ISE 589/OR591

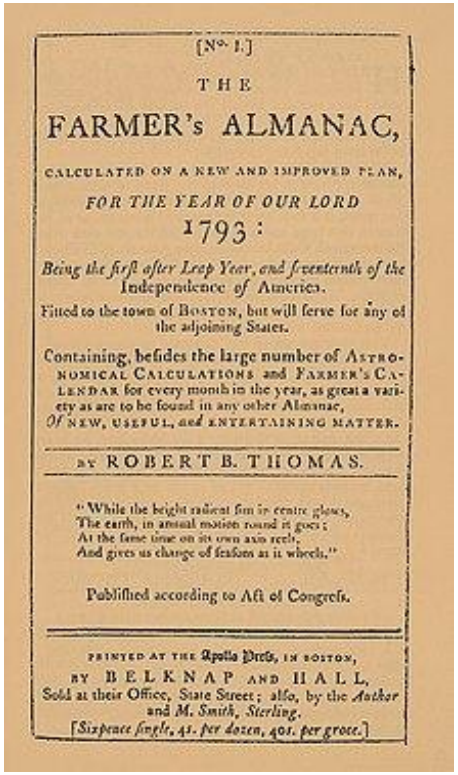
LECTURE 4 - PREDICTION

Unconstrained optimization models for machine learning

1. Least squares estimation
2. Linear regression

Prediction

- Prediction (forecasting) is a process of finding the hidden relationship between Input data and output data for future telling
- It is often understood as learning from past experiences to tell outcomes of future events



[No. I.]

THE

FARMER's ALMANAC,

CALCULATED ON A NEW AND IMPROVED PLAN,

FOR THE YEAR OF OUR LORD

1793:

Being the first after Leap Year, and seventeenth of the Independence *of* America.

Fitted to the town of BOSTON, but will serve for any of the adjoining States.

Containing, besides the large number of ASTRONOMICAL CALCULATIONS and FARMER's CALENDAR for every month in the year, as great a variety as are to be found in any other Almanac, *Of* NEW, USEFUL, *and* ENTERTAINING MATTER.

BY ROBERT B. THOMAS.

"While the bright radiant sun in centre glows,
The earth, in annual motion round it goes;
At the same time on its own axis reels,
And gives us change of seasons as it wheels."

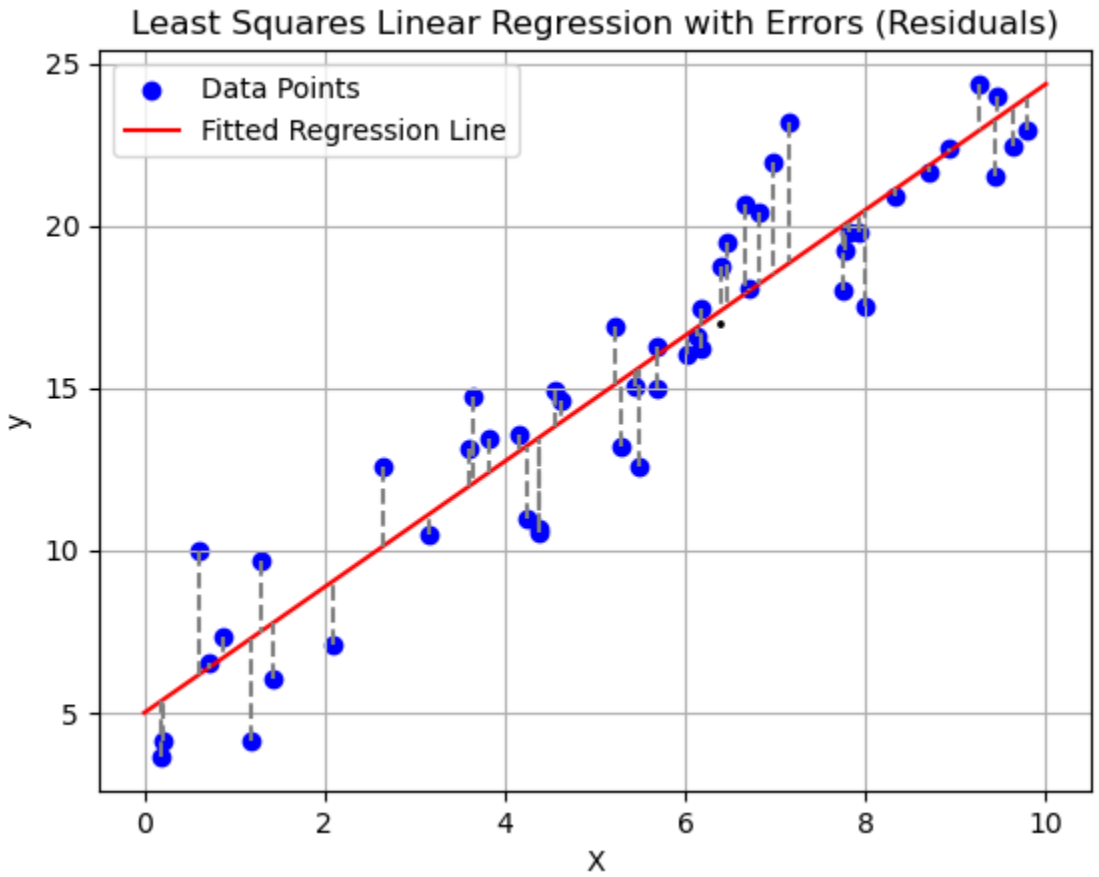
Published according to Act of Congress.

PRINTED AT THE Apollo Press, IN BOSTON,
BY BELKNAP AND HALL,
Sold at their Office, State Street; also, by the *Author* and *M. Smith, Sterling.*
[*Sixpence single, 4s. per dozen, 40s. per groce.*]

Prediction

- **Predictive analytics**, or **predictive AI**, encompasses a variety of statistical techniques from data mining, predictive modeling, and machine learning that analyze current and historical facts to make predictions about future or otherwise unknown events. https://en.wikipedia.org/wiki/Predictive_analytics
- Some useful approaches
  - Moving average models (such as single moving average, autoregressive integrated moving average, ARIMA)
  - Time series models
  - Least-squares models
  - Linear regression
  - Support vector regression (SVR)

Linear regression



https://medium.com/physics-and-machine-learning/misconceptions-about-least-square-regression-1131841d240f

Basics of data and linear algebra – vector norm

- (norm) A vector norm $||\cdot||$ is a function from $\mathbb{R}^n$ to $\mathbb{R}$ such that (i) $||\boldsymbol{x}|| \geq 0$ for any $\boldsymbol{x} \in \mathbb{R}^n$, and $||x||= 0$ if and only if $\boldsymbol{x} = 0$; (ii) $||\boldsymbol{x^1} + \boldsymbol{x^2}|| \leq ||\boldsymbol{x^1}|| + ||\boldsymbol{x^2}||$ for any $\boldsymbol{x^1}, \boldsymbol{x^2} \in \mathbb{R}^n$; (iii) $||\alpha\boldsymbol{x}|| = \alpha||\boldsymbol{x}||$ for any $\alpha \in \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}^n$.
- (commonly used norm)
  - l_0 norm: $\|\boldsymbol{x}\|_0 = \sum_{i=1}^{n} \mathbb{1}(x_i \neq 0)$;
  - l_1 norm: $\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i|$;
  - l_2 norm: $\|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$;
  - l_p norm: $\|\boldsymbol{x}\|_p = (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}$, $p \geq 1$;
  - l_∞ norm: $\|\boldsymbol{x}\|_\infty = \max_{i=1,\ldots,n} |x_i|$.

Gradient functions with 2-norm functions

Gradient of function $f(\cdot)$ at point $\boldsymbol{x} \in \mathbb{R}^n$: $\nabla f(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\boldsymbol{x})}{\partial x_n} \end{pmatrix}$

1. $\nabla\left(\boldsymbol{c}^T\boldsymbol{x}\right) = \boldsymbol{c}, \quad \forall\, \boldsymbol{c}, \boldsymbol{x} \in \mathbb{R}^n$
2. $\nabla(\|\boldsymbol{x}\|_2) = \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|_2}, \quad \forall\, \boldsymbol{x} \in \mathbb{R}^n$
3. $\nabla(\|\boldsymbol{x}\|_2^2) = 2\boldsymbol{x}, \quad \forall\, \boldsymbol{x} \in \mathbb{R}^n$
4. $\nabla(\boldsymbol{x}^T M\boldsymbol{x}) = (M + M^T)\boldsymbol{x}, \quad \forall\, M \in \boldsymbol{M_{n\times n}}$ and $\boldsymbol{x} \in \mathbb{R}^n$

5. $\nabla\left(\|A\boldsymbol{x} - \boldsymbol{b}\|_2^2\right) = 2\left(A^TA\right)\boldsymbol{x} - 2A^T\boldsymbol{b}$, $\forall A \in M_{m\times n}, \boldsymbol{b} \in \mathbb{R}^m$ and $\boldsymbol{x} \in \mathbb{R}^n$

System of linear equations

- Given a system of m linear equations in n real variables:

$$\begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \Leftrightarrow A\boldsymbol{x} = \boldsymbol{b}$$

- Facts from linear algebra:

  Case 1. When $m \leq n$ and A has full rank (rows of A are linear independent), the system has feasible solutions

  Case 2. When $m > n$, too many equations to satisfy may lead to infeasibility of the system!

Least squares optimization model

- Case 1. When $m \leq n$ and A has full rank, the system $A\boldsymbol{x} = \boldsymbol{b}$ has feasible solutions

  Question: How to find a feasible solution?

  - Linear programming model (how?)
  - Least squares optimization model

- Case 2. When $m > n$, too many equations to satisfy may lead to infeasibility of the system!

  Question: What to do?

  - Find an approximate solution

Least squares optimization model

- Case 1. When $m \leq n$ and A has full rank, the system $A\boldsymbol{x} = \boldsymbol{b}$ has feasible solutions
- Task: Find a feasible solution to the system

- Reasoning: (i) $\boldsymbol{x}$ is feasible if and only if $A\boldsymbol{x} - \boldsymbol{b} = \boldsymbol{0}$
  (ii) $\|\boldsymbol{v}\|_2 = 0$ if and if $\mathbf{vector}\ \boldsymbol{v} = \boldsymbol{0}$

- Optimization model:

$$\text{Minimize } \|A\boldsymbol{x} - \boldsymbol{b}\|_2^2$$
$$\text{s.t.} \quad \boldsymbol{x} \in \mathbb{R}^n$$

Least squares optimization model

- Optimization model:

$$\text{Minimize } \|A\boldsymbol{x} - \boldsymbol{b}\|_2^2$$
$$\text{s.t.} \quad \boldsymbol{x} \in \mathbb{R}^n$$

Solution method:

- Optimal solutions of a unconstrained convex optimization problem achieved at the point with zero gradient!

$$\nabla \|A\boldsymbol{x} - \boldsymbol{b}\|_2^2 = 0 \Leftrightarrow 2\left(A^T A\right)\boldsymbol{x} - 2A^T\boldsymbol{b} = 0$$
$$\Leftrightarrow \left(A^T A\right)\boldsymbol{x} = A^T\boldsymbol{b} \quad \text{(Normal Equation)}$$

If $\left(A^T A\right)^{-1}$ exists, then $\boldsymbol{x} = \left(A^T A\right)^{-1} A^T\boldsymbol{b}$.

Otherwise, solve the normal equation, e.g., Cholesky factorization method for $A^T A$ being positive semi-definite.

Least squares optimization model

- Case 2. When $m > n$, too many equations to satisfy may lead to infeasibility of the system!
- Task: Find an approximate solution of the system

Question: How to define an approximate solution?

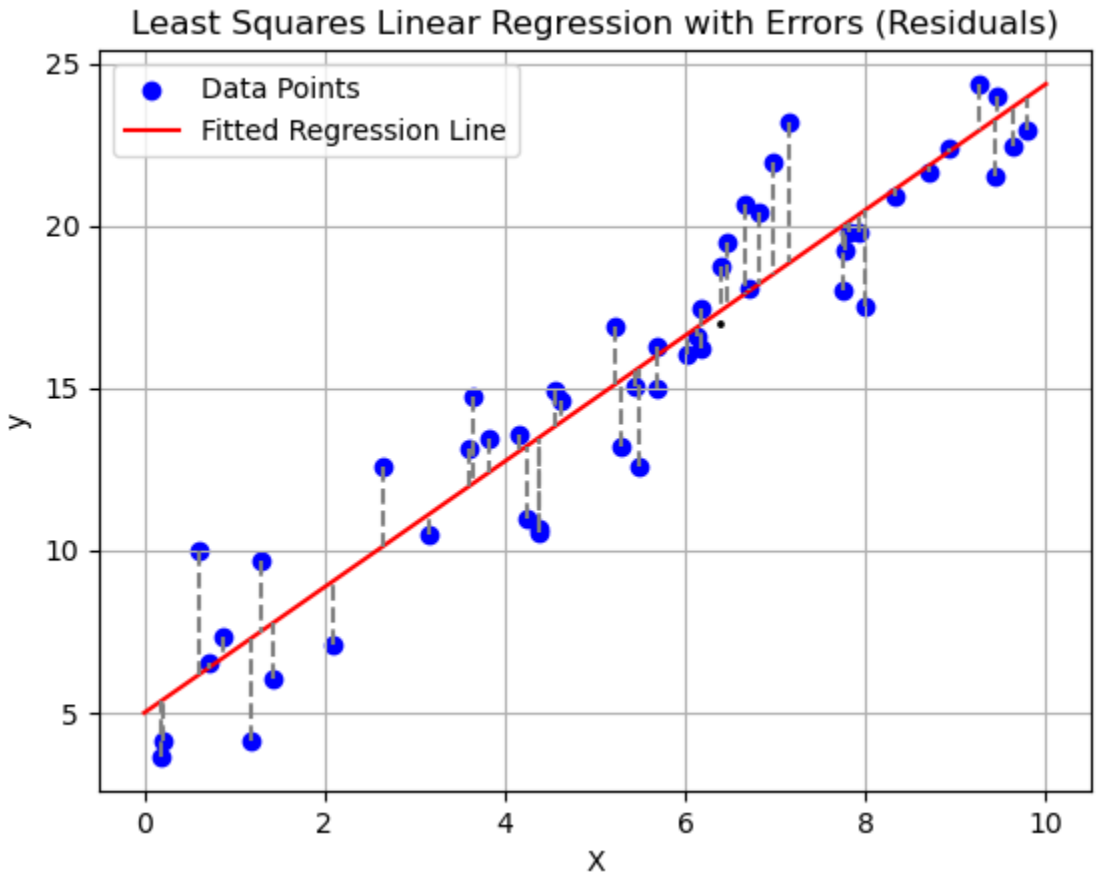
Find $\boldsymbol{y} \in \mathbb{R}^n$ such that

$$A\boldsymbol{y} - \boldsymbol{b} \approx \mathbf{0}$$

- Fact:

Different interpretations of "$\approx$" lead to different models!

Linear regression



https://medium.com/physics-and-machine-learning/misconceptions-about-least-square-regression-1131841d240f

Least squares linear regression

- Linear regression model

Dataset: $S = \{(\boldsymbol{x}^i, y^i): i = 1, \dots, N\}$ for $\boldsymbol{x}^i \in \mathbb{R}^n$ and $y^i \in \mathbb{R}$.

Assumption:

- there is a hidden linear relationship between the input variables and output variables over S

Task: From the N observations (samples), find a linear hyperplane $y = \boldsymbol{m}^T \boldsymbol{x} + b$ ($\boldsymbol{m} \in \mathbb{R}^n, b \in \mathbb{R}$) in the $(n+1)$-dimensional space of $(\boldsymbol{x}, y)$, such that

$$y^i \approx \widehat{y^i} = \boldsymbol{m}^T \boldsymbol{x}^i + b, i = 1, \dots, N$$

Least squares linear regression

- Hidden linear relationship

$$y^i \approx \widehat{y^i} = \boldsymbol{m}^T \boldsymbol{x}^i + b, i = 1, \dots, N$$

- Least squares of errors:

1. define errors $e_i = |\, y^i - \widehat{y^i}|$ for $i = 1, \dots, N$

2. minimize the squares of errors $\sum_{i=1}^{N} \left(y^i - \widehat{y^i} \right)^2$

- Least squares optimization model

$$\text{Minimize } \sum_{i=1}^{N} \left(\boldsymbol{m}^T \boldsymbol{x}^i + b - y^i \right)^2$$

$$\text{s.t. } \boldsymbol{m} \in \mathbb{R}^n, b \in \mathbb{R}$$

Model reformulation

- Optimization model :

$$\text{Minimize } \sum_{i=1}^{N}\left(\boldsymbol{m}^T\boldsymbol{x}^i + b - y^i\right)^2$$
$$\text{s.t. } \quad \boldsymbol{m} \in \mathbb{R}^n, b \in \mathbb{R}$$

- Arrange data:

$$A = \begin{pmatrix} (\boldsymbol{x}^1)^T & 1 \\ \vdots & \vdots \\ (\boldsymbol{x}^N)^T & 1 \end{pmatrix} \in M_{N\times(n+1)}, z = \begin{pmatrix} \boldsymbol{m} \\ b \end{pmatrix} \in \mathbb{R}^{n+1}, \boldsymbol{y} = \begin{pmatrix} y^1 \\ \vdots \\ y^N \end{pmatrix} \in \mathbb{R}^N$$

- Reformulated optimization model :

$$\text{Minimize } \|A\boldsymbol{z} - \boldsymbol{y}\|_2^2$$
$$\text{s.t. } \quad \mathbf{z} \in \mathbb{R}^{n+1}$$

A is given by input data, $\boldsymbol{y}$ is given by output data, and z^* determines the hidden linear relationship

Least squares linear regression

- Optimization model:

$$\text{Minimize } \|A\boldsymbol{z} - \boldsymbol{y}\|_2^2$$
$$\text{s.t.} \quad \boldsymbol{z} \in \mathbb{R}^{n+1}$$

- Solution

$$\nabla \|A\boldsymbol{z} - \boldsymbol{y}\|_2^2 = 0 \Leftrightarrow 2\left(A^T A\right)\boldsymbol{z} - 2A^T\boldsymbol{y} = 0$$
$$\Leftrightarrow \left(A^T A\right)\boldsymbol{z} = A^T\boldsymbol{y} \quad \text{(Normal Equation)}$$

$\boldsymbol{z}^*$ is a solution to a linear system of $n + 1$ variables, which is independent of the number of data points N!

Data manipulation – average effect

Data manipulation: Divide by N on both sides for average effect:

$$\frac{1}{N}A^TA = \frac{1}{N}\begin{pmatrix} x^1 & \cdots & x^N \\ 1 & \cdots & 1 \end{pmatrix}\begin{pmatrix} (x^1)^T & \frac{1}{\vdots} \\ \vdots & \vdots \\ (x^N)^T & 1 \end{pmatrix} = \frac{1}{N}\begin{pmatrix} \sum_{i=1}^N x^i(x^i)^T & \sum_{i=1}^N x^i \\ \sum_{i=1}^N (x^i)^T & N \end{pmatrix}$$

$$\frac{1}{N}A^Ty = \frac{1}{N}\begin{pmatrix} x^1 & \cdots & x^N \\ 1 & \cdots & 1 \end{pmatrix}y = \frac{1}{N}\begin{pmatrix} \sum_{i=1}^N y^i x_1^i \\ \vdots \\ \sum_{i=1}^N y^i x_n^i \\ \sum_{i=1}^N y^i \end{pmatrix}$$

1. The linear relationship $\begin{pmatrix} m \\ b \end{pmatrix}$ of the input vector x and output variable y is determined by the average behavior of x^i, $x^i(x^i)^T$, y^i and $y^i x^i$.
2. The dimensionality of the underlying problem depends solely on the number of features/attributes of the input and output variables $(n+1)$.
   It is independent of the sample size (N).

Discussion and extensions

- Other models for linear regression?
  - Least squares optimization model is commonly used
  - Support vector regression (SVR) model to be introduced
- Nonlinear regression?
  - increasing dimensionality for nonlinearity

Example: $y = ax^2 + bx + c$

$quadratic\ \ relation$ between $input\ x \in \mathbb{R}$ and $output\ y$

$$y = (a, b, c)\begin{pmatrix} u_1 \\ u_2 \\ 1 \end{pmatrix} \text{ where } u_1 = x^2, u_2 = x$$

$linear\ relation$ between $input\ \boldsymbol{u} \in \mathbb{R}^{\mathbf{2}}$ and $output\ y$!

Data regression - Support vector regression

- https://medium.com/analytics-vidhya/support-vector-regression-svr-model-a-regression-based-machine-learning-approach-f4641670c5bb