THE UNIVERSITY OF CHICAGO

Glossário do Módulo 4

Aprendizagem sem supervisão: é um sistema no qual apenas são proporcionados dados de entrada a um algoritmo, sem os valores de saída correspondentes para guiá-lo.
Aprendizagem supervisionada: é um sistema no qual os dados de entrada e saída são fornecidos e que atua como base de aprendizagem para o processamento de dados futuro.

Classificação: em estatística, sua finalidade é categorizar os dados de acordo com seus atributos. Por exemplo, se os gostos de uma população são analisados de acordo com sua idade, em vez de agrupar os dados em todas as idades, geralmente são escolhidas categorias que incluem intervalos mais amplos, por exemplo, de 0 a 10 anos, de 10 a 20, etc**.**

Clusterização: consiste em agrupar os dados em subconjuntos de acordo com seus atributos. Na classificação, o analista define quais categorias/grupos possui, e os dados são associados de acordo com estas definições. Entretanto, na clusterização, que é utilizada em *machine learning*, o próprio sistema determina, com base nos dados e atributos que recebe, quais seriam as categorias ideais, de modo que as distribuí de acordo a categoria correspondente.

***Data furnace*:** método proposto para aquecer casas ou outras áreas, incorporando centros de armazenamento de dados, cujo funcionamento produz calor.

***Data mining*:** processo de extração e descoberta de informações e conhecimentos úteis a partir de grandes volumes de dados.

Detecção de anomalias: este processo serve para identificar casos em um estudo que, por apresentarem condições diferentes das usuais, alterariam de forma errónea os resultados obtidos. Como exemplo, o objetivo é descobrir o tempo médio de produção de um carro a partir de mais de 1.000 dados. Todas as marcas que levam entre 6 e 8 horas, exceto 2: uma indica 1 hora e a outra 20 horas. Deduzimos que ambos os dados seriam "raros". Há muitas ferramentas que realizam análises para detectar estas anomalias ou dados incomuns. O objetivo é identificá-los e excluí-los para que as análises sejam muito mais precisas.

Processamento de linguagem natural: área de inteligência artificial que dá às máquinas a capacidade de ler, compreender e deduzir o significado da linguagem humana.

Regressão: refere-se à análise estatística que é realizada quando uma variável é comparada com outras. A partir dela, podemos estudar as possíveis dependências. Por exemplo, poderíamos determinar a probabilidade de alguém não pagar uma hipoteca com base em sua idade e renda. Nesse caso, em um gráfico, colocaríamos a idade em um eixo e a renda no outro. Desta forma, analisaríamos a evolução desta probabilidade. Geralmente, quanto maior a idade e o nível de renda, menor o risco de inadimplência.