Projeto 1 FSI

***Resumo*—Relatório do projeto 1 da disciplina de Fundamentos de Sistemas Inteligentes.**

***Index Terms*—Aprendizagem de máquina, dados rotulados e não rotulados, regressão, classificação, algoritmos e resultados.**

I. Introdução

Este documento relato sobre o estudo de aprendizagem de máquina, seus subtipos, função de transferência, métodos de classificação e particularmente do uso de dois algoritmos para classificar números manuscritos,o LDA(linear discriminant analysis) e o K-nn(k vizinhos mais próximos) e analisar os seus resultado a partir de testes.

II. Conceitos Estudados

A. Aprendizagem de Máquina

Aprendizagem de máquina pode ser explicada como a capacidade de computadores de aprender e detectar padrões a partir da análise de dados e, com isso tomar decisões sozinhas. Com isso é possível fazer predições pode ser usado em diversas áreas, não só da computação, isso faz com que o estudo de aprendizagem de máquina seja de extrema relevância. Ela pode ser dividida em 3 ramos, supervisionada, não supervisionada e semi-supervisionada.

B. Aprendizagem Supervisionada

Na aprendizagem supervisionada o computador passa por uma fase de "treino", onde ele analisa dados rotulados, sabendo a entrada e a saída esperada, tentando achar a função de transferência, para depois poder fazer suas predições a partir desse treino. Como exemplo desse modelo de aprendizagem temos o algoritmo dos k vizinhos mais próximos, LDA (linear discriminant analysis) e regressão lógica.

C. Aprendizagem não Supervisionada

Diferentemente da supervisionada, no modelo não supervisionado o computador não passa pela fase de "treino", a partir somente de dados não rotulados ele tenta achar algum tipo de padrão ou semelhança entre eles. como exemplos temos mineração de dados e a rede neural artificial, que utiliza um modelo baseado na estrutura neural de organismos inteligentes que aprendem com experiência.[1]

D. Aprendizagem semi-Supervisionada

Nesse modelo é usada uma junção entre os dois modelos anteriores, utilizando em parte dados rotulados e outra parte de dados não rotulados, sendo a maior parte de dados não rotulados, pois é mais fácil e rápido obter dados não rotulados.

[1]http://conteudo.icmc.usp.br/pessoas/andre/research/neural/

E. Regressão e Classificação

Regressão é uma subcategoria de aprendizagem supervisionada, é utilizada quando a saída tem um aspecto contínuo, tentando prever essa saída dado esse aspecto. Já classificação é utilizada quando é preciso dar um rótulo à entrada, uma classe. Para se avaliar a qualidade do ajuste feito através de regressão pode-se analisar o Erro Médio Quadrático(fig 1.), que quanto menor melhor o ajuste, sendo que ele nunca será negativo. Para classificação usa-se a Taxa de erro, ou a a Taxa de Erro de Bayes caso sejam dados de teste.

.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

Figura 1. Erro Médio Quadrático.

F. Dados Rotulados e não Rotulados

Para a avaliação dos diferente tipos de aprendizagem de máquina é preciso saber diferenciar primeiramente os dados, os chamados de dados rotulados são um conjunto de dados obtidos através de um procedimento definido, onde eles são vinculados a um ou mais rótulos. Por exemplo, as imagens de treino são chamados de dados rotulados pois o conjunto de imagens é vinculado ao número a qual cada imagem corresponde, já as imagens de teste são dados não rotulados pois não são vinculados a nenhum rótulo.

G. Mineração de Dados

Um conceito importante para o assunto abordado é mineração de dados, é um termo relativamente recente que está "em alta"hoje em dia. Mineração de dados consiste em analisar grande quantidade de dados não rotulados e achar padrões, tendências existentes nessa grande massa de dados.

H. Tipos de Problemas

Problemas descritivos, são problemas onde é preciso utilizar mineração e agregação de dados para analisar dados não rotulados. Já para trabalhar com problemas preditivos é preciso utilizar modelos estatísticos e técnicas de previsão para analisar dados rotulados e tentar prever o que acontecera caso analise dados não rotulados. Por fim, para os problemas prescritivos é utilizado otimização e algoritmos de simulação para simular cenários e sugerir a melhor opção dado o cenário.[2]

[2]Slides disponibilizados pelo professor da disciplina em aprender.ead.unb.br/course/view.php?id=52

I. Função de Transferência

Função de transferência se refere a uma função que dada uma certa entrada ela nos dê uma saída de acordo com a entrada vide fig 2., isso é importante pois se é achada essa função, é possível achar a saída dada qualquer valor de entrada,ou pra qualquer que sejam os dados de entrada é possível achar a saída.

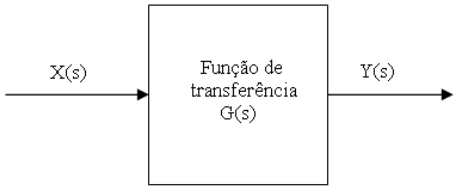

Figura 2. Função de Transferência.

J. Métodos Paramétricos e não Paramétricos

Métodos paramétricos são métodos onde é escolhido uma forma para a função de transferência, já os não paramétricos não assumem uma forma definida para a função de transferência.

III. Classificação de Números Manuscritos

O problema proposto no projeto 1 é ,dadas 70.000 imagens de dígitos manuscritos disponibilizadas, utilizar 60.000 como treino, ou seja, como dados rotulados, e as outras 10.000 como teste, dados não rotulados e utilizar dois algoritmos de classificação supervisionada, o LDA e o K-nn(com 3 valores diferentes para o k), para classificar os dígitos, e depois comparar com as reais classificações e analisar esses resultados.

IV. Algoritmos

A. K-nn

O algoritmo dos k vizinhos mais próximos é um dos algoritmos mais simples de classificação, ele utiliza um método não paramétrico de classificação, onde dado um x qualquer, ele analisa os k vizinhos mais próximos no espaço de atributos(obtidos através dos dados rotulados na fase de "treino") utilizando uma métrica de distância Euclidiana,vista na fig. 3, de Hamming ou discreta, e dependendo da classe dos vizinhos analisados atribui à x a classe de maioria. No caso específico do algoritmo utilizado para os testes neste projeto foi utilizado a métrica de distância Euclidiana. È interessante notar também que dependendo do valor de k o resultado de x pode ser bem variado. Um exemplo pode ser visto na fig. 4 onde se k=3 será atribuído a classe B e se k=6 a classe A.

B. LDA

LDA é um algoritmo de classificação que procura uma combinação linear de variáveis que melhor separa as classes de forma a minimizar a distância dentro das mesmas, ou seja minimizar a matriz de espalhamento de cada classe, enquanto

$$d(\mathbf{p},\mathbf{q}) = d(\mathbf{q},\mathbf{p}) = \sqrt{(q_1-p_1)^2+(q_2-p_2)^2+\cdots+(q_n-p_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n}(q_i-p_i)^2}.$$

Figura 3. Distância Euclidiana.

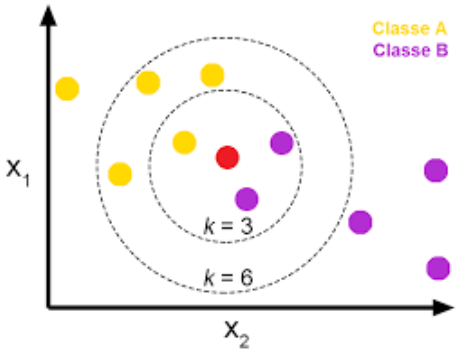

Figura 4. Exemplo k-nn.

que maximiza a distância entre as médias de classes diferentes, maximiza a matriz de espalhamento entre classes.

V. Resultados

Usando os algoritmos de classificação LDA e K-nn com o arquivo teste disponibilizado pelo professor no projeto 1, tivemos bons resultados em termos da acurácia dos algoritmos, no caso dos testes feitos por nós aplicando k=10 obtemos uma precisão de aproximadamente 93,7167%(matriz de confusão pode ser vista na tabela 1), já com k=5 94,2517%(matriz de confusão pode ser vista na tabela 2) e com k=4 94,115%(matriz de confusão pode ser vista na tabela 3), logo é possível observar que k sendo 10 é um valor alto, pois diminuindo pra 5 a acurácia aumenta, mas se abaixarmos ainda mais, para 4 a acurácia acaba caindo também, logo 5 é um bom valor para k. Já com o LDA a precisão fica em torno de 85%(matriz de confusão pode ser vista na tabela 4), valor mais baixo que com o K-nn, logo para esse problema o K-nn se mostrou mais preciso.

Tabela I
MATRIZ DE CONFUSÃO DE K-NN COM K=10

	0	1	2	3	4	5	6	7	8	9
0	5842	7	5	3	3	16	34	1	5	7
1	1	6682	19	7	11	1	4	13	1	3
2	86	178	5419	39	20	11	25	149	18	13
3	14	76	64	5722	3	68	7	71	52	54
4	7	118	6	1	5448	4	23	17	2	216
5	28	56	5	145	17	5003	84	10	13	60
6	51	45	2	2	5	39	5769	1	4	0
7	6	155	11	1	39	10	2	5953	0	88
8	28	239	39	169	39	168	47	38	4966	118
9	31	50	9	76	92	13	5	233	14	5426

Classes Reais x Classes Obtidas pelo K-nn

Tabela II
MATRIZ DE CONFUSÃO DE K-NN COM K=5

	0	1	2	3	4	5	6	7	8	9
0	5846	8	7	3	2	10	34	2	5	6
1	0	6682	23	3	11	0	3	15	1	4
2	74	140	5487	37	15	10	19	146	17	13
3	15	50	80	5744	1	78	5	64	54	40
4	7	93	6	1	5463	1	29	17	3	222
5	28	39	4	138	15	5032	86	9	20	50
6	43	36	4	0	6	38	5787	0	4	0
7	8	111	16	2	45	7	1	5982	1	92
8	30	187	43	150	43	156	39	38	5073	98
9	31	41	9	71	102	24	4	199	13	5455

Classes Reais x Classes Obtidas pelo K-nn

Tabela III
MATRIZ DE CONFUSÃO DE K-NN COM K=4

	0	1	2	3	4	5	6	7	8	9
0	5860	6	5	1	2	11	26	2	6	4
1	0	6680	23	5	14	0	2	16	1	1
2	90	141	5507	36	11	9	15	120	14	15
3	14	51	83	5748	0	74	6	61	56	38
4	8	95	8	0	5551	1	22	19	2	136
5	31	40	4	180	17	4998	82	5	21	43
6	56	43	3	0	6	37	5766	0	7	0
7	4	112	22	5	47	4	1	6003	2	65
8	38	187	71	177	44	177	43	32	4994	88
9	36	39	11	68	154	23	3	239	14	5362

Classes Reais x Classes Obtidas pelo K-nn

Tabela IV
MATRIZ DE CONFUSÃO DO LDA

	0	1	2	3	4	5	6	7	8	9
0	5522	5	35	54	39	105	37	3	110	13
1	0	6510	48	21	23	33	6	14	78	9
2	78	198	4812	185	136	30	180	84	213	42
3	23	99	236	4954	41	251	19	127	210	171
4	9	50	43	7	5067	48	45	5	97	471
5	64	77	39	310	93	4215	135	31	325	132
6	64	81	88	8	76	137	5302	3	152	7
7	40	137	42	64	210	30	5	5209	34	494
8	29	417	52	225	84	309	30	20	4514	171
9	47	27	15	90	350	35	1	362	77	4945

Classes Reais x Classes Obtidas pelo LDA