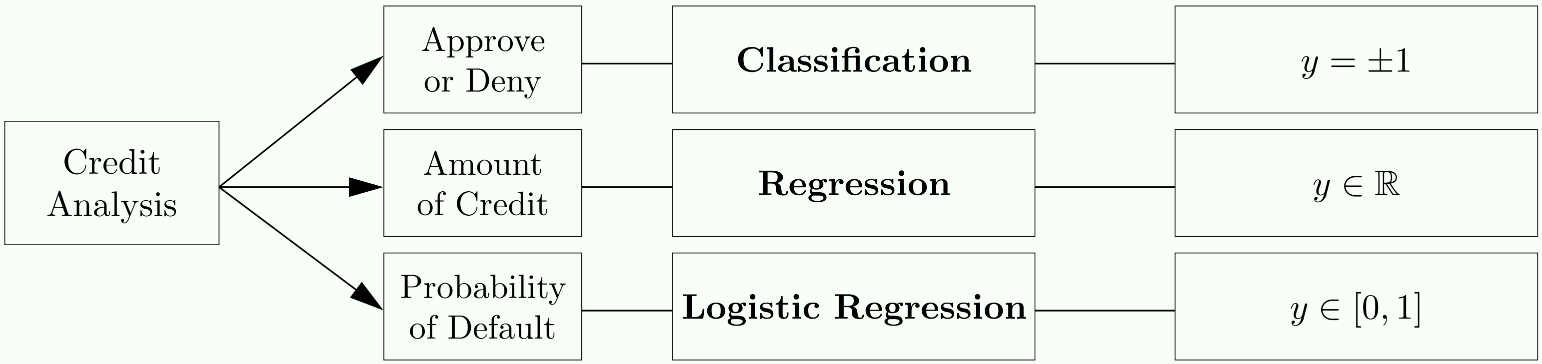
Learning From Data
Lecture 8
Linear Classification and Regression

Linear Classification
Linear Regression

M. Magdon-Ismail
CSCI 4100/6100

Three Learning Problems



- Linear models are perhaps ***the*** fundamental model.

- The linear model is the first model to try.

The Linear Signal

linear in $\mathbf{x}$: gives the line/hyperplane separator

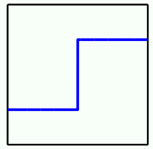
$$s = \mathbf{w}^{\mathrm{T}}\mathbf{x}$$

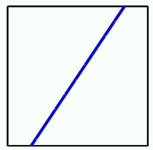
linear in $\mathbf{w}$: makes the algorithms work

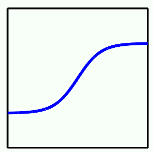
$\mathbf{x}$ is the augmented vector: $\mathbf{x} \in \{1\} \times \mathbb{R}^d$

The Linear Signal

$s = \mathbf{w}^{\mathrm{T}}\mathbf{x} \longrightarrow$

 $\rightarrow \mathrm{sign}(\mathbf{w}^{\mathrm{T}}\mathbf{x})$ $\{-1, +1\}$

 $\rightarrow \mathbf{w}^{\mathrm{T}}\mathbf{x}$ $\mathbb{R}$

 $\rightarrow \theta(\mathbf{w}^{\mathrm{T}}\mathbf{x})$ $[0, 1]$

$$\boldsymbol{y = \theta(s)}$$

Linear Regression

age	32 years
gender	male
salary	40,000
debt	26,000
years in job	1 year
years at home	3 years
...	...

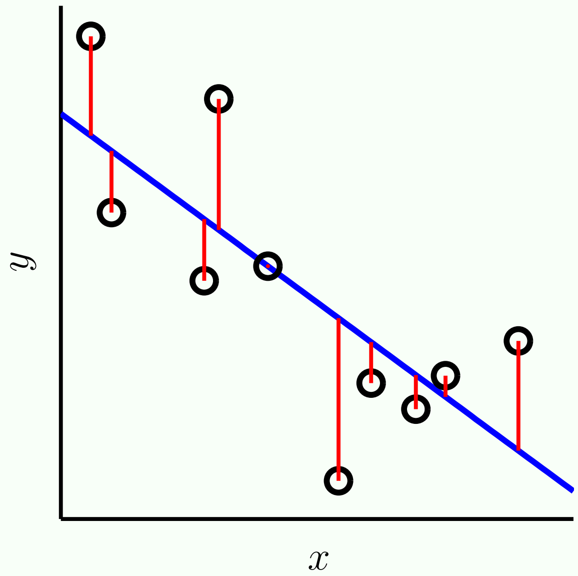
Classification: Approve/Deny

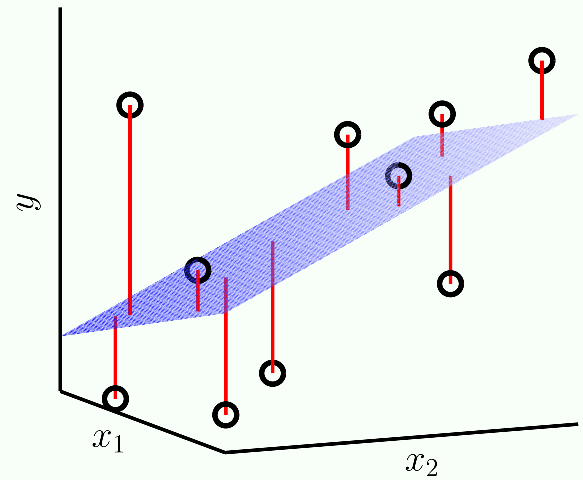
Regression: Credit Line (dollar amount)

regression $\equiv y \in \mathbb{R}$

$$h(\mathbf{x}) = \sum_{i=0}^{d} w_i x_i = \mathbf{w}^{\mathrm{T}}\mathbf{x}$$

Least Squares Linear Regression





$$y = f(\mathbf{x}) + \epsilon \qquad \longleftarrow \quad \text{noisy target } P(y|\mathbf{x})$$

$$\left.\begin{array}{ll} \text{in-sample error} & E_{\text{in}}(h) = \frac{1}{N}\sum_{n=1}^{N}(h(\mathbf{x}_n) - y_n)^2 \\[2em] \text{out-of-sample error} & E_{\text{out}}(h) = \mathbb{E}_{\mathbf{x}}[(h(\mathbf{x}) - y)^2] \end{array}\right\} \quad h(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\mathbf{x}$$

Using Matrices for Linear Regression

$$\underbrace{\mathrm{X} = \begin{bmatrix} \text{---}\mathbf{x}_1\text{---} \\ \text{---}\mathbf{x}_2\text{---} \\ \vdots \\ \text{---}\mathbf{x}_N\text{---} \end{bmatrix}}_{\text{data matrix, } N \times (d+1)} \qquad \underbrace{\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{\text{target vector}} \qquad \underbrace{\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_N \end{bmatrix} = \begin{bmatrix} \mathbf{w}^{\mathrm{T}}\mathbf{x}_1 \\ \mathbf{w}^{\mathrm{T}}\mathbf{x}_2 \\ \vdots \\ \mathbf{w}^{\mathrm{T}}\mathbf{x}_N \end{bmatrix} = \mathrm{X}\mathbf{w}}_{\text{in-sample predictions}}$$

$$\begin{aligned} E_{\text{in}}(\mathbf{w}) &= \frac{1}{N}\sum_{n=1}^{N}(\hat{y}_n - y_n)^2 \\ &= \tfrac{1}{N}\| \, \hat{\mathbf{y}} - \mathbf{y} \, \|_2^2 \\ &= \tfrac{1}{N}\| \, \mathrm{X}\mathbf{w} - \mathbf{y} \, \|_2^2 \\ &= \tfrac{1}{N}(\mathbf{w}^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{X}\mathbf{w} - 2\mathbf{w}^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{y}) \end{aligned}$$

Linear Regression Solution

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{N}\left(\mathbf{w}^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathrm{X}\mathbf{w} - 2\mathbf{w}^{\mathrm{T}}\mathrm{X}^{\mathrm{T}}\mathbf{y} + \mathbf{y}^{\mathrm{T}}\mathbf{y}\right)$$

Vector Calculus: To minimize $E_{\text{in}}(\mathbf{w})$, set $\nabla_{\mathbf{w}} E_{\text{in}}(\mathbf{w}) = \mathbf{0}$.

$$\nabla_{\mathbf{w}}(\mathbf{w}^{\mathrm{T}}\mathrm{A}\mathbf{w}) = (\mathrm{A} + \mathrm{A}^{\mathrm{T}})\mathbf{w}, \qquad \nabla_{\mathbf{w}}(\mathbf{w}^{\mathrm{T}}\mathbf{b}) = \mathbf{b}.$$

$\mathrm{A} = \mathrm{X}^{\mathrm{T}}\mathrm{X}$ and $\mathbf{b} = \mathrm{X}^{\mathrm{T}}\mathbf{y}$:

$$\nabla_{\mathbf{w}} E_{\text{in}}(\mathbf{w}) = \frac{2}{N}(\mathrm{X}^{\mathrm{T}}\mathrm{X}\mathbf{w} - \mathrm{X}^{\mathrm{T}}\mathbf{y})$$

Setting $\nabla E_{\text{in}}(\mathbf{w}) = \mathbf{0}$:

$$\mathrm{X}^{\mathrm{T}}\mathrm{X}\mathbf{w} = \mathrm{X}^{\mathrm{T}}\mathbf{y} \qquad \longleftarrow \text{ normal equations}$$

$$\mathbf{w}_{\text{lin}} = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathbf{y} \qquad \longleftarrow \text{ when } \mathrm{X}^{\mathrm{T}}\mathrm{X} \text{ is invertible}$$

Linear Regression Algorithm

Linear Regression Algorithm:

1. Construct the matrix X and the vector $\mathbf{y}$ from the data set $(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_N, y_N)$, where each $\mathbf{x}$ includes the $x_0 = 1$ coordinate,

$$\underbrace{\mathrm{X} = \begin{bmatrix} —\mathbf{x}_1— \\ —\mathbf{x}_2— \\ \vdots \\ —\mathbf{x}_N— \end{bmatrix}}_{\text{data matrix}}, \qquad \underbrace{\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix}}_{\text{target vector}}.$$

2. Compute the pseudo inverse $\mathrm{X}^\dagger$ of the matrix X. If $\mathrm{X}^{\mathrm{T}}\mathrm{X}$ is invertible,

$$\mathrm{X}^\dagger = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}$$

3. Return $\mathbf{w}_{\text{lin}} = \mathrm{X}^\dagger\mathbf{y}$.