LING 506 - TOPICS IN COMPUTATIONAL LINGUISTICS

Introductory Machine Learning

Yan Tang

Department of Linguistics, UIUC

Week 12

Last week…

- Regression analysis
- Linear regression
  - A parametric regression technique
- Root Mean Square Error (RMSE) and Mean Square Error (MSE)

Last week…

- Normal Equation

$$\hat{c} = (X^T X)^{-1} \cdot (X^T y)$$

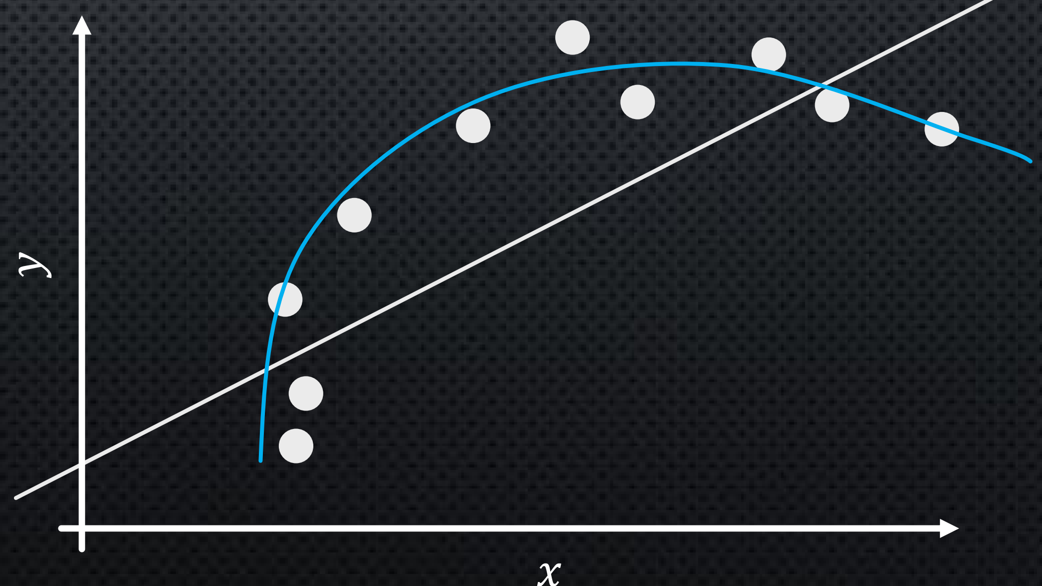
- Gradient Descent (GD)
  - Batch GD
  - Stochastic GD
  - Mini-batch GD

Polynomial regression

- When data is nonlinear, can linear models still be of use?
- Polynomial Regression
  - Adds new features by including powers of each raw feature

$$y = c_0 \cdot x^0 + c_1 \cdot x^1 + c_2 \cdot x^2 + \ldots + c_d \cdot x^d$$

d: the degree of the polynomial



Polynomial regression

- Polynomial regression is also a multi-linear regression model
- The the number of added features can be tremendous!
  - $N = \frac{(n+d)!}{n!d!}$

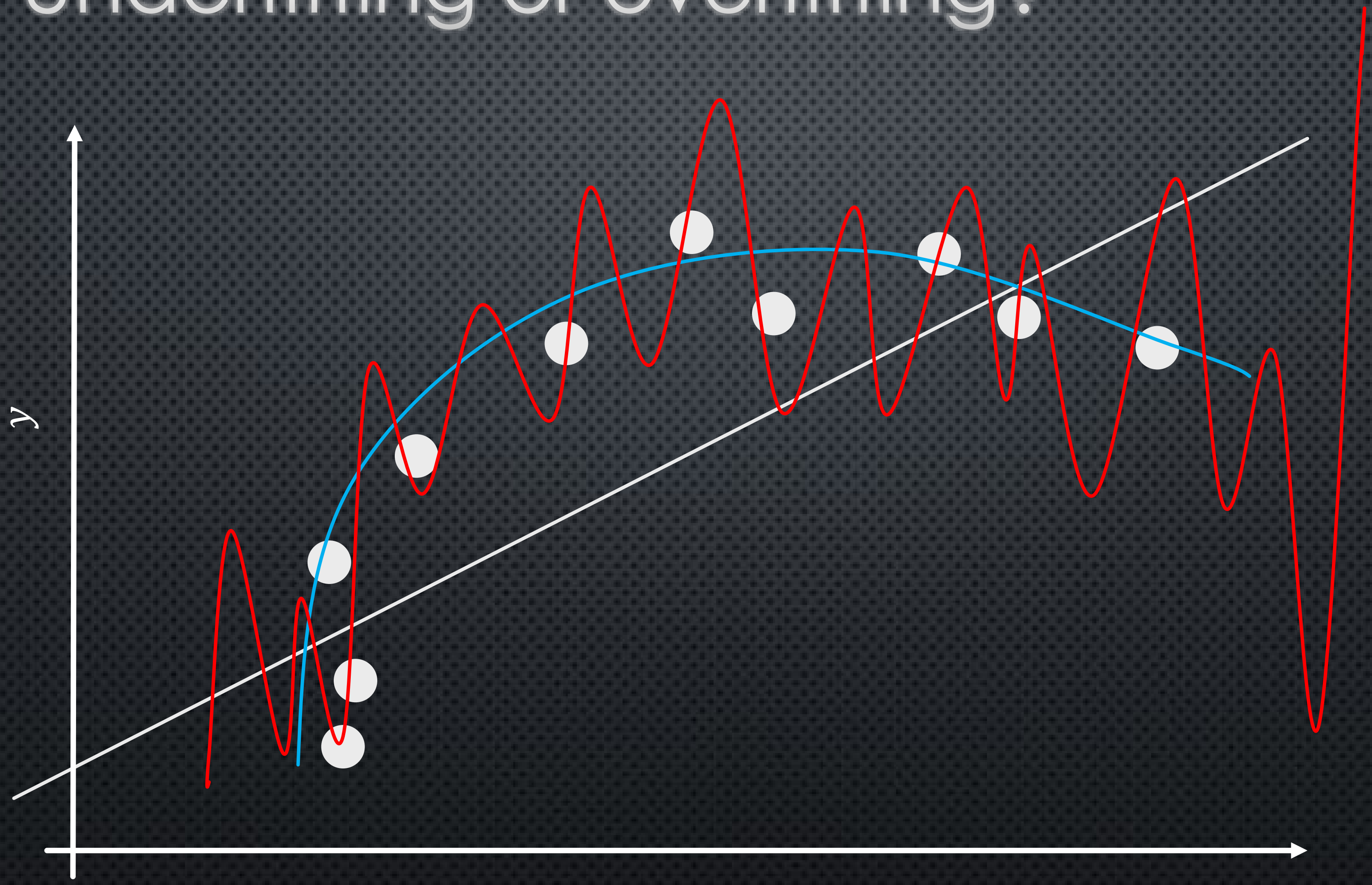
N: the total polynomial features

n: the number of raw features

d: the degree of polynomial

- Excessively high degrees imposed to a large number of features may lead to the “combinatorial explosion” of the number of polynomial features

Underfitting or overfitting?



Learning curves

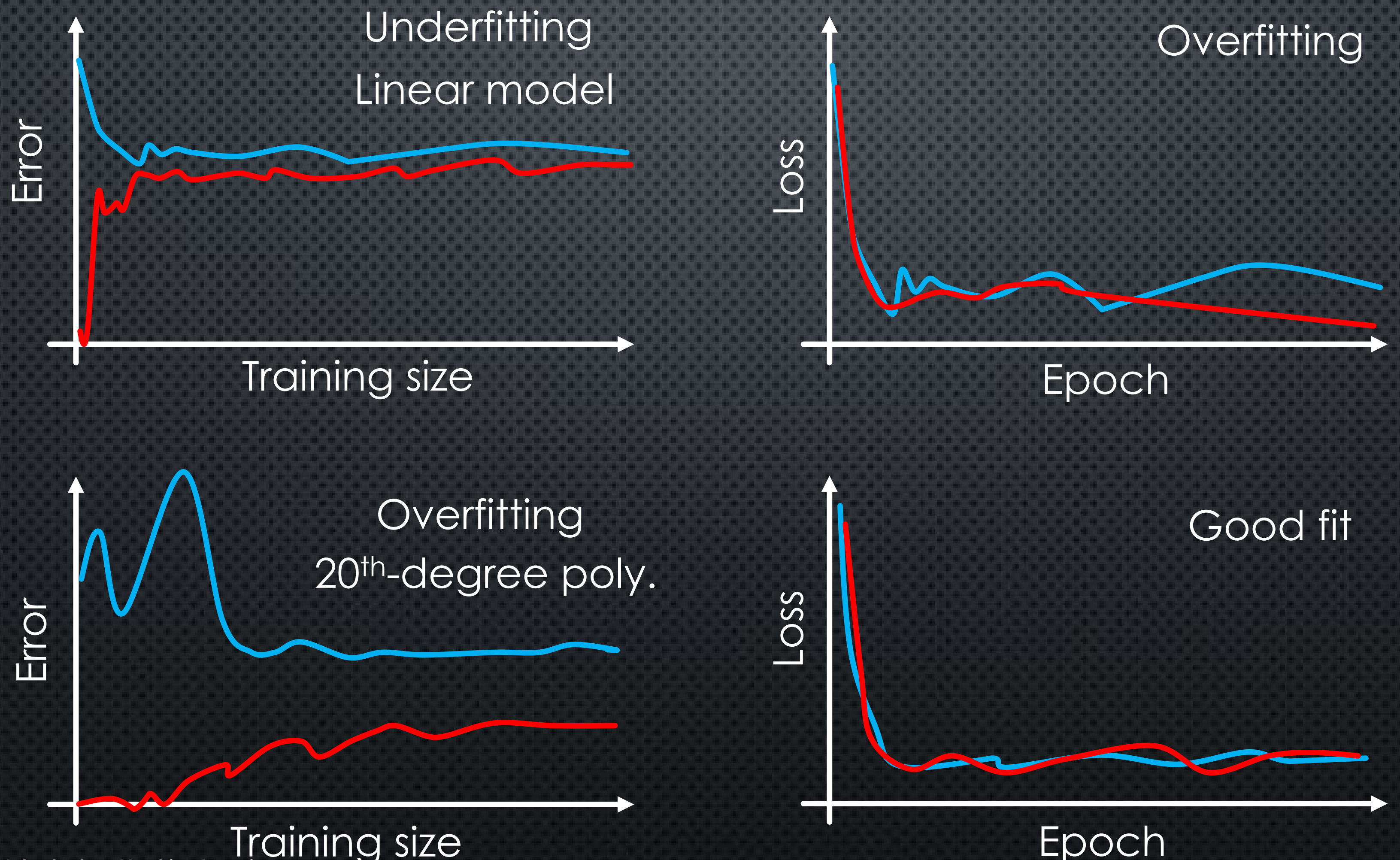
- A learning curve is a plot of model learning performance on the training set and validation over experience or time
  - widely used for algorithms that learn incrementally over samples and time

- The Metrics for evaluation:
  - Classification accuracy being maximising
  - Loss or error being minimising; more common

Learning curves

- Two curves on one plot:
  - **Train Learning Curve**: calculated from the training dataset
  - **Validation Learning Curve**: calculated from a separate validation dataset

Learning curves



Dealing with underfitting and overfitting

- When underfitting
  - Use more complex model
  - Engineering for better feature
  - However, adding more training samples is not helpful
- When overfitting
  - Add more balanced training data
  - Simplify the model structure
- Trade-off between *bias* (simple model) and *variance* (complex model)

Regularised Linear Models

- Regularisation helps reducing overfitting
- A simple regularisation for polynomial regression is to reduce the number of degrees
- For general linear models, regularisation is to constrain the range of the linear coefficients, i.e. the weights of features

Ridge Regression

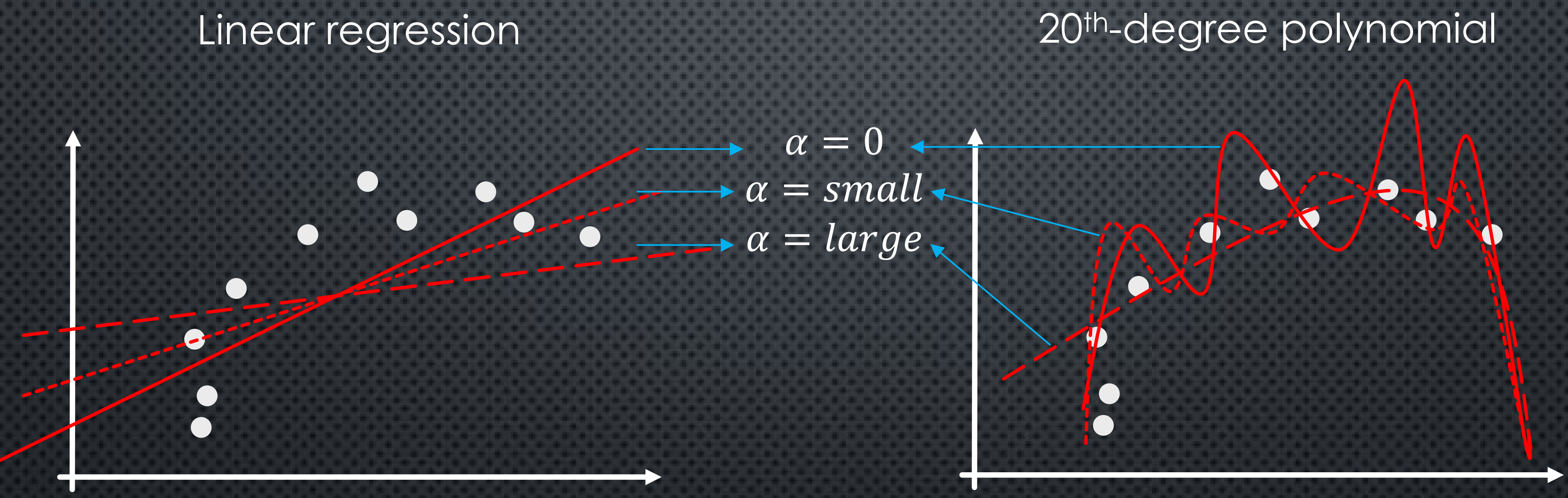
- Ridge Regression (Tikhonov regularisation)
  - A regularised version of Linear Regression
  - Adds a regularisation term to the cost function

$$C = MSE(c) + \boxed{\frac{\alpha}{2}\sum_{i=1}^{n} c_i^2}$$

α :the factor controlling the extent to which the model is regularised

Important: the regularisation term is for training only!

Ridge Regression: examples



Ridge Regression: closed-form and Gradient Decent solutions

- The closed-form solution of Ridge Regression
  - Only two extra terms added to the Normal Equation

$$\hat{c} = (X^T X + \alpha I)^{-1} \cdot (X^T y)$$

I: a $(n+1)\times(n+1)$ identify matrix

- The revised local gradient in Gradient Decent for Ridge Regression:

$$\nabla MSE(c)' = \nabla MSE(c) + \alpha c$$

Lasso Regression

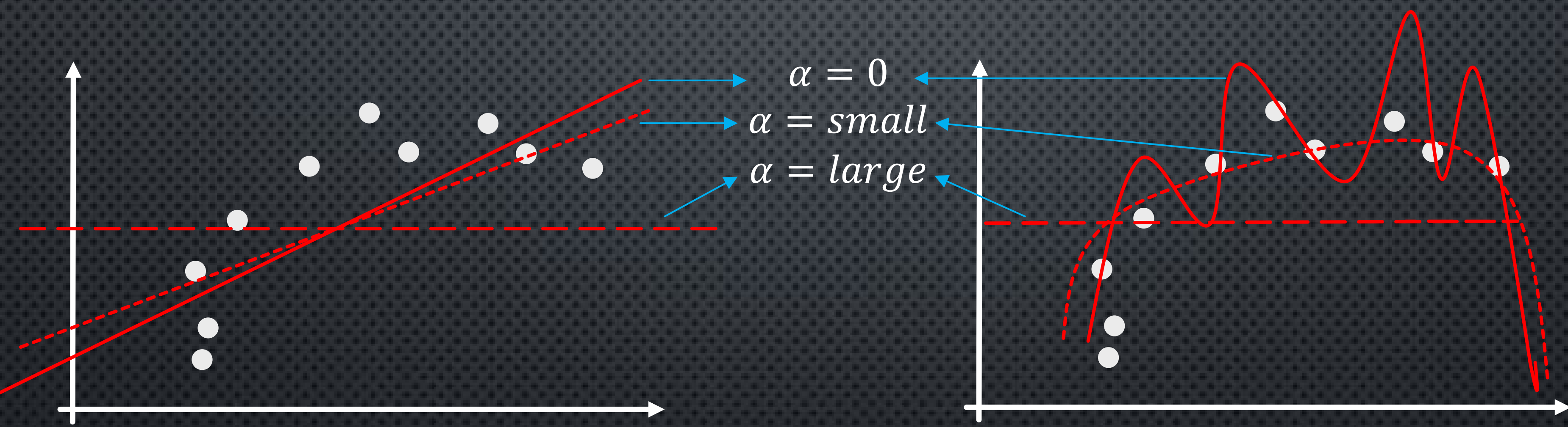
- Least absolute Shrinkage and Selection Operator Regression (Lasso Regression)
  - A regularised version of Linear Regression
  - Adds a regularisation term to the cost function

$$C = MSE(c) + \boxed{\alpha \sum_{i=1}^{n} |c_i|}$$

- Lasso Regression tends to eliminate the coefficients of the less important features

Lasso Regression: examples

Linear regression | 20th-degree polynomial



- The revised local gradient in GD for Lasso Regression

$$\nabla MSE(c)' = \nabla MSE(c) + \alpha \cdot \mathrm{sign}(c), \, where \ \mathrm{sign}(c_i) = \begin{cases} -1 \ if \, c_i < 0 \\ 0 \ if \, c_i = 0 \\ 1 \ if \, c_i > 0 \end{cases}$$

Elastic Net

- A balanced approach between *Ridge Regression* and *Lasso Regression*

$$C = MSE(c) + r\alpha \sum_{i=1}^{n} |c_i| + (1 - r)\frac{\alpha}{2}\sum_{i=1}^{n} c_i^2$$

r: the mix ratio

When $r = 0$, it is Ridge Regression

When $r = 1$, it is Lasso Regression

Ridge, Lasso or Elastic Net?

- Avoid unregularised Linear Regression
- Ridge can be used as a default
- Lasso or Elastic Net for cases where not all the features are important
  - Elastic Net is preferred; a good balance when several features are correlated