Statistical Learning for Data Science

Lecture 02

唐晓颖

电子与电气工程系

南方科技大学

February 15, 2023

Machine Learning and Statistical Learning

- ---Simon Blomberg:
  - *From R's fortunes package: To paraphrase provocatively, 'machine learning is statistics minus any checking of models and assumptions'.*

- ---Andrew Gelman:
  - *In that case, maybe we should get rid of checking of models and assumptions more often. Then maybe we'd be able to solve some of the problems that the machine learning people can solve but we can't!*

Machine Learning and Statistical Learning

- Purpose

The purpose of statistics is to make an inference about a population based on a sample.

Machine learning is used to make repeatable predictions by finding patterns within data.

- Data

Machine learning requires large amounts of data in order to make accurate predictions. Models are built using training data, fine tuned using a validation dataset, and are evaluated with a test dataset. All of these steps help the machine “learn.”

Statistics does not involve multiple subsets of data because you are not trying to make a prediction. The point of modeling in this situation is to display the relationship between data and the outcome variable. In addition, statistics relies on significance tests to determine the direction and magnitude of a relationship, discounting noise and confounding variables.

Machine Learning and Statistical Learning

- Interpretability

Machine learning: large number of variables, models can be simultaneously extremely accurate and almost impossible to understand. Statistical models: typically easier to understand because they are based on fewer variables, and the accuracy of relationships is supported by tests of statistical significance.

- Dimensional Differences

Statistics emphasizes statistical derivation of low-dimensional spatial problems （confidence intervals, hypothesis tests, optimal estimators）

Machine learning emphasizes high-dimensional prediction problems

- Areas of greater interest for each of statistics and machine learning

Statistics: survival analysis, spatial analysis, multiple testing, minimax theory, deconvolution, semiparametric inference, bootstrapping, time series.

Machine learning : online learning, semi-supervised learning, manifold learning, active learning, boosting.

Machine Learning and Statistical Learning (Professional term)

- Statistics — Machine Learning

Statistics	Machine Learning
Estimation	Learning
Hypothesis	Classifier
Data point	Example/Instance
Regression	Supervised Learning
Classification	Supervised Learning
Covariate	Feature
Response	Label

Object of statistical learning -- Data

Data is a collection of facts, such as numbers, words, measurements, observations or even just descriptions of things.

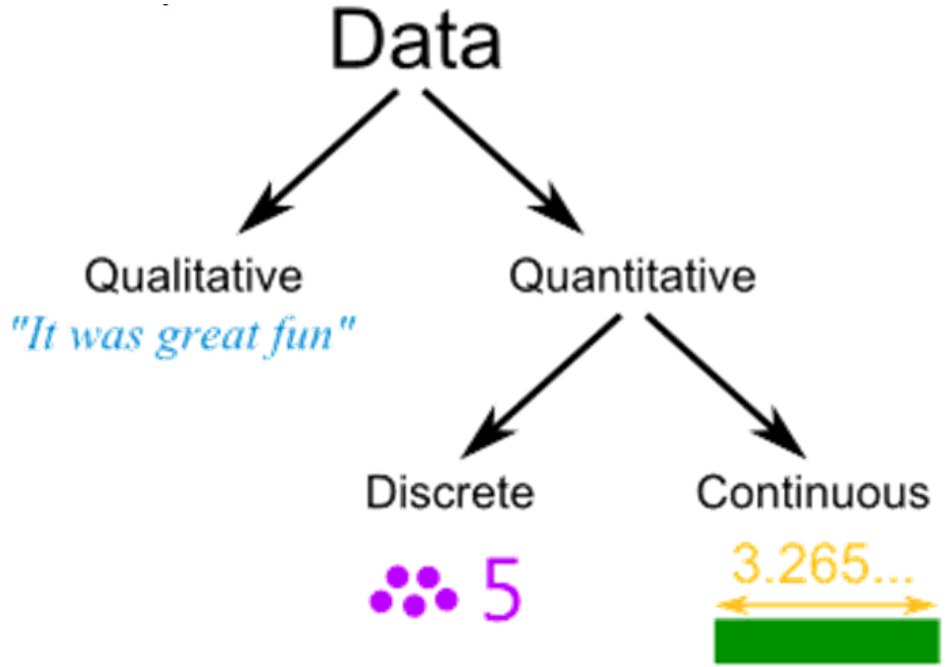
- The basic assumption of data is that similar data have certain statistical regularity.

Data are **measured**, **collected** and **reported**, and **analyzed**, whereupon it can be **visualized** using graphs, images or other analysis tools.

Data as a general concept refers to the fact that some existing **information** or **knowledge** is **represented** or **coded** in some form suitable for better usage or processing.

Qualitative vs Quantitative

- **Qualitative data** is descriptive information (it *describes* something)
- **Quantitative data**, is numerical information (numbers).



Data

Qualitative

- It is brown and black
- It has long hair
- It seems to be full of energy

Quantitative

- Discrete:
  - It has 4 legs
  - It has 2 brothers
- Continuous:
  - Its weight is in the range of (15kg, 30kg)
  - Its height is in the range of (300mm, 600mm)

The purpose of statistical learning is to predict and analyze data (especially unknown data).

Statistical Learning

- Methods of statistical learning
  - Category:
    - Supervised learning
    - Unsupervised learning
    - Semi-supervised learning

Supervised learning

- Instance, feature vector, feature space
- Feature vectors of the input instance x:

$$x = (x^{(1)}, x^{(2)}, \cdots, x^{(i)}, \cdots, x^{(n)})^{\mathrm{T}}$$

$$x_i = (x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)})^{\mathrm{T}}$$

- Training Set:

$$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$$

- Input and output variables:
  - classification problem, regression problem, labeling problem

Supervised learning

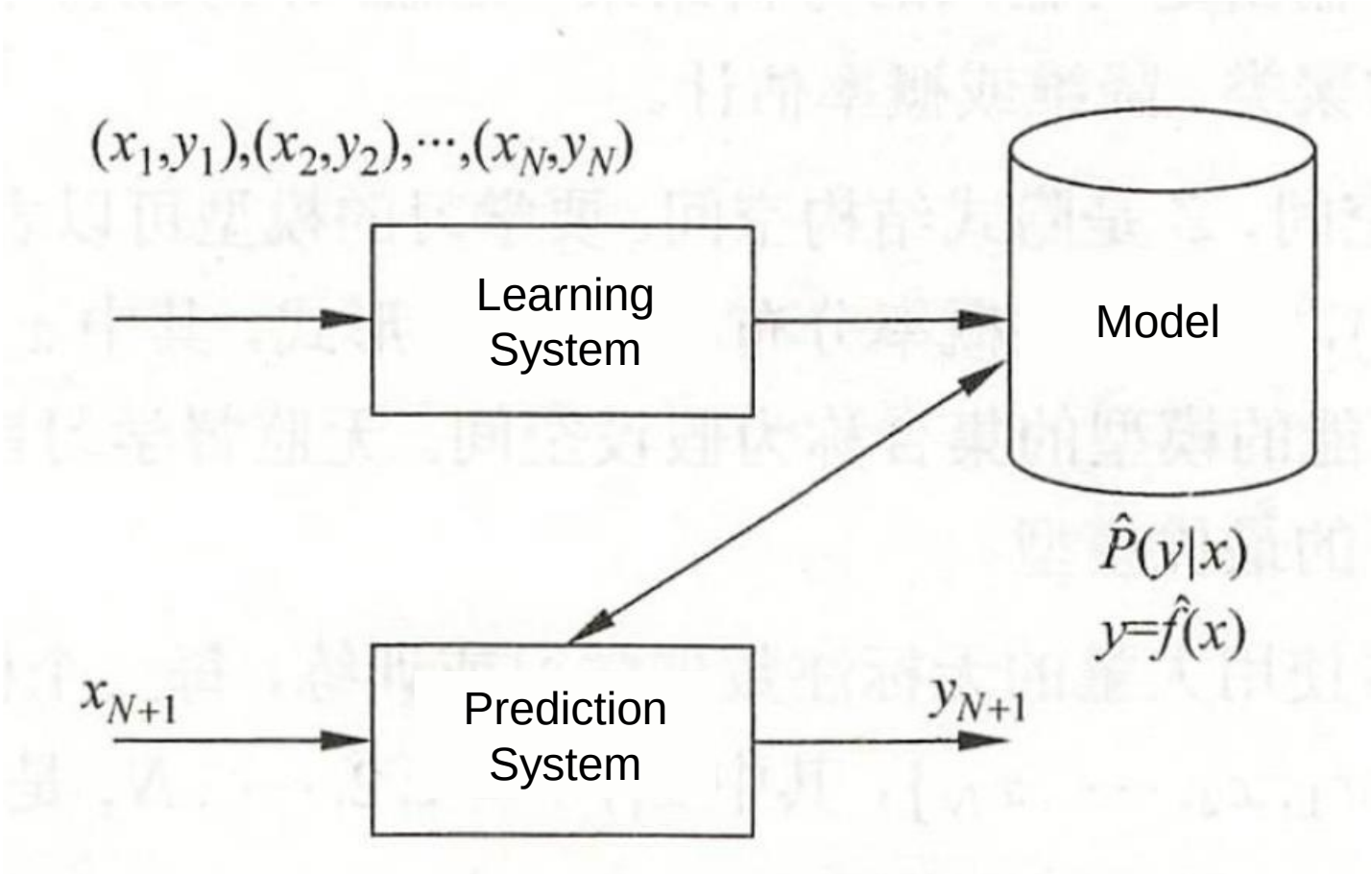
监督学习分为学习和预测两个过程，由学习系统与预测系统完成。在学习过程中，学习系统利用给定的训练数据集，通过学习（或训练）得到一个模型，表示为条件概率分布 $\hat{P}(Y|X)$ 或决策函数 $Y = \hat{f}(X)$。条件概率分布 $\hat{P}(Y|X)$ 或决策函数 $Y = \hat{f}(X)$ 描述输入与输出随机变量之间的映射关系。在预测过程中，预测系统对于给定的测试样本集中的输入 x_{N+1}，由模型 $y_{N+1} = \arg\max_y \hat{P}(y|x_{N+1})$ 或 $y_{N+1} = \hat{f}(x_{N+1})$ 给出相应的输出 y_{N+1}。

在监督学习中，假设训练数据与测试数据是依联合概率分布 $P(X,Y)$ 独立同分布产生的。

学习系统（也就是学习算法）试图通过训练数据集中的样本 (x_i, y_i) 带来的信息学习模型。具体地说，对输入 x_i，一个具体的模型 $y = f(x)$ 可以产生一个输出 $f(x_i)$，而训练数据集中对应的输出是 y_i。如果这个模型有很好的预测能力，训练样本输出 y_i 和模型输出 $f(x_i)$ 之间的差就应该足够小。学习系统通过不断地尝试，选取最好的模型，以便对训练数据集有足够好的预测，同时对未知的测试数据集的预测也有尽可能好的推广。

Supervised learning

- Formalization of problems



$$y_{N+1} = \arg\max_{y} \hat{P}(y \mid x_{N+1})$$

$$y_{N+1} = \hat{f}(x_{N+1})$$

Unsupervised learning

无监督学习可以用于对已有数据的分析，也可以用于对未来数据的预测。分析时使用学习得到的模型，即函数 $z = \hat{g}(x)$，条件概率分布 $\hat{P}(z|x)$，或者条件概率分布 $\hat{P}(x|z)$。预测时，和监督学习有类似的流程。由学习系统与预测系统完成

在学习过程中，学习系统从训练数据集学习，得到一个最优模型，表示为函数 $z = \hat{g}(x)$，条件概率分布 $\hat{P}(z|x)$ 或者条件概率分布 $\hat{P}(x|z)$。在预测过程中，预测系统对于给定的输入 x_{N+1}，由模型 $z_{N+1} = \hat{g}(x_{N+1})$ 或 $z_{N+1} = \arg\max_z \hat{P}(z|x_{N+1})$ 给出相应的输出 z_{N+1}，进行聚类或降维，或者由模型 $\hat{P}(x|z)$ 给出输入的概率 $\hat{P}(x_{N+1}|z_{N+1})$，进行概率估计。

Unsupervised learning

- Training set:

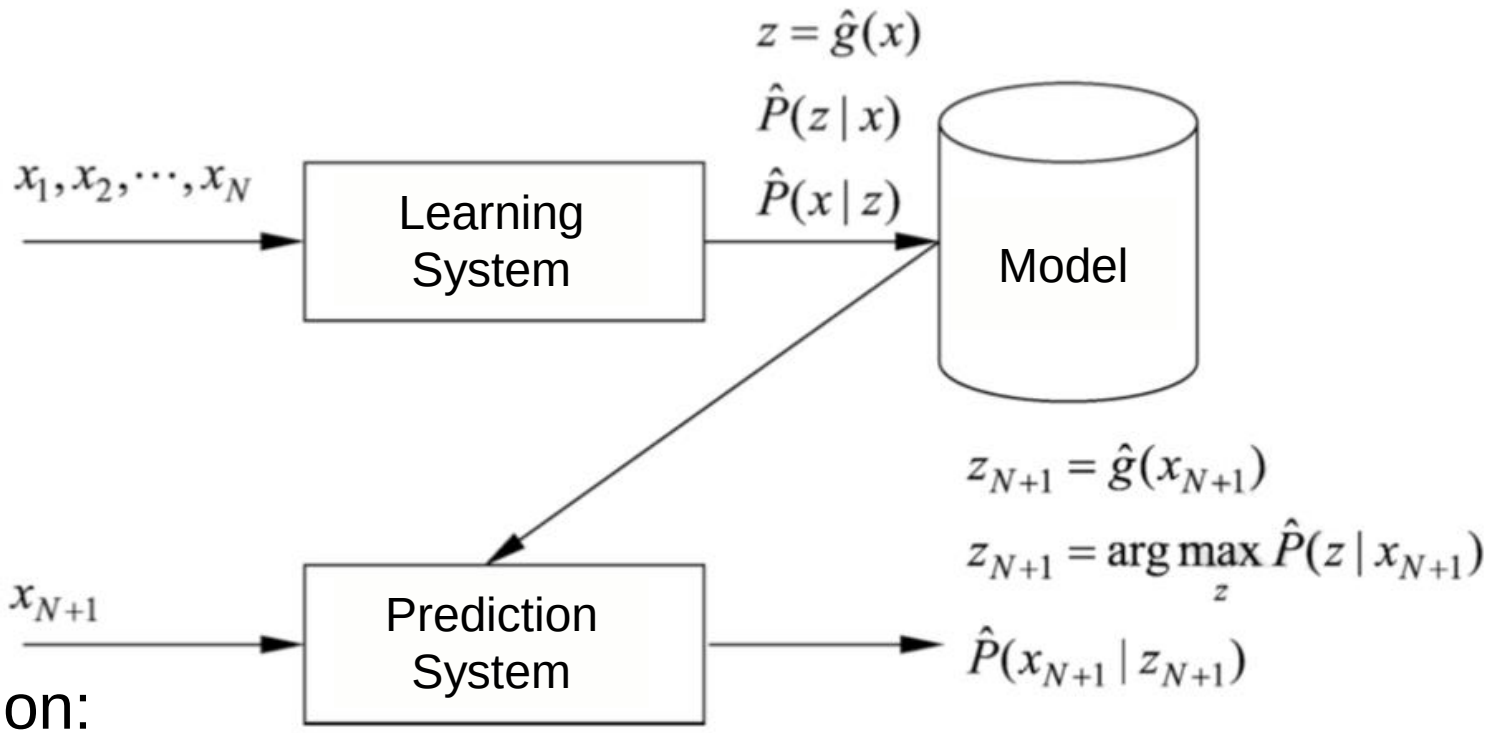
$$U = \{x_1, x_2, \cdots, x_N\}$$

- Model function:

$$z = g(x)$$

- Conditional probability distribution:

$$P(z|x)$$



Semi-supervised learning

- Small amount of labeled data, large amount of unlabeled data
- Use information from unlabeled data to assist labeled data for supervised learning
- Lower cost

Three elements of statistical learning

- Method = Model + Strategy + Algorithm
  - Model:
    - The set of decision functions $\mathcal{F} = \{f|Y = f(X)\}$
    - Parameter space $\mathcal{F} = \{f|Y = f_\theta(X), \theta \in \mathbf{R}^n\}$
    - Set of conditional probabilities $\mathcal{F} = \{P|P(Y|X)\}$
    - Parameter space $\mathcal{F} = \{P|P_\theta(Y|X), \theta \in \mathbf{R}^n\}$

Three elements of statistical learning

- Method = Model + Strategy + Algorithm
  - Strategy:
    - Loss function: how good or bad a prediction is
    - Risk function: how good or bad the model prediction is in the average sense
    - 0-1 loss function

$$L(Y, f(X)) = \begin{cases} 1, & Y \neq f(X) \\ 0, & Y = f(X) \end{cases}$$

    - quadratic loss function

$$L(Y, f(X)) = (Y - f(X))^2$$

    - absolute loss function

$$L(Y, f(X)) = |Y - f(X)|$$

    - logarithmic loss function/loglikelihood loss function

$$L(Y, P(Y|X)) = -\log P(Y|X)$$

Three elements of statistical learning

- Method = Model + Strategy + Algorithm

  - Strategy:

    - Expectation of loss function

$$R_{\exp}(f) = E_P[L(Y, f(X))]$$
$$= \int_{\mathcal{X}\times\mathcal{Y}} L(y, f(x)) P(x, y) \mathrm{d}x\mathrm{d}y$$

    - risk function, expected loss

    - P(x,y) can be derived directly from P(x|y), which is generally unknown
    - Given a training set

$$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$$

    - empirical risk, empirical loss

$$R_{\mathrm{emp}}(f) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i))$$

Three elements of statistical learning

- Method = Model + Strategy + Algorithm

  - Strategy: Empirical Risk Minimization and Structural Risk Minimization
    - Optimal model of empirical risk minimization

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i))$$

    - Empirical risk minimization may not work well when the sample size is small, and may result in "over-fitting".

    - Structural risk minimization, a strategy proposed to prevent over-fitting, is equivalent to regularization, adding the regularization term, or penalty term.

$$R_{\text{srm}}(f) = \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f)$$

Three elements of statistical learning

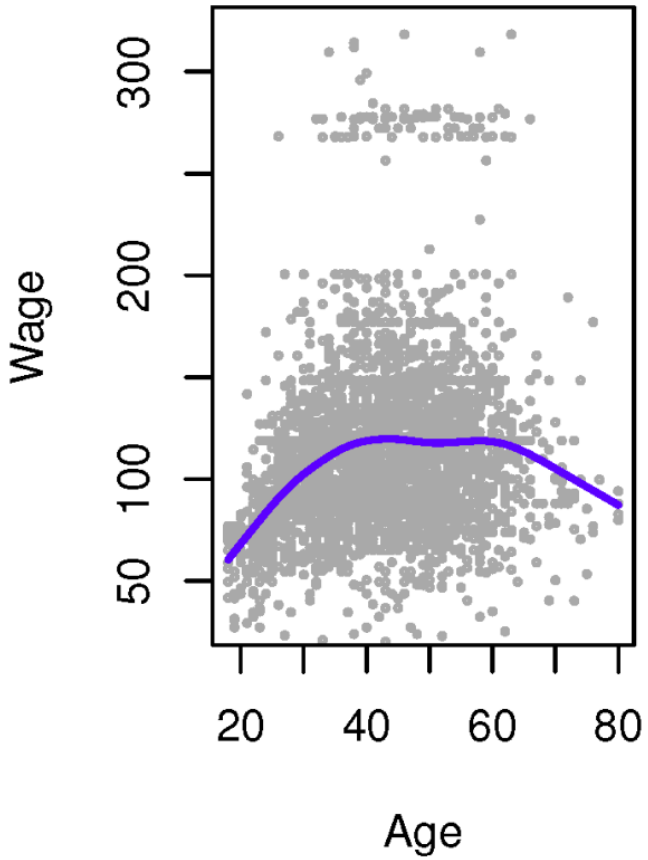
- To find the optimal model is to solve the optimization problem

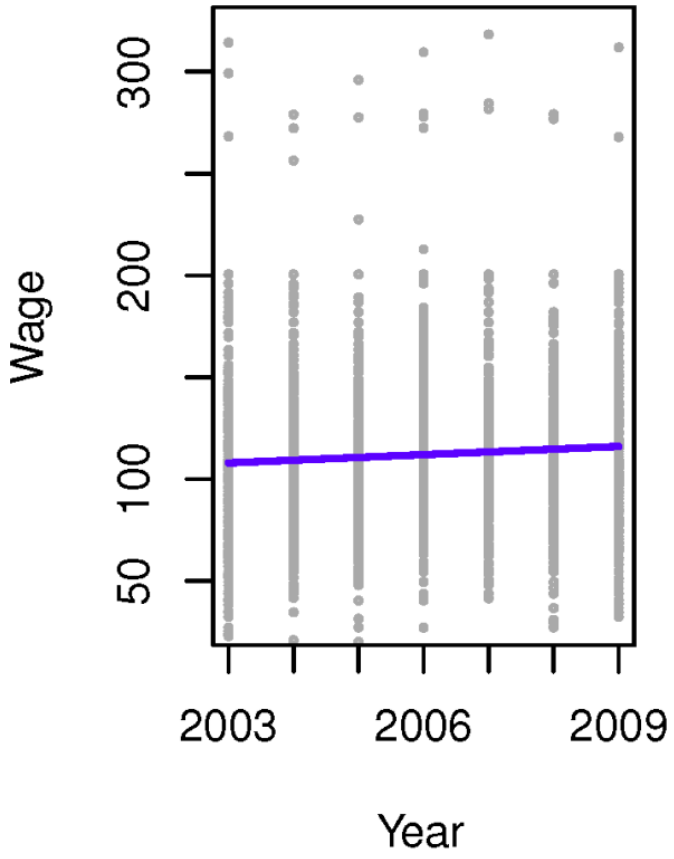
$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f)$$

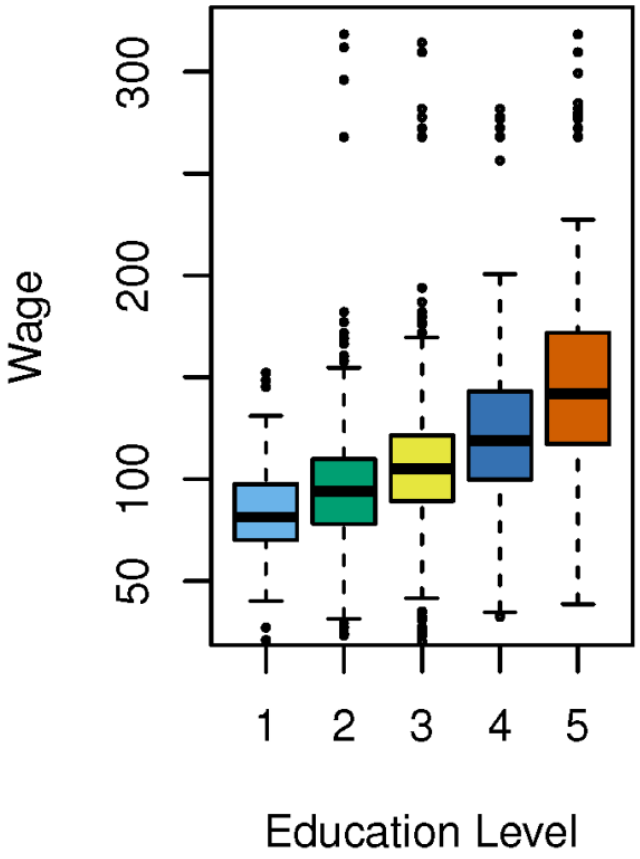
- Algorithm
  - If the optimization problem has an explicit analytic equation, the algorithm is relatively simple
  - But usually the analytic formula does not exist, so a numerical calculation is needed

Statistical Learning Problems

- Establish the relationship between salary and demographic variables in population survey data

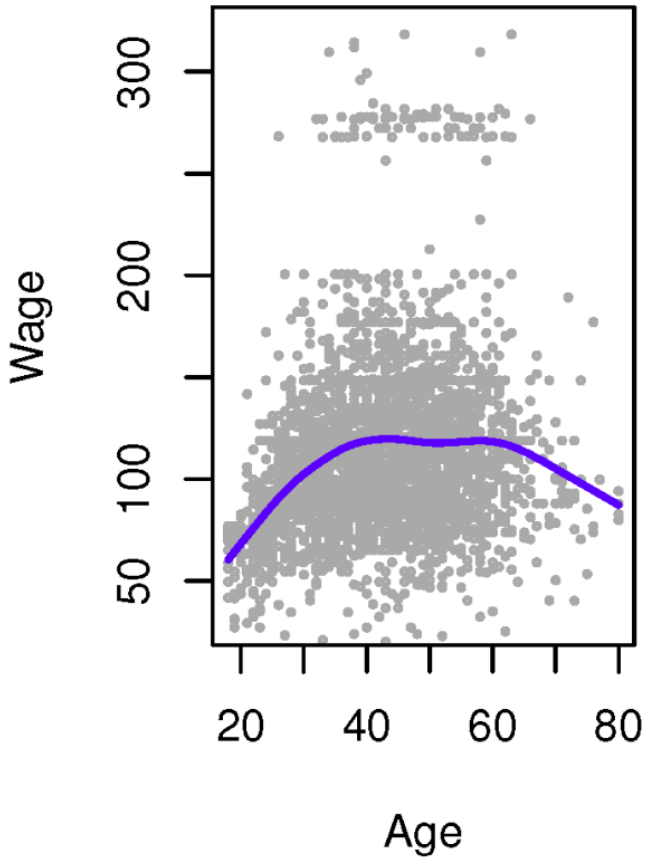






Statistical Learning Problems

- Establish the relationship between salary and demographic variables in population survey data



- Wage increases with age but then decreases again after approximately age 60
- Significant variability exists from person to person
- It is unlikely to have an accurate estimation of a person's wage based on his/her age alone