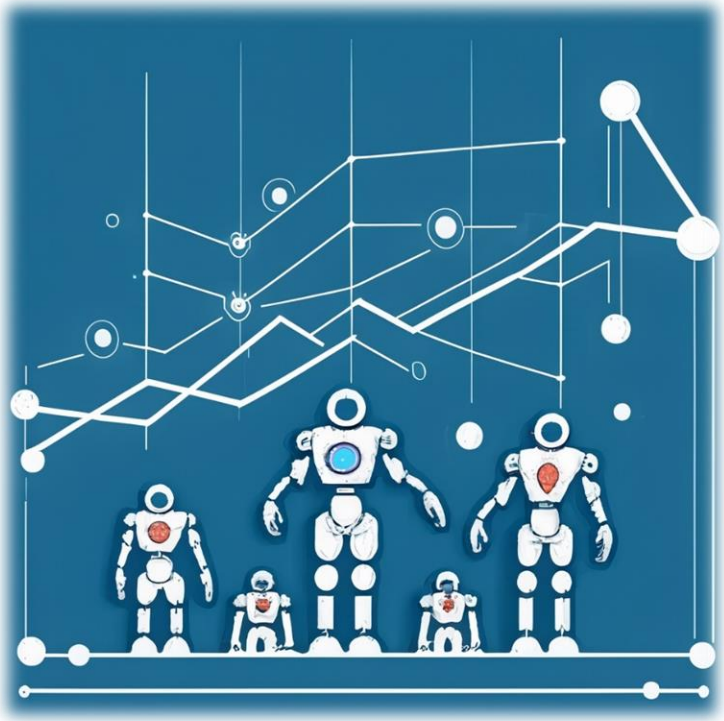
Module 5
Linear Regression (ML approach)



Jon M. Huntsman School of Business
Utah State University

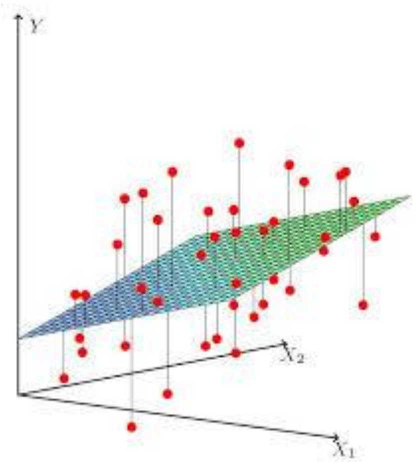
Pedram Jahangiry

Course Modules

- Module 1- Introduction to Deep Learning
- Module 2- Setting up Machine Learning Environment
- Module 3- Linear Regression (Econometrics approach)
- Module 4- Machine Learning Fundamentals
- Module 5- Linear Regression (Machine Learning approach)
- Module 6- Penalized Regression (Ridge, LASSO, Elastic Net)
- Module 7- Logistic Regression
- Module 8- K-Nearest Neighbors (KNN)
- Module 9- Classification and Regression Trees (CART)
- Module 10- Bagging and Boosting
- Module 11- Dimensionality Reduction (PCA)
- Module 12- Clustering (KMeans – Hierarchical)

Topics

- Part I
  - Linear Regression



- Part II
  - Polynomial Regression

Road map

- ML Algorithm
  - Supervised
    - Regression
      - Linear / Polynomial
      - Penalized regression
      - KNN
      - SVR
      - Tree-based Regression models
    - Classification
      - Logistic regression
      - KNN
      - SVC
      - Tree-based Classification models
  - Unsupervised
    - Dimensionality Reduction
      - Principle Component Analysis (PCA)
    - Clustering
      - K-Mean
      - Hierarchical

Linear Regression

Linear Models

The linear model is an important example of a parametric model

- We have a collection of labeled examples $\{(X_i\,, y_i)\}_{i=1}^{N}$, where
  - N is the size of the collection
  - X_i is the D-dimensional feature vector
  - y_i is a real-valued target

$$f_{w,b}(X) = WX + b$$

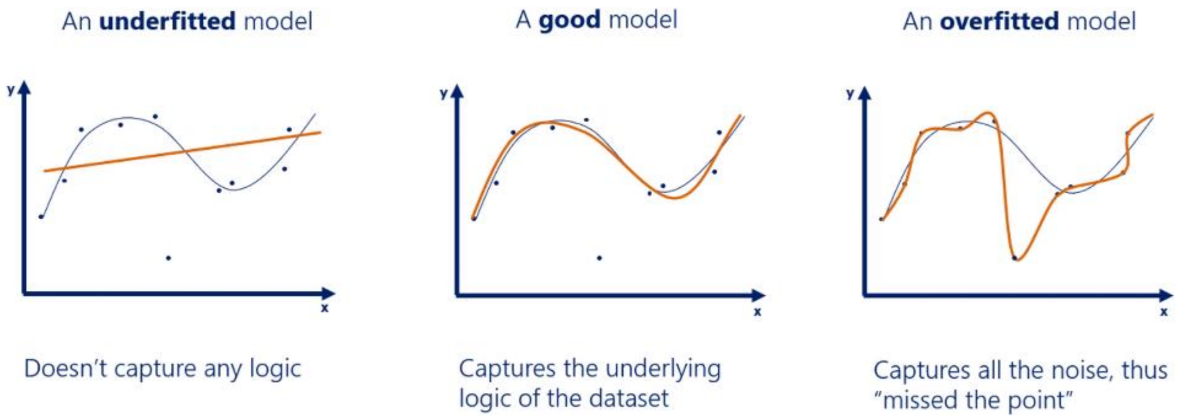
The model $f_{w,b}$ is a linear combination of features and parameterized by W and b

- W is a D-dimensional vector of parameters
- b is a real number

Linear Models (cont'd)

$$f_{w,b}(X) = WX + b$$

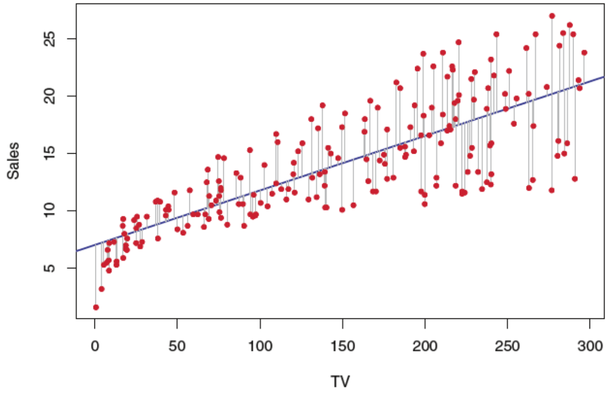
- The model is specified with D+1 parameter.
- We estimate the parameters (W^*, b^*) by fitting the model to training data.
- Although it is almost never correct, a linear model often serves as a simple and interpretable approximation the unknown true $f(X)$.
- It may seem overly simplistic, but linear regression is extremely useful both conceptually and practically.
- Linear regression models rarely overfit.



JON M. HUNTSMAN SCHOOL OF BUSINESS
UtahStateUniversity

Pedram Jahangiry

The optimization problem



The optimization problem is defined as:

$$Min_{w,b}\ MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - f_{w,b}(X_i)\right)^2$$

$\left(y_i - f_{w,b}(X_i)\right)^2$ is called the objective function, or the **loss function**! Or the squared error loss.

Why quadratic loss function? Why not using absolute value or cube?

1. More convenient (well-behaved derivative).
2. There exists a closed form solution.

The solution to this optimization problem is w^* and b^* . Now we can make predictions!

Jon M. Huntsman School of Business
Utah State University

Pedram Jahangiry

Linear Regression Evaluation Metrics

$$R^2 = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{SS_{residuals}}{SS_{total}}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}|$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y})^2}$$

Adjusted

versions

$$Adjusted\ R^2 = 1 - (1 - R^2) * \frac{n - 1}{n - k - 1}$$

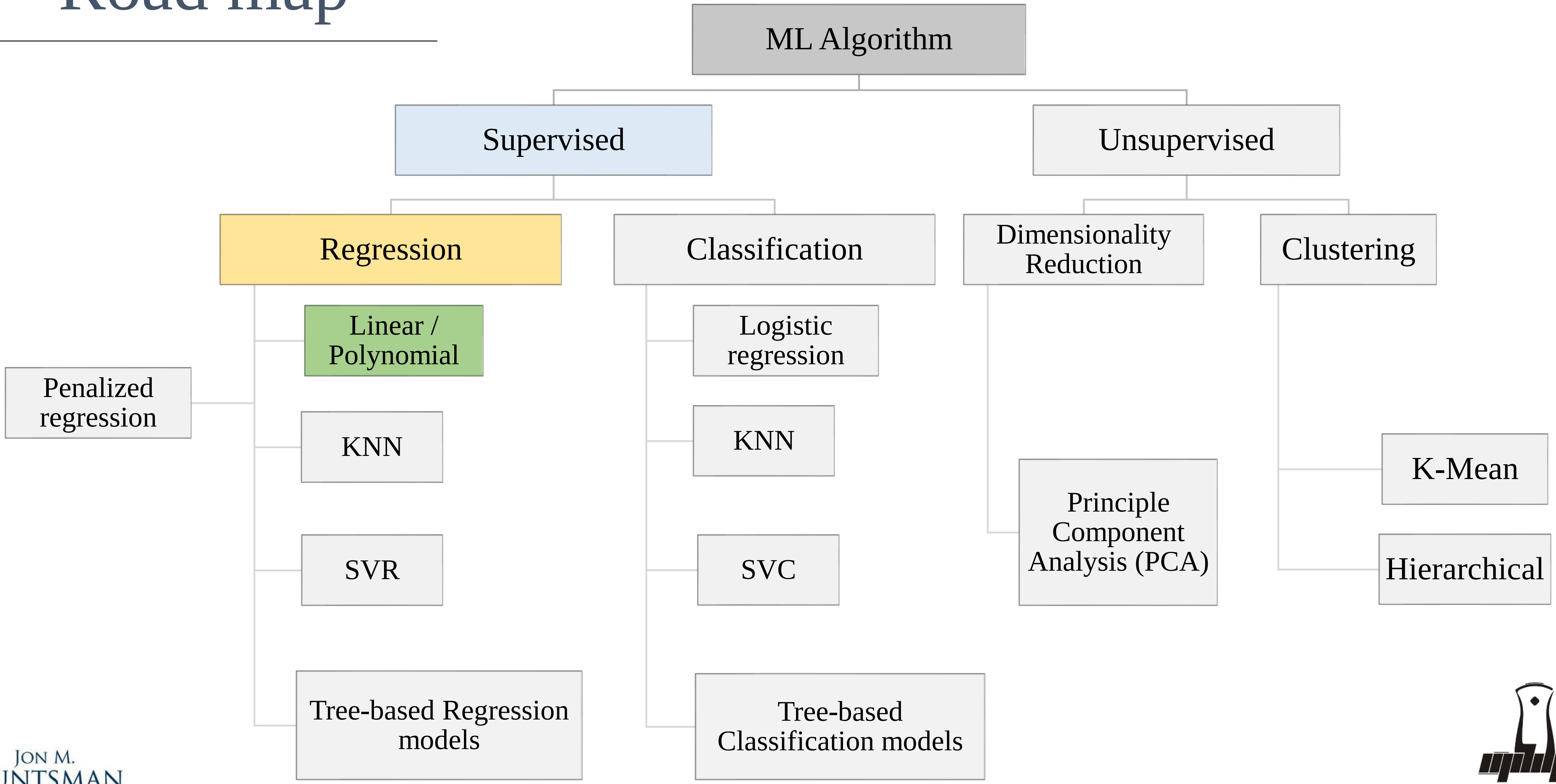
$$AIC = \frac{2K}{N} - \frac{2\ln(\hat{L})}{N}$$

$$BIC = \ln(N)\,K - 2\ln(\hat{L})$$

- **AIC**: *Akaike information criterion*
- **BIC**: *Bayesian information criterion*
- *K:* number of estimated parameters
- $\hat{L}$: Maximum value of the likelihood function

Polynomial Regression

Road map



ML Algorithm
Supervised
Unsupervised
Regression
Classification
Dimensionality Reduction
Clustering
Linear / Polynomial
Penalized regression
KNN
SVR
Tree-based Regression models
Logistic regression
KNN
SVC
Tree-based Classification models
Principle Component Analysis (PCA)
K-Mean
Hierarchical

Polynomial regression model

The polynomial regression model is a special case of multiple linear regression models!

- Create new variables $X_1 = X,\ X_2 = X^2$, ... etc and then treat as multiple linear regression.
- Not really interested in the coefficients; more interested in the fitted function!

$$\hat{f}(X) = f_{w,b}(X) = b + w_1 x + w_2 x^2 + \ \ldots..w_d x^d$$

- W is a d-dimensional vector of parameters
- b is a real number
- d is the polynomial degree of the model (we either fix the d at some reasonably low value, else use **cross-validation** to choose d)

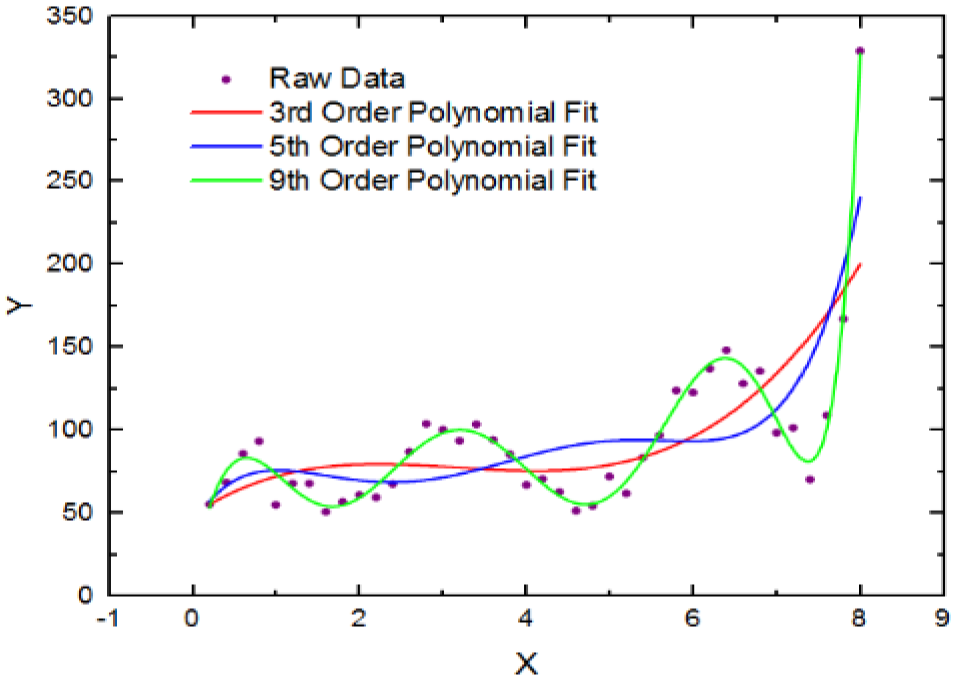
JON M. HUNTSMAN SCHOOL OF BUSINESS UtahStateUniversity

Pedram Jahangiry

The optimization problem

The optimization problem is defined as:

$$Min_{w,b}\ MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{y_i})^2 = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - f_{w,b}(X_i)\right)^2$$

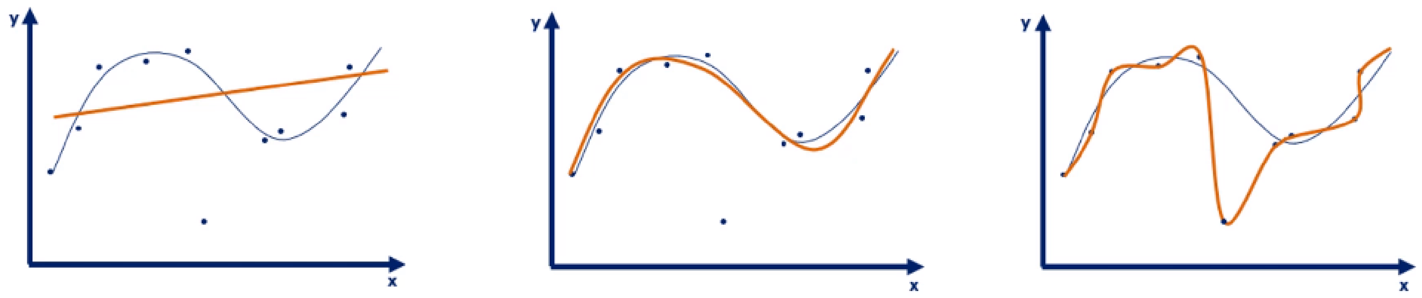


- We use the same **loss function** as in linear regression!
- The solution to this optimization problem is w^* and b^*
- Now we can make predictions!

JON M. HUNTSMAN SCHOOL OF BUSINESS
UtahStateUniversity

Pedram Jahangiry

Tuning the hyperparameter d!

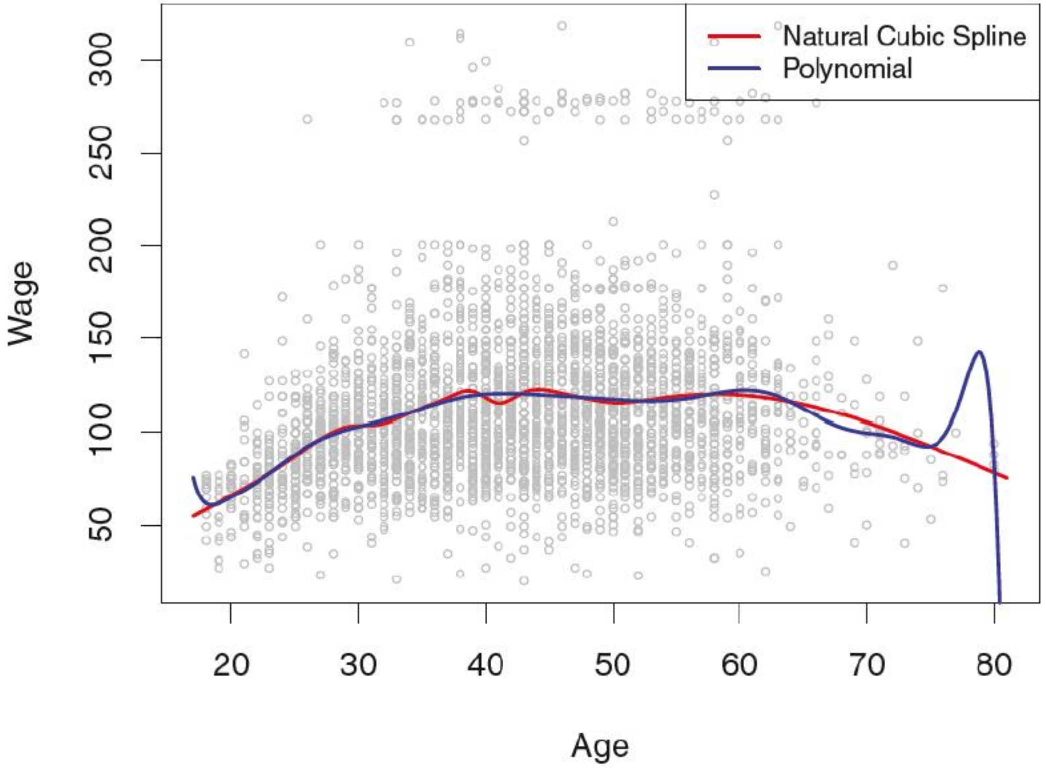


$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2$$

$$RMSE = \sqrt{MSE}$$

Caveats!

- In general quadratic loss functions are sensitive to outliers
- Polynomials have notorious tail behavior
- Polynomials are global fit!
- Solution: piecewise polynomial, splines and local regressions.



JON M. HUNTSMAN SCHOOL OF BUSINESS
UtahStateUniversity

Pedram Jahangiry

Course Modules

- ✓ Module 1- Introduction to Deep Learning
- ✓ Module 2- Setting up Machine Learning Environment
- ✓ Module 3- Linear Regression (Econometrics approach)
- ✓ Module 4- Machine Learning Fundamentals
- ✓ Module 5- Linear Regression (Machine Learning approach)
- Module 6- Penalized Regression (Ridge, LASSO, Elastic Net)
- Module 7- Logistic Regression
- Module 8- K-Nearest Neighbors (KNN)
- Module 9- Classification and Regression Trees (CART)
- Module 10- Bagging and Boosting
- Module 11- Dimensionality Reduction (PCA)
- Module 12- Clustering (KMeans – Hierarchical)