1. Conceitos Básicos

A Aprendizagem de Máquinas é o ramo da inteligência artificial que tem como base técnicas que permitem que a máquina aprenda a solução para um determinado problema a partir de dados fornecidos como exemplo, tornando esta capaz de identificar padrões e tomar decisões com o mínimo de intervenção humana. Segundo Arthur Samuel (1959), o aprendizado de máquina é o "campo de estudo que dá aos computadores a habilidade de aprender sem serem explicitamente programados".

Na programação tradicional **"Dados + Programa = Saída"** enquanto em Aprendizagem de Máquina podemos dizer que **"Dados + Saída = Programa".**

Onde encontramos AM?
- carros autônomos
- sistemas de recomendação
- *chatbots*, ou robôs de conversação
- identificação por biometria (face, íris, voz, ...)
- OCRs
- dentre outras aplicações.

E você sabe quando utilizar Aprendizagem de máquina (AM)?
Quando não for possível uma solução determinística para um problema e existir dados suficientes para gerar um modelo cognitivo com boa capacidade de generalização.

Em geral, como são os algoritmos de AM?
Algoritmos de aprendizagem de máquina são formados basicamente por 3 componentes principais:
- Representação: como devemos representar o problema?
- Avaliação: como devemos avaliar o modelo ou hipótese criada?
- Otimização: qual estratégia utilizar para obter as hipóteses no espaço de solução do problema?

E quais são os tipos de aprendizagem de máquina?
Supervisionada: consiste em treinar a máquina a partir de exemplos para os quais conhecemos a saída (resposta ou solução). Tarefas principais: classificação e regressão. Na classificação a saída é categórica (discreta), por exemplo, classificar imagem de objetos em categorias como carro, bicicleta, bola, etc. Já na regressão a saída é um valor contínuo, por exemplo, estimar a idade de uma pessoa ou prever o valor de mercado de um veículo.

Não Supervisionada: os dados de treinamento não possuem a saída (resposta ou solução). A principal tarefa aqui é o agrupamento (*clustering*) que consiste em encontrar grupos de dados (*clusters*) segundo algum critério de similaridade. Um exemplo seria agrupar clientes de um banco ou seguradora com base na similaridade de suas operações (compras) na tentativa de descobrir grupos de clientes e identificar seus perfis.

Semi-supervisionada: os dados de treinamento possuem alguns poucos exemplos com as respectivas saídas (respostas ou solução). Há várias situações reais onde obter dados com os rótulos é muito difícil e caro, nestes casos as técnicas semi-supervisionadas representam uma alternativa na busca de solucionar tarefas como classificação e regressão.

Por reforço: não há uso de uma base com exemplos de treinamento, a máquina aprende a partir de recompensas recebidas por suas ações em determinado ambiente. As técnicas deste tipo são muito

utilizadas na robótica e em jogos digitais.

2. Terminologia

Há conceitos básicos importantíssimos que devem ser assimilados para que possamos trabalhar de forma adequada. Na **seção 2 do artigo *Reviewing some Machine Learning Concepts and Method*** temos a descrição dos principais conceitos que devemos conhecer, alguns dos quais foram traduzidos aqui:

a) **Bases de dados (ver Figura 1):** base contendo instâncias (Ti) a serem utilizadas no processo de aprendizagem. A base de um determinado problema normalmente é dividida em treinamento e teste. A base de treinamento é utilizada na etapa de construção do modelo. Desta, ainda é comum separarmos uma parte para validação. Aprende-se com o exemplo de treinamento e ajustam-se parâmetros (busca de otimização) ou mesmo encerra-se o processo de treinamento utilizando-se a base de validação. Já a base de teste é uma caixa-preta, e deve ser utilizada somente na avaliação final de desempenho dos modelos (hipóteses) criados.

b) **Instância:** um exemplo de treinamento, validação ou teste formado por atributos (*features*) de entrada (Xi) e atributo alvo (y).

c) **Atributo (*feature*)**: uma instância é composta de atributos (Xi) que podem ser de diferentes tipos (inteiro, real, categórico, ...). Estes representam características que descrevem uma instância.

Figura 1 - Formato Base de dados (aprendizagem)

	X_1	X_2	...	X_m	Y
T1	x_{11}	x_{12}	...	x_{1m}	y_1
T2	x_{21}	x_{22}	...	x_{2m}	y_2
...	...	...	...	...	...
Tn	x_{n1}	x_{n2}	...	x_{nm}	y_n

d) **Função alvo:** *y=f(X)*, função a ser aprendida e que mapeia o vetor de atributos X ao valor de saída *y*.

e) **Hipótese**: uma aproximação de *f(X)*, uma solução candidata (ou modelo).

f) **Indutor**: técnica ou algoritmo utilizado no aprendizado.

g) **Espaço de hipóteses**: conjunto de possíveis aproximações de *f(X)* que um indutor pode criar.

h) **Classificador**: um modelo de aprendizado cuja saída é discreta (categórica).

i) **Regressor**: um modelo de aprendizado cuja saída é contínua.

j) **Agrupamento** (*clustering*): tarefa não supervisionada que busca agrupar dados por similaridade

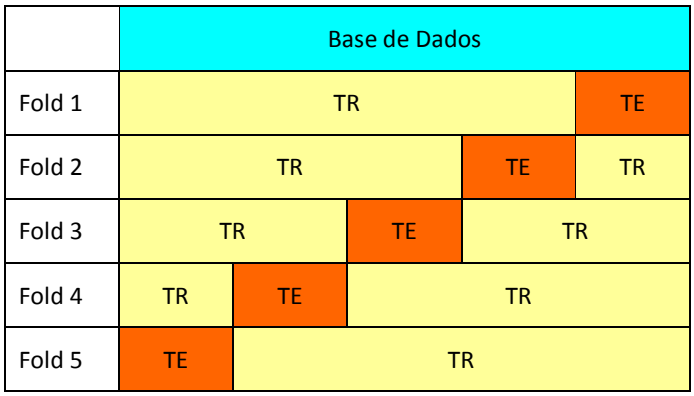
k) **Principais opções para avaliação de modelo supervisionado (classificação ou regressão)**

Validação cruzada (ver Figura 2): divide-se a base de dados do problema em *N folds* (ou

pastas) normalmente de maneira estratificada (mantendo a distribuição original da base). Selecionados um *fold* para teste e treina-se um modelo com os *N-1 folds* restantes. Repetimos isto *N* vezes selecionando a cada vez um novo *fold* para teste. Isto garantirá que todos os exemplos da base de dados sejam usados como teste.

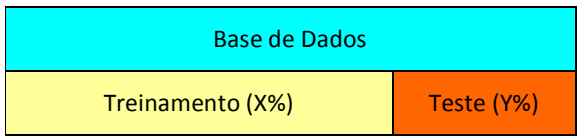
Holdout (*percentage split*): dividir a base em treinamento (X%) e teste (Y%) de forma estratificada, usualmente algo em torno de 60% para treinamento e 40% para teste (ver Figura 3). Para evitar viés na definição do teste, utiliza-se seleção aleatória de exemplos e ainda replicação que consiste em repetir o processo de divisão das bases *N* vezes, calculando-se a acurácia com base na média dos experimentos realizados.

Figura 2 - Esquema Validação Cruzada (5-Fold)



Fonte: O autor

Figura 3 - Esquema Holdout



Fonte: O autor

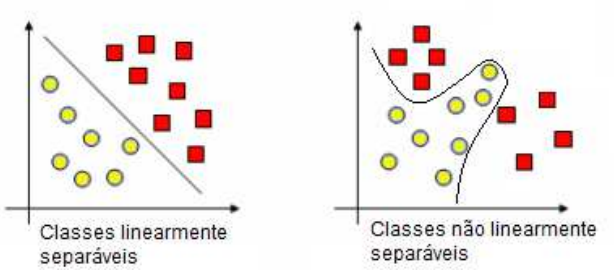
Comum: (X=60% Y=40%) ou : (X=70% Y=30%)

Resubstituição: uso da própria base de treinamento para avaliar o modelo criado. Não garante nada com relação à capacidade de generalização do modelo, mas é útil para demonstrar se o aprendizado convergiu (se os dados de treinamento são suficientes para a criação do modelo).

Leave-One-Out: deixa-se uma instância da base de dados de fora a qual será usada para teste, treina-se o modelo com o restante. Repete-se o processo *N* vezes retirando-se uma instância diferente para teste, sendo *N* a quantidade de instâncias na base.

I) **Dificuldade do problema de classificação:** há problemas cujas classes são linearmente separáveis, assim como há problemas cuja fronteira entre classes é mais complicada de ser definida. A Figura 4 mostra um problema onde é possível a separação linear e outro cuja fronteira não é linear.

Figura 4 – Tipos de Problemas de Classificação



3. Métricas de Avaliação

3.1) No caso de um classificador: as medidas mais utilizadas são **acurácia** (**taxa geral de acerto**) e **matriz de confusão**. Contudo, para problemas desbalanceados, ou seja, onde há um desiquilíbrio na quantidade de exemplos por classe, recomenda-se medir **precisão**, **revocação** e **f-measure** (f1-score).

Por exemplo: considere um problema de duas classes (indivíduo tem ou não tem câncer). Pode acontecer da acurácia ser alta, porém em situação de desbalanceamento, o sistema estar classificando muito bem uma classe e outra não. O que pode ser um problema muito grave.

A matriz de confusão apresentada considera um problema binário, contudo no caso de múltiplas classes (W), podemos usar uma matriz WxW. Você tem na diagonal principal de uma matriz de confusão os acertos, enquanto nas demais posições confusões entre classes (erros).

Para estimar (VP, VN, FP e FN) em problemas de múltiplas classes podemos considerar uma estratégia um contra todos, na qual cada classe seria em dado momento considerada (+) e as demais como (-).

A **precisão (precision)** tenta responder a seguinte questão "qual a proporção de classificações positivas estão corretas?". Um modelo que não produz falso positivo (FP) tem precisão igual a 1.

A **revocação (recall)** tenta responder a seguinte questão: "qual a proporção das amostras positivas foram classificadas corretamente?" Um modelo que não produza falso negativo (FN) tem revocação igual a 1.

f-mesure (f1-score): consiste na média harmônica entre precisão e revocação, variando de 0 a 1 (sendo 1 perfeito).

Figura 4: Matriz confusão para problema binário, duas classes (+) e (-)

Classe Predita (saída do classificador)			
Classe real (rótulo)		(+)	(-)
	(+)	VP	FN
	(-)	FP	VN

- VP (Verdadeiro Positivo): corretamente classificados
- VN (Verdadeiro Negativo): corretamente classificados
- FP (Falso Positivo): erroneamente classificados como positivos
- FN (Falso Negativo): erroneamente classificados como negativos

Taxa de acerto (Accuracy) = (TP+TN) / (TP+TN+FP+FN)

Precisão (Precision): P = TP / (TP+FP)

Revocação (Recall): R = TP / (TP + FN)

F-measure: 2/(1/P+1/R)

Figura 5: Matriz de confusão para problema multiclasses (10 classes no exemplo, de 0 a 9)

	0	1	2	3	4	5	6	7	8	9
0	988	0	0	0	0	0	3	0	2	0
1	0	986	0	0	0	0	1	0	0	0
2	1	8	993	3	0	0	0	21	0	0
3	1	0	1	995	0	7	0	12	1	2
4	2	2	1	0	983	0	0	2	0	23
5	0	1	0	1	0	971	24	1	0	1
6	6	1	0	0	2	0	966	0	1	0
7	0	1	4	0	1	0	0	961	0	1
8	2	1	1	1	0	12	6	2	995	9
9	0	0	0	0	14	10	0	1	1	964

Taxa_Acerto (Acurácia_Global) = Total_Acertos / Total_Exemplos, sendo o Total_Acertos a soma da diagonal principal da matriz.

3.2) No caso de um regressor: as medidas mais utilizadas são:

- **coeficiente de correlação** (r), varia de -1 a 1 (-1: forte correlação negativa; 0: não existe correlação; 1: forte correlação positiva).
- **coeficiente de determinação** (r2): varia entre 0 e 1, indicando o quanto o modelo consegue explicar os valores observados. quanto maior o r2, mais explicativo é o modelo com relação à amostra.

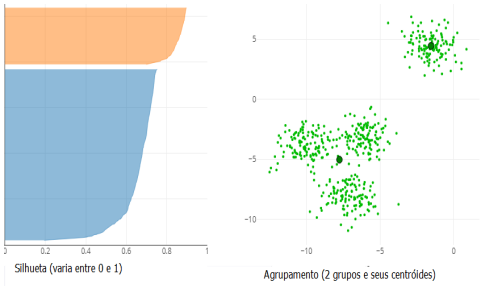
3.3) No caso de agrupamento: temos a variância intra e inter-grupos (clusters) e a medida da silhueta (varia entre -1 e 1).

1: indica ponto interno ao cluster e longe da borda (fronteira) entre grupos (clusters)

0: indica ponto na fronteira entre grupos ou bem próximo dela.

-1: indica ponto associado a cluster errado.

Figura 6: Índice silhueta para avaliar agrupamento (2 clusters no exemplo)



Fonte: O autor

Conceitos traduzidos com base em:

- Baranauskas J. A., Monard M. C. **Reviewing some Machine Learning Concepts and Methods**. Relatório Técnico do Instituto de Ciências Matemáticas e de Computação (ICMC). São Carlos (SP). Fev/2000. Disponível em http://www.ppgia.pucpr.br/~alceu/am/1%20-%20Introduction%20to%20ML/MainConcepts.pdf.