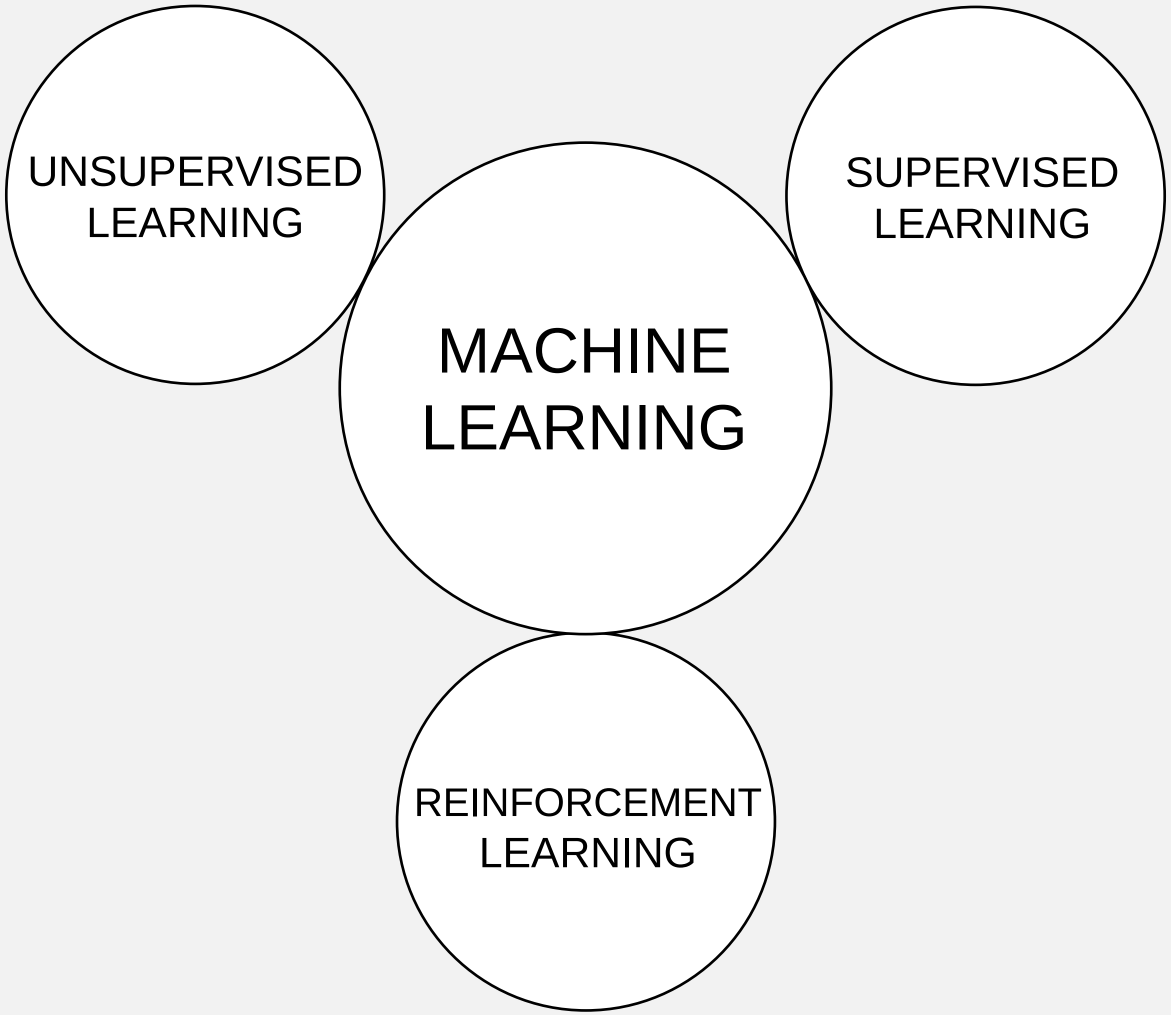
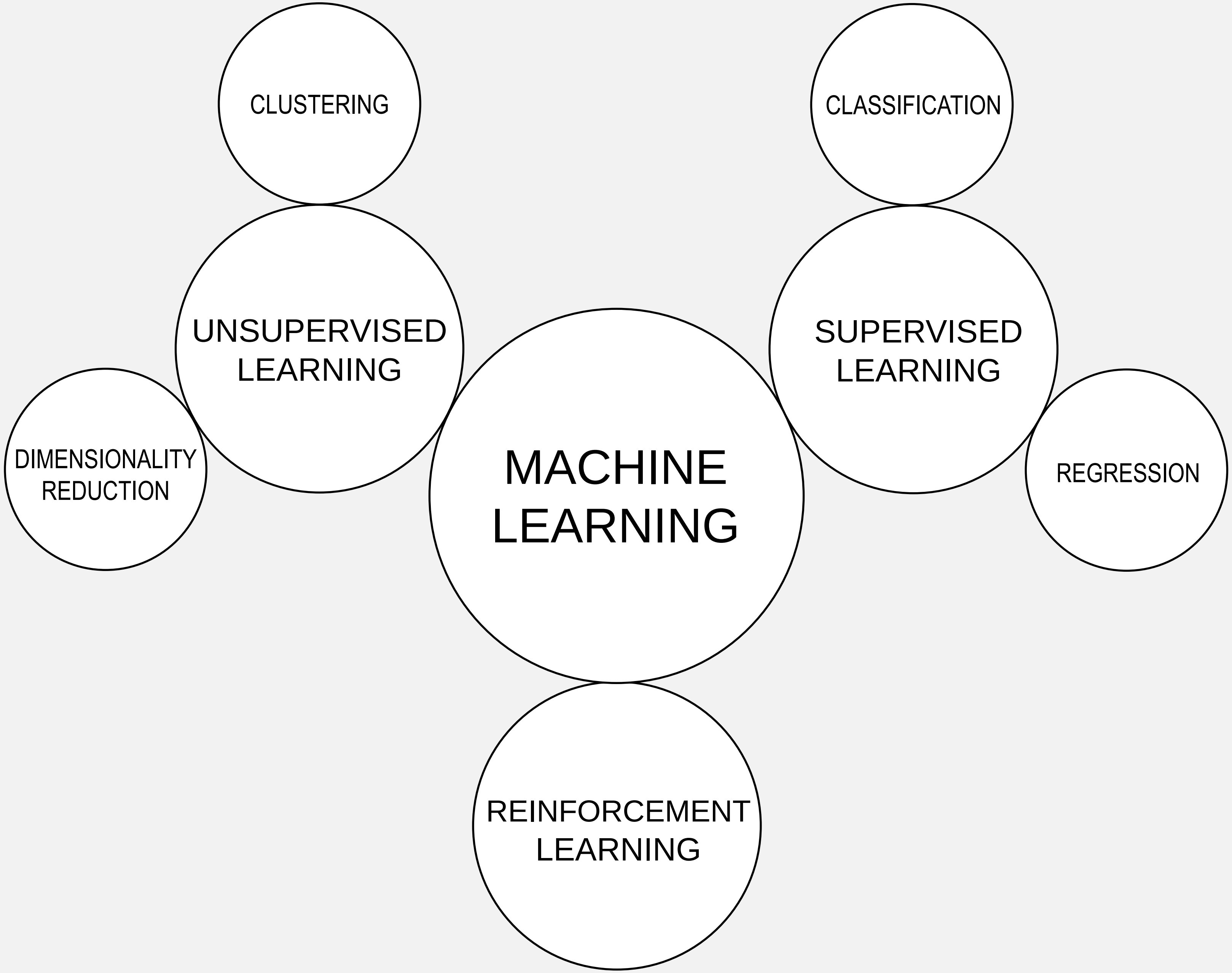
REGRESSION

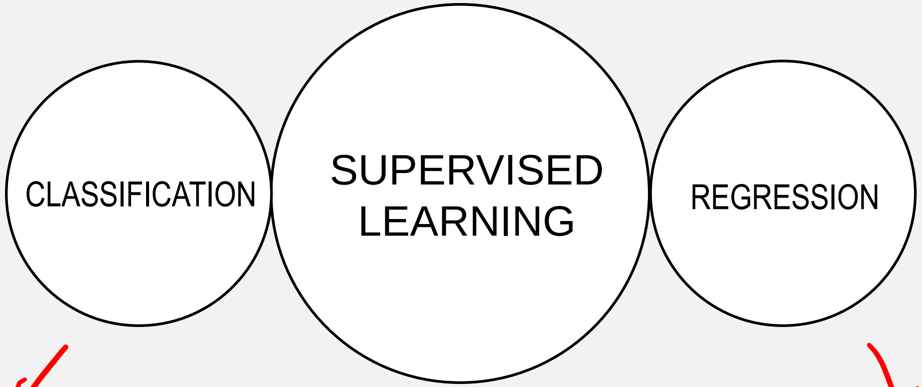
Lecture 8

MAL1, 2024

MACHINE
LEARNING


UNSUPERVISED LEARNING
SUPERVISED LEARNING
MACHINE LEARNING
REINFORCEMENT LEARNING


CLUSTERING
CLASSIFICATION
UNSUPERVISED LEARNING
SUPERVISED LEARNING
DIMENSIONALITY REDUCTION
MACHINE LEARNING
REGRESSION
REINFORCEMENT LEARNING



response variable is a class

response variable is a number

REGRESSION

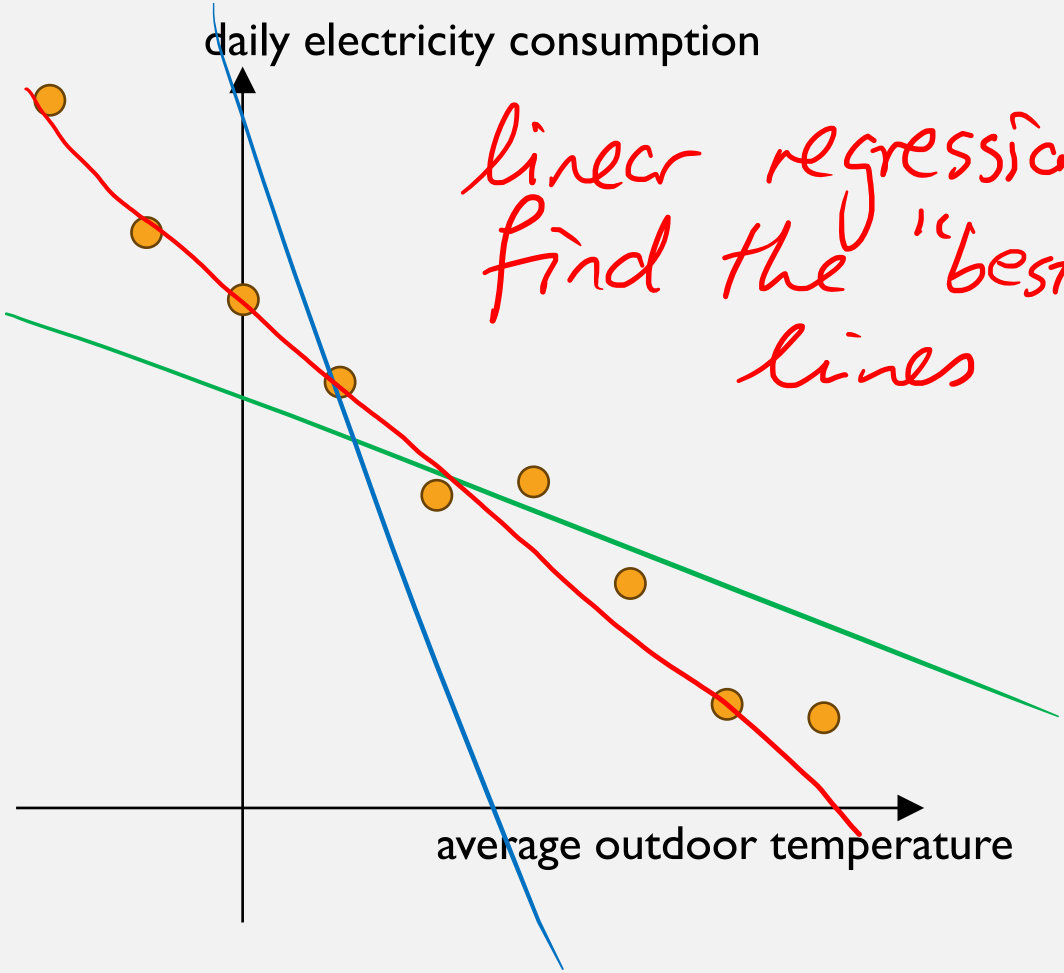
- Linear regression
- Performance metrics
- Polynomial regression
- Regularization

REGRESSION

average outdoor temperature (°C)	daily electricity consumption (kWh)
-10	46.5
-5	37.9
0	33.2
5	27.5
10	20.3
15	21.1
20	14.2
25	6.3
30	5.6

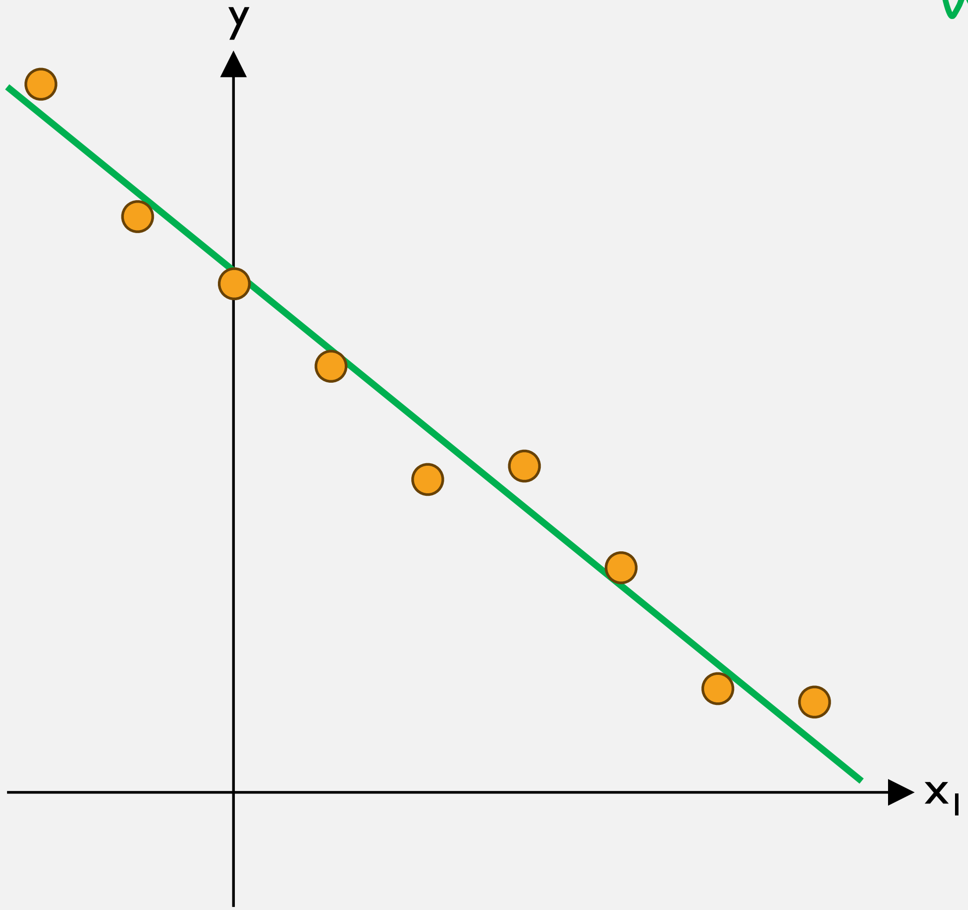
feature

response variable



REGRESSION

x_1	y
$\theta_0 + {-10} \times \theta_1 \approx$	46.5
$\theta_0 + {-5} \times \theta_1 \approx$	37.9
$\theta_0 + 0 \times \theta_1 \approx$	33.2
$\theta_0 + 5 \times \theta_1 \approx$	27.5
$\theta_0 + 10 \times \theta_1 \approx$	20.3
$\theta_0 + 15 \times \theta_1 \approx$	21.1
$\theta_0 + 20 \times \theta_1 \approx$	14.2
$\theta_0 + 25 \times \theta_1 \approx$	6.3
$\theta_0 + 30 \times \theta_1 \approx$	5.6



with $\hat{y} = \theta_0 + \theta_1 x_1$, find the "best" values of θ_0 and θ_1

$$\underbrace{\theta_0 = 33.72 \qquad \theta_1 = -1.009}_{\text{finding these numbers}}$$

= training the model

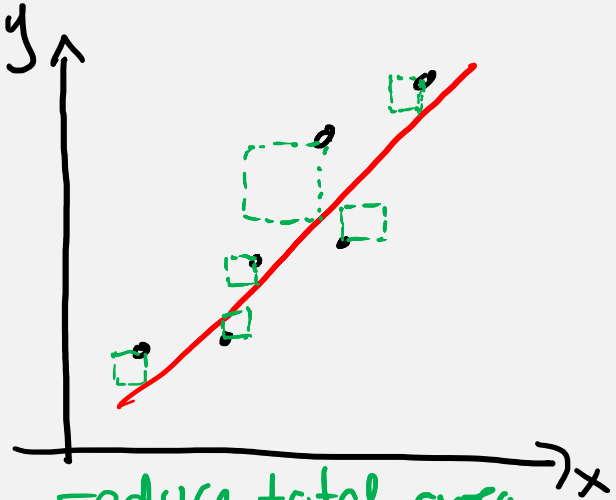
FINDING THETA'S

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \dots = \underbrace{\theta^T x}_{\text{matrix notation}}$$

prediction

features

The "best" line reduces/minimizes



reduce total area

$$SSE = \sum_i \left(y^{(i)} - \hat{y}^{(i)}\right)^2$$

sum of squared errors

observation

prediction

FINDING THETA'S

$$\text{SSE} = \sum_i \left(y^{(i)} - \hat{y}^{(i)}\right)^2 = \sum_i \left(y^{(i)} - \theta^T x^{(i)}\right)^2$$

minimize → take the derivative wrt. all θ's and set equal to zero:

$$\frac{\partial}{\partial \theta_j} \text{SSE} = \frac{\partial}{\partial \theta_j} \sum_i \left(y^{(i)} - \theta^T x^{(i)}\right)^2 = 2 \sum_i \left(y^{(i)} - \theta^T x^{(i)}\right) x_j^{(i)} = 0$$

summarize in matrix form:

$$2\boldsymbol{X}^T(\boldsymbol{X\theta} - \boldsymbol{y}) = 0$$

$$X^T X \theta = X^T y$$

$$\theta = (X^T X)^{-1} X^T y$$

"normal equation"

FINDING THETA'S

$$\boldsymbol{\theta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

$$\boldsymbol{X} = \begin{bmatrix} 1 & -10 \\ 1 & -5 \\ 1 & 0 \\ 1 & 5 \\ 1 & 10 \\ 1 & 15 \\ 1 & 20 \\ 1 & 25 \\ 1 & 30 \end{bmatrix} \quad \boldsymbol{y} = \begin{bmatrix} 46.5 \\ 37.9 \\ 33.2 \\ 27.5 \\ 20.3 \\ 21.1 \\ 14.2 \\ 6.3 \\ 5.6 \end{bmatrix}$$

$$\boldsymbol{X}^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -10 & -5 & 0 & 5 & 10 & 15 & 20 & 25 & 30 \end{bmatrix}$$

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{bmatrix} 9 & 90 \\ 90 & 2400 \end{bmatrix}$$

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{bmatrix} \frac{8}{45} & -\frac{1}{150} \\ -\frac{1}{150} & \frac{1}{1500} \end{bmatrix}$$

$$(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \begin{bmatrix} \frac{11}{45} & \frac{19}{90} & \frac{8}{45} & \frac{13}{90} & \frac{1}{9} & \frac{7}{90} & \frac{2}{45} & \frac{1}{90} & -\frac{1}{45} \\ -\frac{1}{75} & -\frac{1}{100} & -\frac{1}{150} & -\frac{1}{300} & 0 & \frac{1}{300} & \frac{1}{150} & \frac{1}{100} & \frac{1}{75} \end{bmatrix}$$

pseudoinverse

$$\boldsymbol{\theta} = \underbrace{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T}_{\text{pseudoinverse}}\boldsymbol{y} = \begin{bmatrix} 33.72 \\ -1.009 \end{bmatrix} \begin{matrix} \leftarrow \theta_0 \text{ intercept} \\ \leftarrow \theta_1 \text{ slope} \end{matrix}$$

X y

x_0	x_1	y
1	-10	46.5
1	-5	37.9
1	0	33.2
1	5	27.5
1	10	20.3
1	15	21.1
1	20	14.2
1	25	6.3
1	30	5.6

$O(n^2)$ instead of $O(n^3)$

in practice, the pseudoinverse is computed using SVD

REGRESSION

- Linear regression
- Performance metrics
- Polynomial regression
- Regularization

SSE AND FRIENDS

- The smaller the SSE, the better the model …

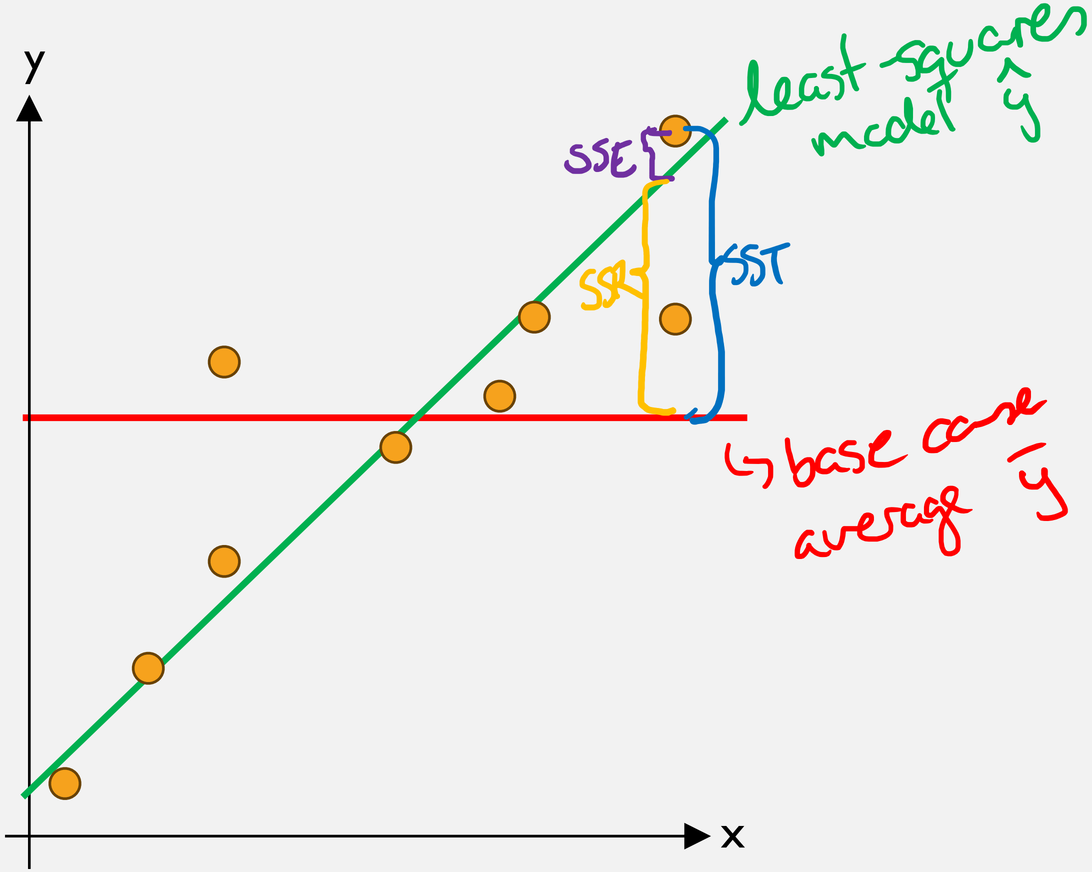
but SSE depends on # of data points

$\rightarrow MSE = \frac{1}{n} SSE$ mean squared error

but MSE depends on scale of response variable

… so what should we use as our performance metric?

SSE AND FRIENDS



total deviation from mean

$$SST = \sum_i (y_i - \bar{y})^2$$

(total sum of squares)

unexplained

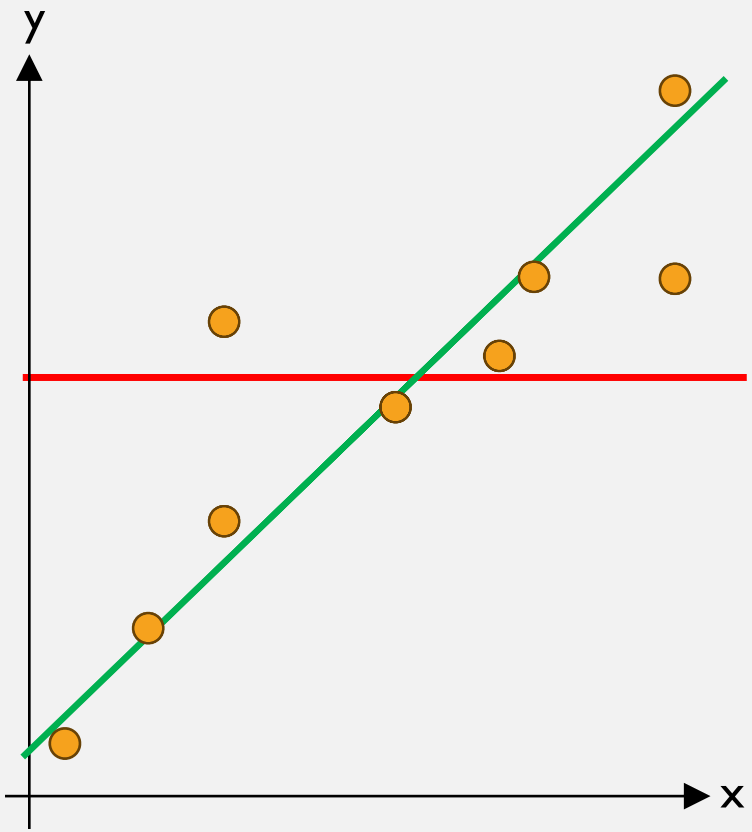
$$SSE = \sum_i (y_i - \hat{y}_i)^2$$

explained

$$SSR = \sum_i (\hat{y}_i - \bar{y})^2$$

(sum of squares due to regression)

SSE AND FRIENDS



$$\mathrm{SST} = \sum_i \left(y^{(i)} - \bar{y}\right)^2$$ total deviation from mean

$$\mathrm{SSE} = \sum_i \left(y^{(i)} - \hat{y}^{(i)}\right)^2$$ unexplained part

$$\mathrm{SSR} = \sum_i \left(\hat{y}^{(i)} - \bar{y}\right)^2$$ explained part

$$r^2 = \frac{SSR}{SST} = \frac{\text{"explained"}}{\text{"total"}}$$

the amount of variance the model is able to explain

$r^2 \leq 1$ – perfect predictive model

CODE EXAMPLE

Jupyter Notebook **Regression - Hitters**

REGRESSION

- Linear regression
- Performance metrics
- Polynomial regression
- Regularization

POLYNOMIAL REGRESSION

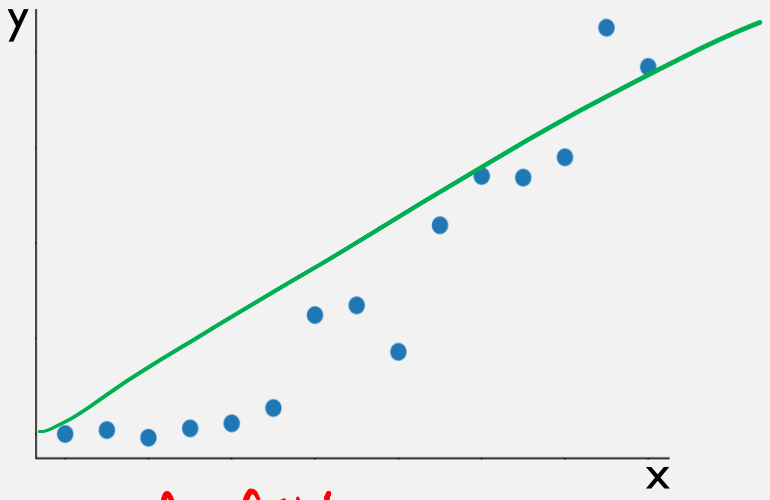
$$\hat{y} = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \dots$$

$x^2, x^3, \dots$ are just new features

$$X = \begin{bmatrix} 1 & x & x^2 & \\ 1 & x & x^2 & \\ 1 & x & x^2 & \dots \\ 1 & x & x^2 & \\ 1 & x & x^2 & \\ 1 & x & x^2 & \\ \vdots & \vdots & \vdots & \end{bmatrix}$$

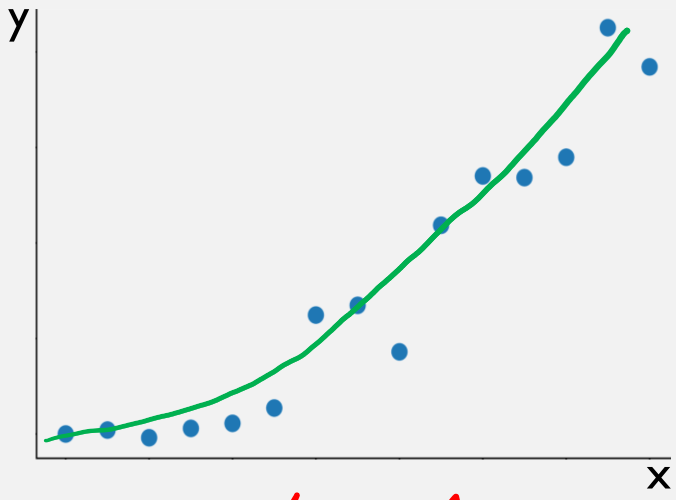
may be a good idea - depending on the underlying relationship

UNDERFITTING AND OVERFITTING



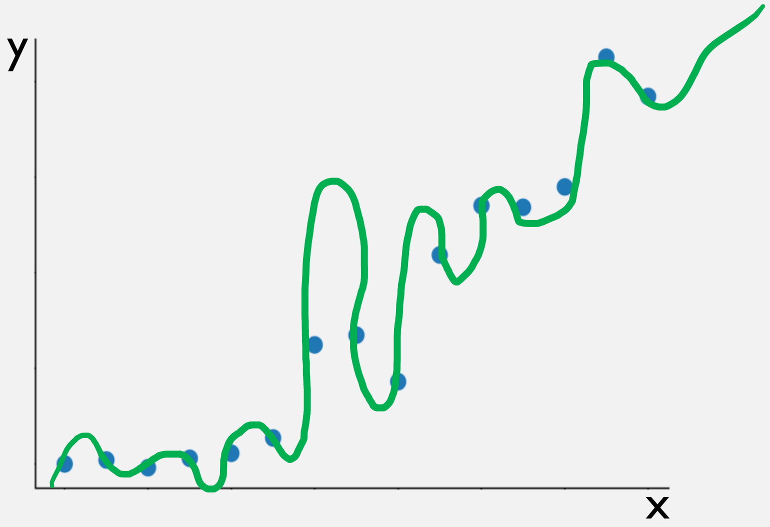
underfitting

high bias
low variance



sweet spot

bias-variance trade-off



overfitting

low bias
high variance

Bias = "Inability to learn from data"
Variance = "Reliance on data"

REGRESSION

- Linear regression
- Performance metrics
- Polynomial regression
- Regularization

REGULARIZATION

Tool to avoid overfitting

Idea: Penalize large coefficient

regularization (hyperparameter)

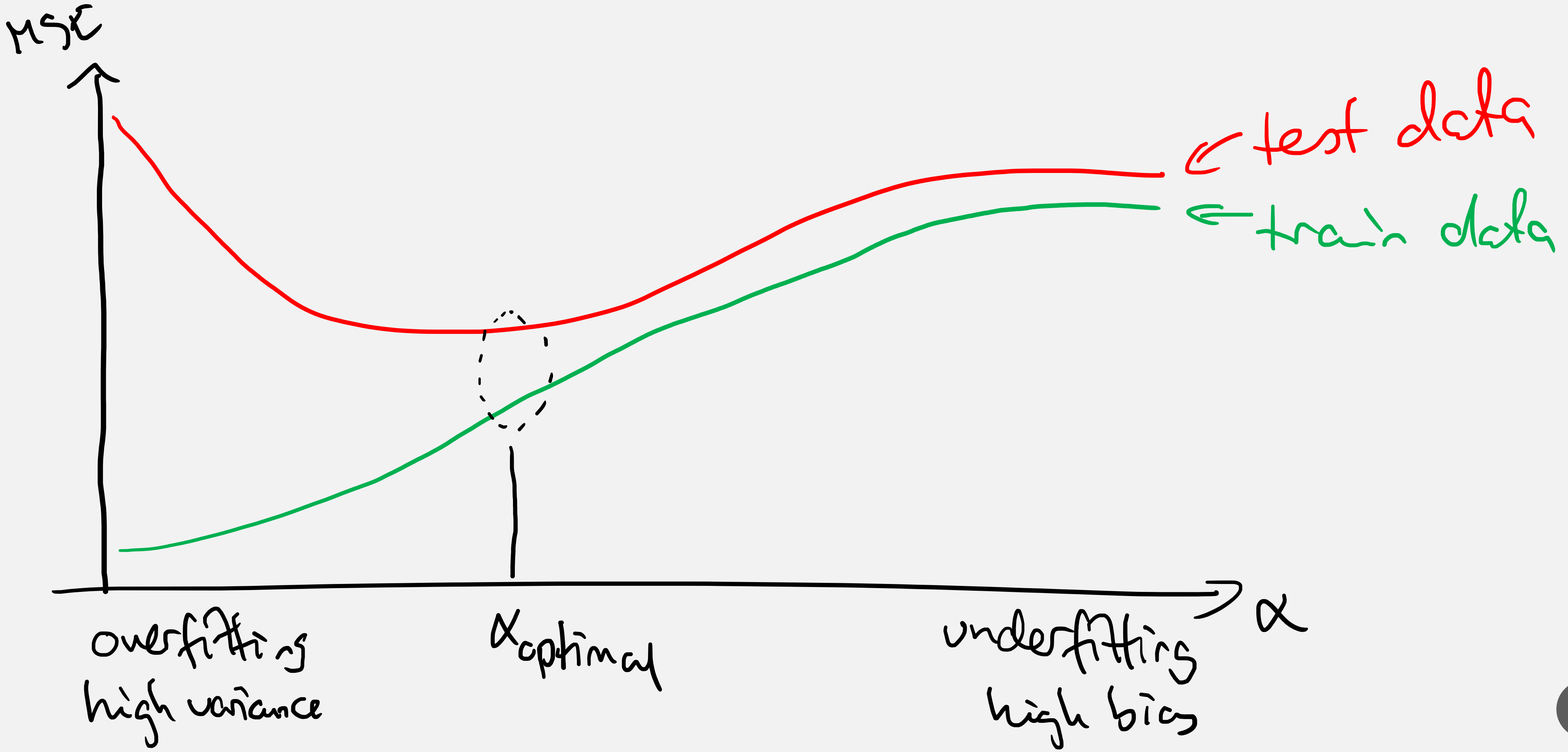
$$\text{loss function} \quad L = MSE + \alpha \underbrace{R(\theta)}_{\text{penalty function of } \theta}$$

common choices

$R(\theta) = \sum_i \theta_i^2$ L_2 regularization → Ridge regression

$R(\theta) = \sum_i |\theta_i|$ L_1 regularization → Lasso regression

THE OPTIMAL REGULARIZATION PARAMETER



RIDGE VS LASSO REGRESSION

Ridge

Drives coefficients to small values overall

Lasso

Drives as many coefficients as possible to zero

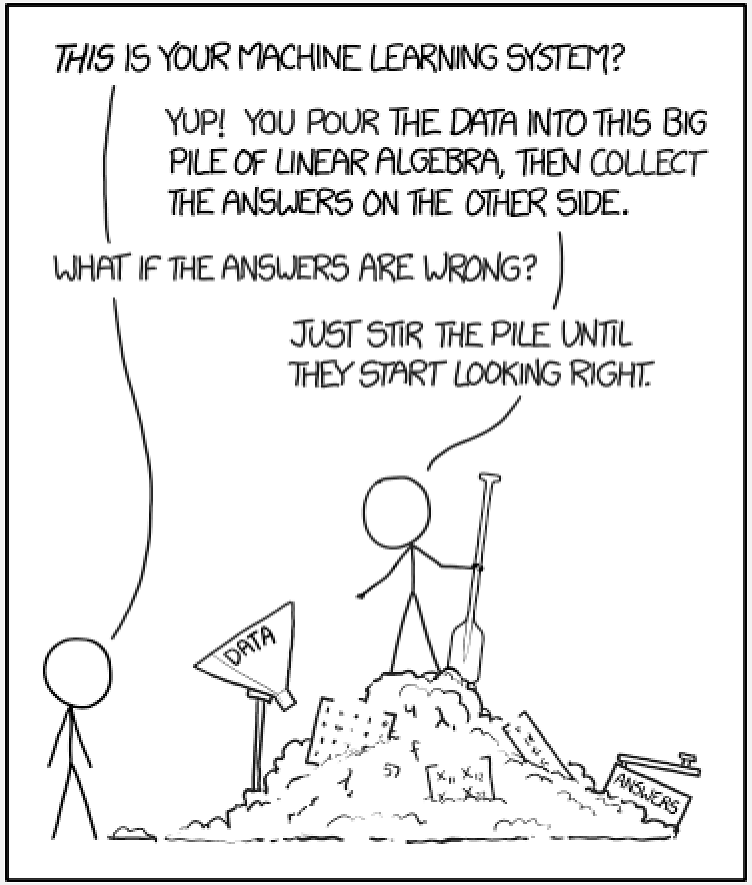
Elastic Net

combination

$$\text{penalty} = \beta \cdot \text{Lasso} + (1-\beta) \cdot \text{Ridge}$$

CODE EXAMPLE

Jupyter Notebook **Regression - Hitters**


THIS IS YOUR MACHINE LEARNING SYSTEM?
YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.
WHAT IF THE ANSWERS ARE WRONG?
JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
DATA
ANSWERS