数据科学基础 I
（Matlab）

东北大学

相关与回归概述

变量间关系

变量之间的关系

确定性关系

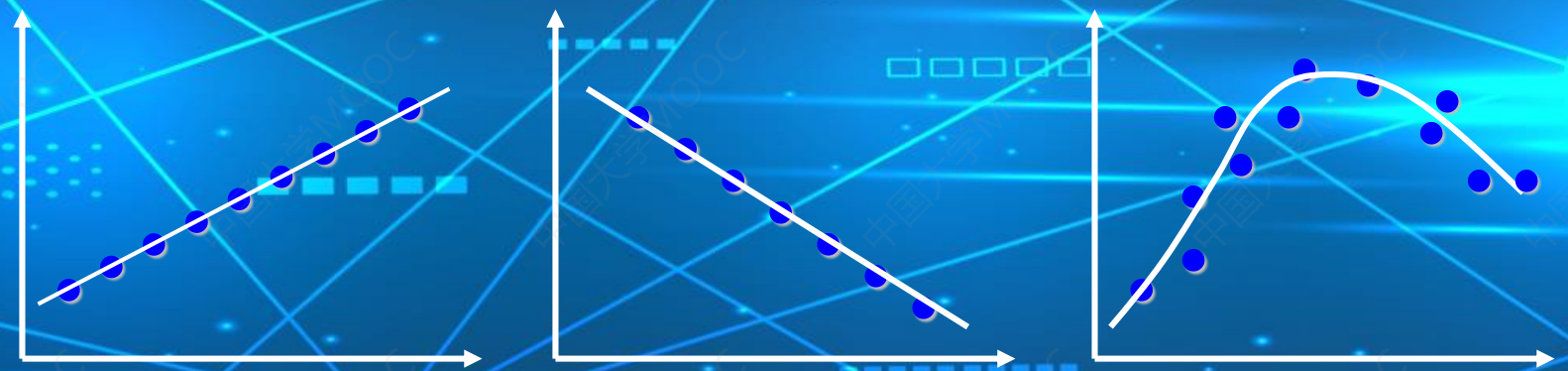
$S = \pi r^2$

相关关系

- 身高和体重

相关与回归概述

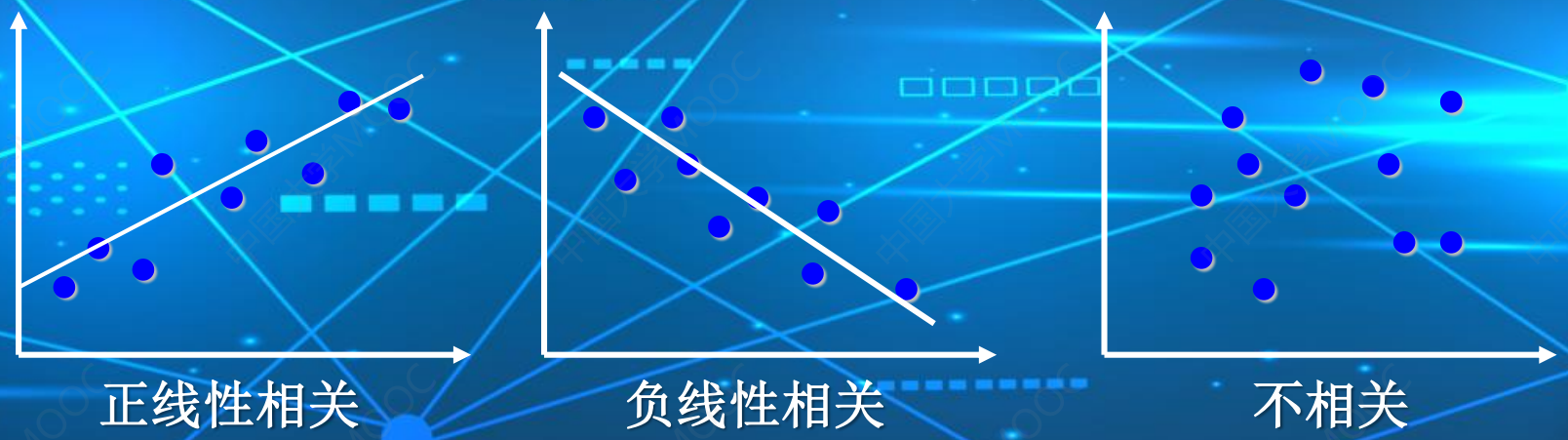
相关关系图示



完全正线性相关　　完全负线性相关　　非线性相关

相关与回归概述

相关关系图示



相关与回归概述

回归问题——统计学角度

- 1. 从一组样本数据出发，确定变量之间的数学关系式。
- 2. 对这些关系式的可信程度进行各种统计检验，并从影响某一特定变量的诸多变量中找出哪些变量的影响显著，哪些不显著。
- 3. 利用所求的关系式，根据一个或几个变量的取值来预测或控制另一个特定变量的取值，并给出这种预测或控制的精确程度。

相关与回归概述

回归问题——机器学习角度

- 回归和分类同为有监督学习问题。
- 回归模型表示的是从输入变量到输出变量之间映射的函数。
- 回归问题分为学习和预测两个过程。

相关与回归概述

回归和分类

- 回归和分类都用于预测，回归预测出一个值，分类预测出一个类别。
- 回归问题的输出是连续值，是定量的。
- 分类问题的输出是离散值，是定性的。
- 回归的目标是找到最优的拟合，分类的目标是找到最优的决策边界。

相关与回归概述

回归

- 回归的主要目标是找到一个（或多个）变量（自变量）与一个特定变量（因变量）的映射（函数）关系。
- 一个自变量的回归称为一元回归，多个自变量则为多元回归。
- 因变量与自变量之间为线性关系的回归称为线性回归，否则为非线性回归。

相关与回归概述

回归问题的步骤

（1）选模型。

- 模型就是函数的集合，例如线性函数模型。

（2）根据某种策略，选用某种算法学习出最优的模型参数

- 例如根据使得平方损失函数最小化的策略，使用最小二乘法学习出最优的函数。

（3）根据学习的模型进行新数据的预测。