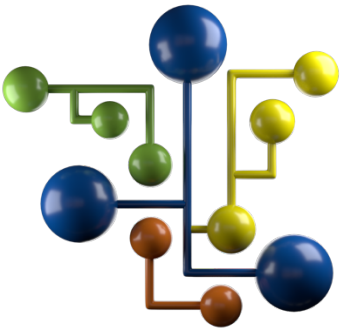


Data Science Academy

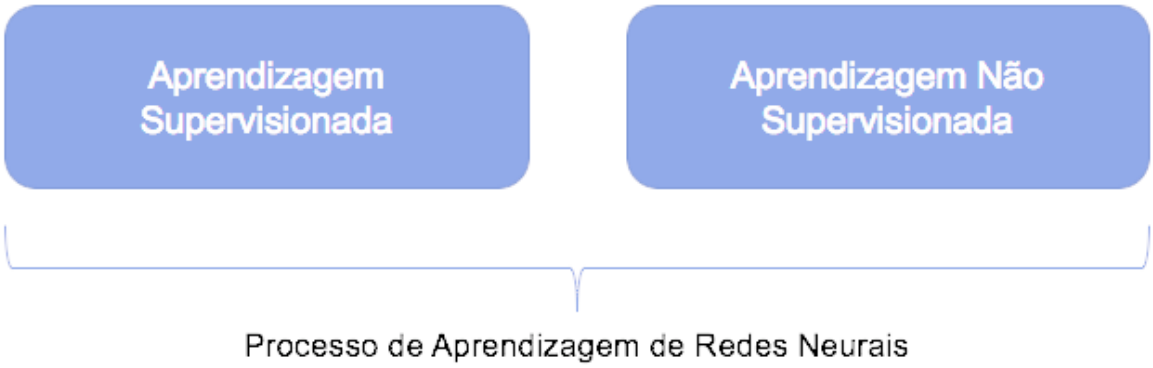
www.datascienceacademy.com.br

Machine Learning

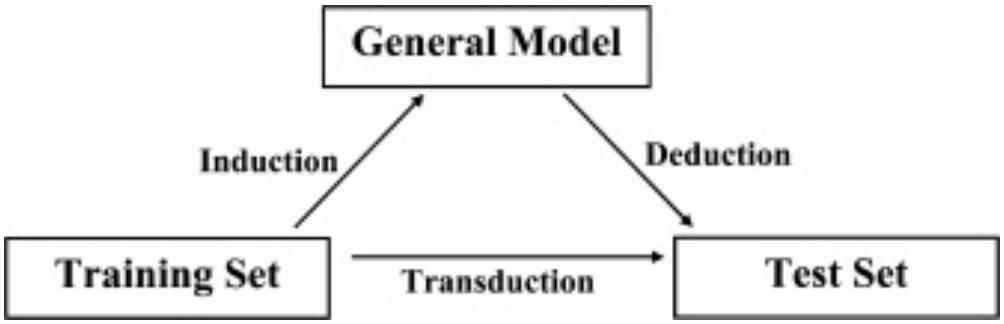
Teoria do Aprendizado Estatístico

Vamos começar fazendo uma breve revisão do processo de aprendizagem de máquina e na sequência vamos estudar a Teoria do Aprendizado Estatístico.

O processo de aprendizado de uma rede neural pode se dar de duas formas: aprendizagem supervisionada e aprendizagem não supervisionada.

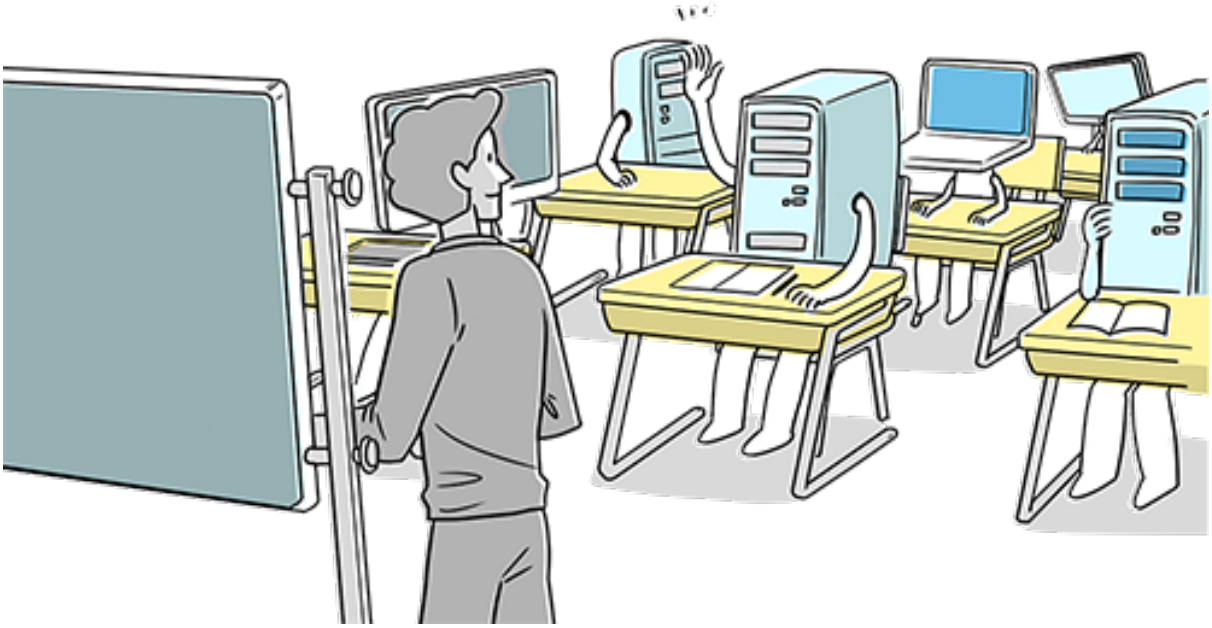


As técnicas de Machine Learning empregam um princípio de inferência denominado indução, no qual obtém-se conclusões genéricas a partir de um conjunto particular de exemplos (que são os dados de treinamento). O aprendizado indutivo pode ser dividido em dois tipos principais: supervisionado e não- supervisionado.

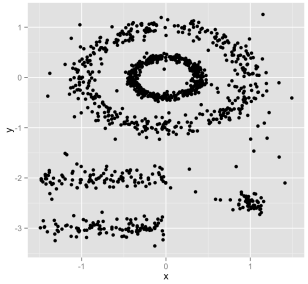


No aprendizado supervisionado tem-se a figura de um professor externo, o qual apresenta o conhecimento do ambiente por conjuntos de exemplos na forma: dados de entrada, dados da saída desejada. O algoritmo de ML extrai a representação do conhecimento a partir desses exemplos. O objetivo é que a representação gerada seja capaz de produzir saídas corretas para novas entradas não apresentadas previamente.

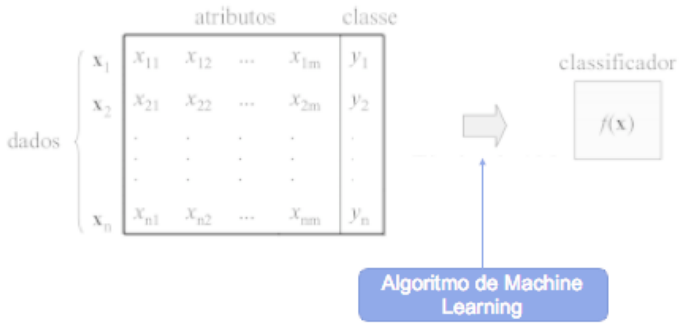
No aprendizado não-supervisionado não há a presença de um professor, ou seja, não existem exemplos rotulados. O algoritmo de Machine Learning aprende a representar (ou agrupar) as entradas submetidas segundo uma medida de qualidade. Essas técnicas são utilizadas principalmente quando o objetivo for encontrar padrões ou tendências que auxiliem no entendimento dos dados.



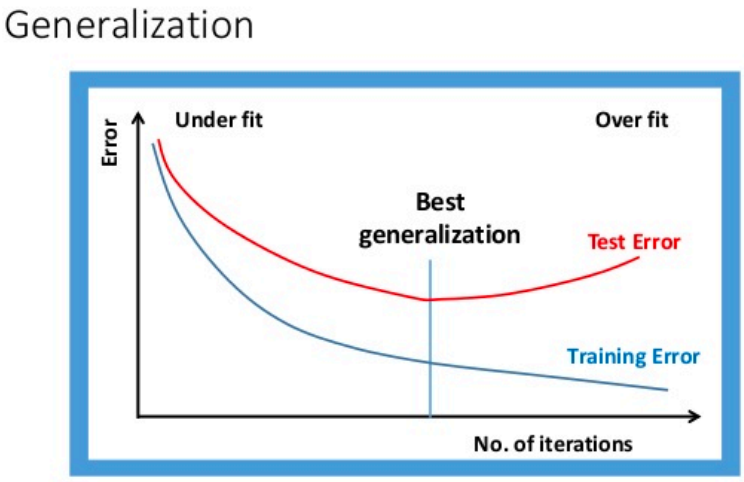
Um requisito importante para as técnicas de Machine Learning é que elas sejam capazes de lidar com dados imperfeitos, denominados ruídos. Muitos conjuntos de dados apresentam esse tipo de caso, sendo alguns erros comuns a presença de dados com rótulos e/ou atributos incorretos. A técnica de Machine Learning deve idealmente ser robusta a ruídos presentes nos dados, procurando não fixar a obtenção dos classificadores sobre esse tipo de caso. Deve-se também minimizar a influência de outliers no processo de indução. Os outliers são exemplos muito distintos dos demais presentes no conjunto de dados. Esses dados podem ser ruídos ou casos muito particulares, raramente presentes no domínio dos novos dados que serão apresentados ao modelo.



Os conceitos referentes à geração de um classificador a partir do aprendizado supervisionado são representados de forma simplificada neste diagrama abaixo. Tem-se nessa figura um conjunto com n dados. Cada dado xi possui m atributos, ou seja, xi = (xi1, . . . , xim). As variáveis yi representam as classes. A partir dos exemplos e as suas respectivas classes, o algoritmo de ML extrai um classificador. Pode-se considerar que o modelo gerado fornece uma descrição compacta dos dados fornecidos. A obtenção de um classificador por um algoritmo de Machine Learning a partir de uma amostra de dados também pode ser considerada um processo de busca. Procura-se, entre todas as hipóteses que o algoritmo é capaz de gerar a partir dos dados, aquela com melhor capacidade de descrever o domínio em que ocorre o aprendizado. Para estimar a taxa de predições corretas ou incorretas (também denominadas taxa de acerto e taxa de erro, respectivamente) obtidas por um classificador sobre novos dados, o conjunto de exemplos é, em geral, dividido em dois subconjuntos disjuntos: de treinamento e de teste. O subconjunto de treinamento é utilizado no aprendizado do conceito e o subconjunto de teste é utilizado para medir o grau de efetividade do conceito aprendido na predição da classe de novos dados.



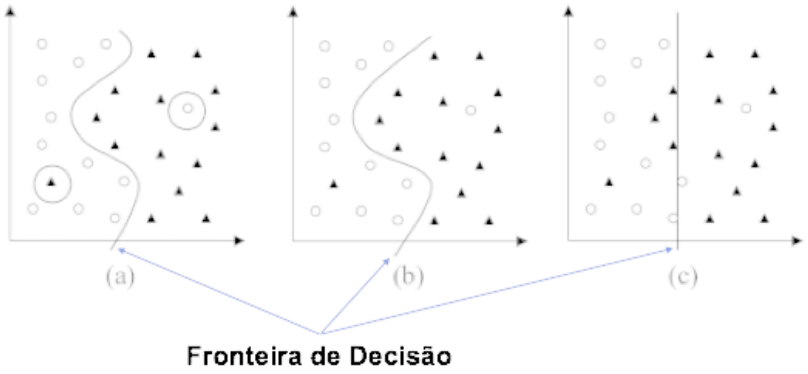
Outro importante conceito empregado em Machine Learning é o de generalização de um classificador, definida como a sua capacidade de prever corretamente a classe de novos dados. No caso em que o modelo se especializa nos dados utilizados em seu treinamento, apresentando uma baixa taxa de acerto quando confrontado com novos dados, tem-se a ocorrência de um superajustamento (overfitting). É também possível induzir hipóteses que apresentem uma baixa taxa de acerto mesmo no subconjunto de treinamento, configurando uma condição de subajustamento (underfitting). Essa situação pode ocorrer, por exemplo, quando os exemplos de treinamento disponíveis são pouco representativos ou quando o modelo obtido é muito simples.



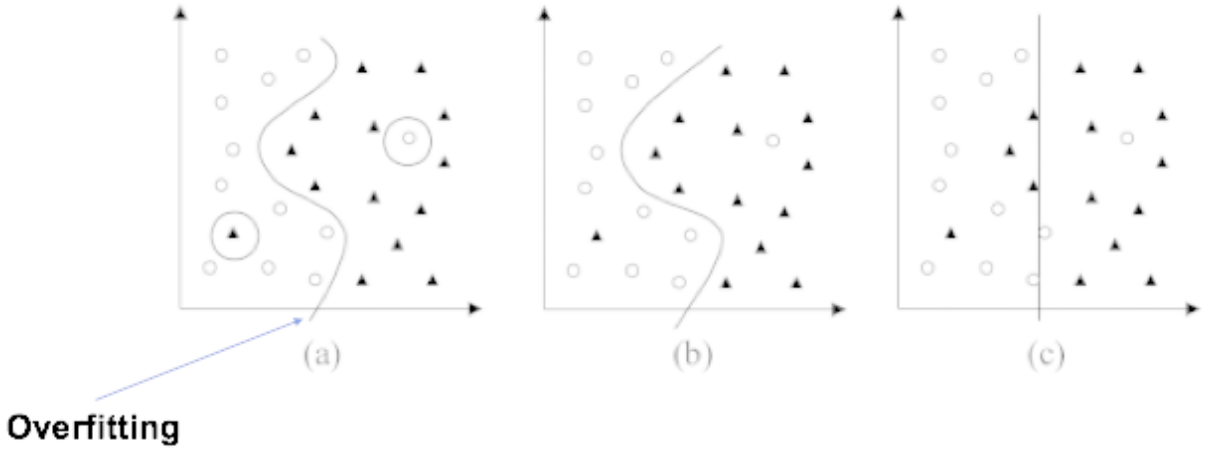
Depois desta breve revisão sobre os conceitos de Aprendizagem de máquina, vamos estudar outro importante conceito por trás de outros algoritmos de Machine Learning. Vamos começar definindo o que é a Teoria do Aprendizado Estatístico, que é a teoria por trás de vários algoritmos de Machine Learning em aprendizagem supervisionada!

Considere f um classificador e F o conjunto de todos os classificadores que um determinado algoritmo de Machine Learning pode gerar. Esse algoritmo, durante o processo de aprendizado, utiliza um conjunto de treinamento T, composto de n pares (xi , yi), para gerar um classificador particular ˆf ∈ F.

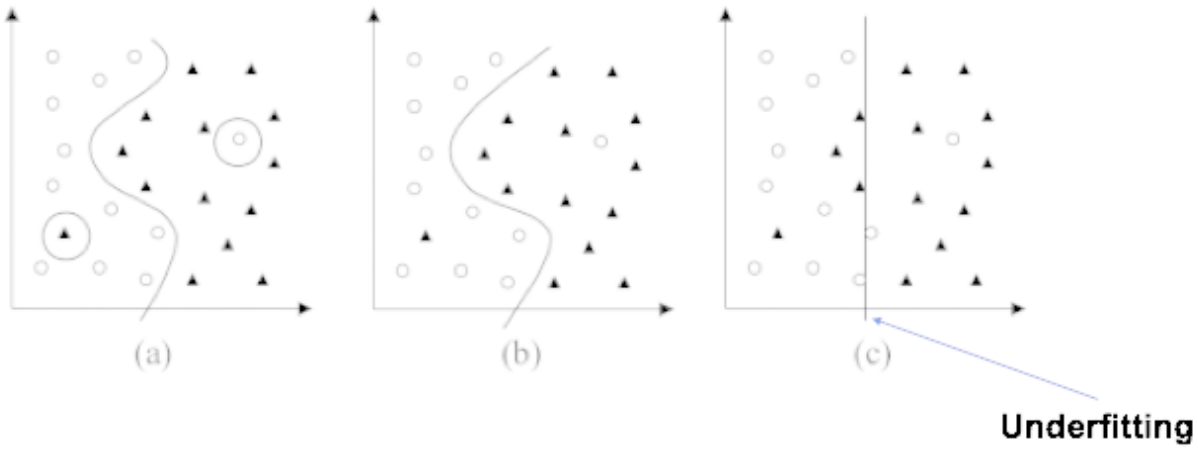
Considere, por exemplo, o conjunto de treinamento desta figura abaixo. O objetivo do processo de aprendizado é encontrar um classificador que separe os dados das classes “círculo” e “triângulo”. As funções ou hipóteses consideradas são ilustradas na figura por meio das bordas, também denominadas fronteiras de decisão, traçadas entre as classes.



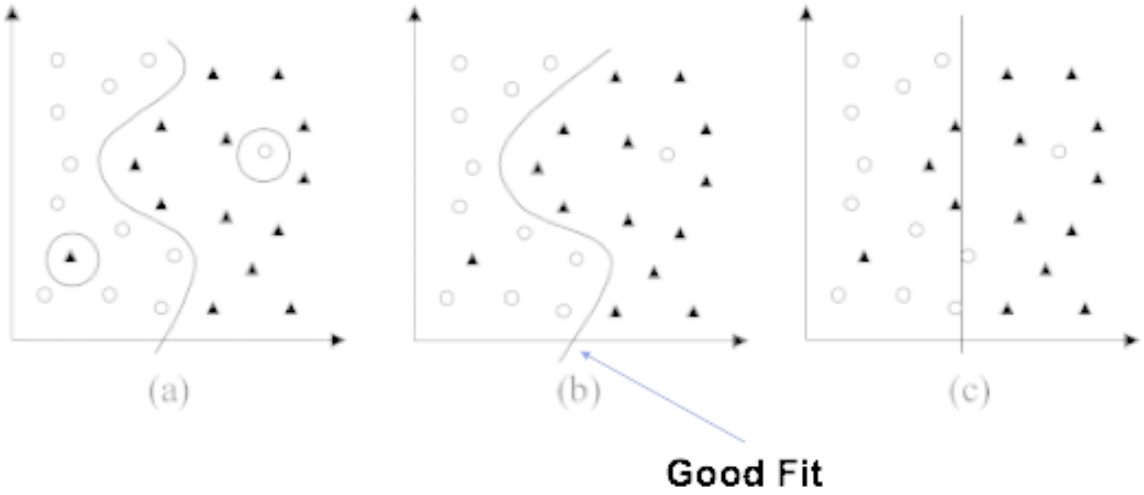
Na imagem da Figura a, tem-se uma hipótese que classifica corretamente todos os exemplos do conjunto de treinamento, incluindo dois possíveis ruídos. Por ser muito específica para o conjunto de treinamento, essa função apresenta elevada suscetibilidade a cometer erros quando confrontada com novos dados. Esse caso representa a ocorrência de um superajustamento do modelo aos dados de treinamento.



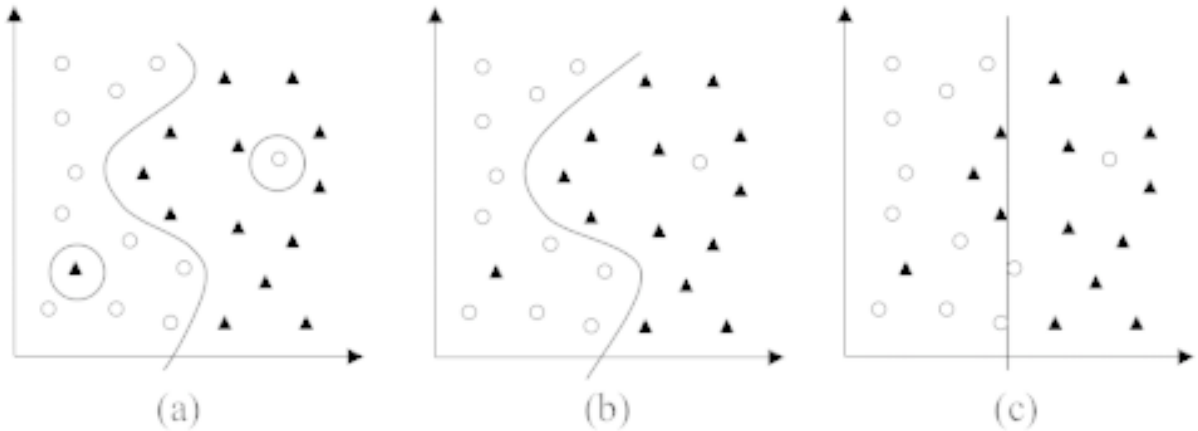
Um outro classificador poderia desconsiderar pontos pertencentes a classes opostas que estejam muito próximos entre si. A ilustração da c representa essa alternativa. A nova hipótese considerada, porém, comete muitos erros, mesmo para casos que podem ser considerados simples. Tem-se assim a ocorrência de um subajustamento, pois o classificador não é capaz de se ajustar mesmo aos exemplos de treinamento.



Um meio termo entre as duas funções descritas é representado na figura b. Esse preditor tem complexidade intermediária e classifica corretamente grande parte dos dados, sem se fixar demasiadamente em qualquer ponto individual.



A Teoria de Aprendizado Estatístico estabelece condições matemáticas que auxiliam na escolha de um classificador particular ˆf a partir de um conjunto de dados de treinamento. Essas condições levam em conta o desempenho do classificador no conjunto de treinamento e a sua complexidade, com o objetivo de obter um bom desempenho também para novos dados do mesmo domínio. Na aplicação da Teoria de Aprendizado Estatístico, assume-se inicialmente que os dados do domínio em que o aprendizado está ocorrendo são gerados de forma independente e identicamente distribuída de acordo com uma distribuição de probabilidade P(x, y), que descreve a relação entre os dados e os seus rótulos.



O erro (também denominado risco) esperado de um classificador f para dados de teste pode então ser quantificado por esta Equação. O risco esperado mede então a capacidade de generalização.

$$R\left(f\right) = \int c(f\left(\mathbf{x}\right), y) dP\left(\mathbf{x}, y\right)$$

Nesta Equação, c(f(x), y) é uma função de custo relacionando a previsão f(x) quando a saída desejada é y. Essa função retorna o valor 0 se x é classificado corretamente e 1 caso contrário. Infelizmente, não é possível minimizar o risco esperado, apresentado na Equação, diretamente, uma vez que em geral a distribuição de probabilidade P(x, y) é desconhecida. Tem-se unicamente a informação dos dados de treinamento. Normalmente utiliza-se o princípio da indução para inferir uma função ˆf que minimize o erro sobre esses dados e espera-se que esse procedimento leve também a um menor erro sobre os dados de teste.

$$R\left(f\right) = \int c(f\left(\mathbf{x}\right), y) dP\left(\mathbf{x}, y\right)$$

O risco empírico de f, fornecido pela Equação 2, mede o desempenho do classificador nos dados de treinamento, por meio da taxa de classificações incorretas obtidas em T. Esse processo de indução com base nos dados de treinamento conhecidos, constitui o princípio de minimização do risco empírico

$$R_{emp}\left(f\right) = \frac{1}{n} \sum_{i=1}^{n} c(f\left(\mathbf{x}_i\right), y_i)$$

Embora a minimização do risco empírico possa levar a um menor risco esperado, nem sempre isso ocorre. Considere, por exemplo, um classificador que memoriza todos os dados de treinamento e gera classificações aleatórias para outros exemplos. Embora seu risco empírico seja nulo, seu risco esperado é 0,5. Nessa direção, a Teoria do Aprendizado Estatístico provê diversos limites no risco esperado de uma função de classificação, os quais podem ser empregados na escolha do classificador. São esses limites estabelecidos pela Teoria do Aprendizado Estatístico, sobre os quais alguns modelos SVM se baseiam.

Referências:

Deep Learning Book
http://www.deeplearningbook.com.br/