beta} \right|_{\beta^t} < \epsilon$.

Least Square solution satisfies Normal Equations

$$\left. \frac{\partial J(\beta)}{\partial \beta} \right|_{\hat{\beta}} = 0 \quad \text{gives} \quad \underbrace{(\mathbf{A}^T \mathbf{A})}_{p \times p} \underbrace{\hat{\beta}}_{p \times 1} = \underbrace{\mathbf{A}^T \mathbf{Y}}_{p \times 1}$$

If $(\mathbf{A}^T \mathbf{A})$ is invertible,

1) If dimension p not too large, analytical solution:

$$\hat{\beta} = (\mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T \mathbf{Y} \qquad \hat{f}_n^L(X) = X \hat{\beta}$$

2) If dimension p is large, computing inverse is expensive $O(p^3)$

Gradient descent since objective is convex ($\mathbf{A}^T \mathbf{A} \succeq 0$)

$$\begin{aligned} \beta^{t+1} &= \beta^t - \frac{\alpha}{2} \left. \frac{\partial J(\beta)}{\partial \beta} \right|_t \\ &= \beta^t - \alpha \mathbf{A}^T (\mathbf{A} \beta^t - \mathbf{Y}) \end{aligned}$$

Linear regression solution satisfies Normal Equations

$$\underbrace{(\mathbf{A}^T \mathbf{A})}_{p \times p} \underbrace{\hat{\boldsymbol{\beta}}}_{p \times 1} = \underbrace{\mathbf{A}^T \mathbf{Y}}_{p \times 1}$$

When is $(\mathbf{A}^T \mathbf{A})$ invertible?

Recall: Full rank matrices are invertible. What is rank of $(\mathbf{A}^T \mathbf{A})$?