Trabalho 3 de Fundamentos de Sistemas Inteligentes - Estudo e Aplicação do Algoritmo de Support Vector Machines para Classificação de Dados de Câncer de Mama

Alex Siqueira Lacerda - 16/0047692, Arthur da Veiga Feitoza Borges - 13/0050725

***Abstract*— Este trabalho visa explicar o algoritmo de aprendizado de máquina Support Vectors Machine *(SVM)*, utilizando-o para classificar um conjunto de dados de diagnósticos de câncer de mama e fazendo análise dos resultados.**

I. INTRODUÇÃO

Uma gigantesca quantidade de dados estão sendo armazenados e disponibilizados na internet para análise ao redor do mundo. Esses dados tem sido coletados ao longo de décadas e, com o advento da era digital, têm sido também digitalizados para aplicações das mais diversas.

Aproveitar de forma inteligente esta grande quantidade de informação é um grande desafio da ciência moderna, em especial: a Ciência da Computação. Foram desenvolvidos nas últimas décadas poderosíssimos algoritmos de aprendizado de máquina, que têm apresentado resultados relevantes na interpretação destes dados, que tem sido excepcionais para o desenvolvimento científico em muitas áreas.

Neste artigo, iremos abordar um desses algoritmos, o *Support Vector Machine (SVM)*, aplicado a um dataset que consiste em dados de diagnósticos de câncer de mama. Estes dados foram obtidos a partir de exames clínicos reais em Hospitais da Universidade de Wisconsin, que foram disponibilizados no banco de datasets da Universidade de Califórnia, Irvine. O link para download está disponível na bibliografia.

A. SVM - Support Vector Machines

Se trata de um algoritmo de classificação que se apoia na intuição inicial de que, quando as classes podem ser separadas por um hiper-plano, uma análise da "fronteira" entre as classes pode ser efetiva para se gerar vetores de suporte que proporcionem o cálculo de tal hiperplano. A consequência disto é que, mesmo com uma grande quantidade de dados, o algoritmo poderia realizar os cálculos mais rapidamente, pois este só analisaria um sub-conjunto das observações, ou seja, só aquelas necessárias para o cálculo dos vetores de suporte.

Uma característica importante são as margens do separador. O algoritmo que inspirou o SVM, o Classificador de Margem Máxima, como o nome sugere, utilizava o raciocínio simples de que seria mais eficiente colocar seu hiperplano exatamente no meio da distância entre as duas classes. Ou seja, o mais longe da "fronteira" de cada uma delas. O SVM estende este conceito para uma margem de tolerância, chamado parâmetro C. Basicamente o classificador estabelece um limite em que amostras de uma classe podem aparecer do outro lado. Isso dá maior robustez ao algoritmo, que passa a conseguir classificar com boa taxa de precisão, mesmo os dados que não sejam tão uniformemente separáveis. Isto possibilita aplicações eficientes para uma maior quantidade de datasets, mesmo aqueles muito grandes e não tão bem distribuídos.

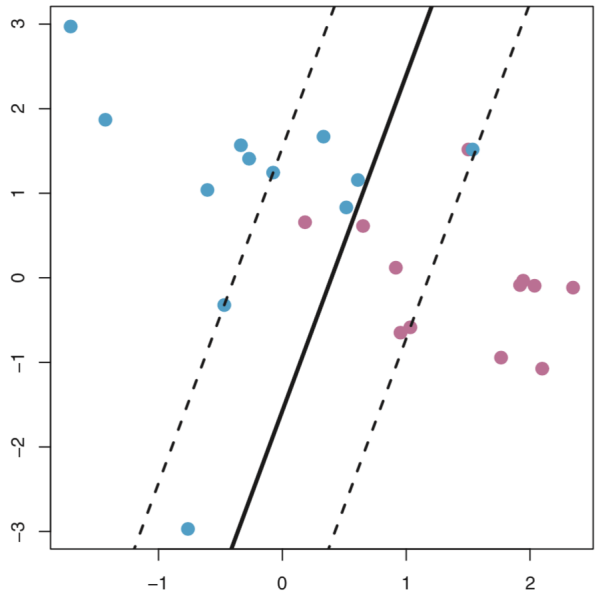

Fig. 1. Funcionamento do parâmetro C. Fonte: [2]

Mas, e quando as classes não são linearmente separáveis?

É muito comum se deparar com conjuntos de dados que não são linearmente separáveis, isto torna mais complexa a aplicação de algoritmos de aprendizado de máquina, devido ao acréscimo de complexidade inerente.

O SVM, para classificação não linear, aumenta o espaço de features, adicionando dimensões não-lineares a partir de equações de kernel. Ou seja, algumas operações não-lineares são feitas sobre as features do espaço de observações. Elas podem ser do tipo polinomial, radial, gaussiano, etc. A partir deste novo espaço de features, o algoritmo realiza a classificação linear da mesma forma que faria no espaço original. Porém, quando transposto para o espaco original, a fronteira de classificação toma feições não-lineares de acordo com as operações que foram feitas.

II. MATERIAL E MÉTODOS

Para cumprirmos com o proposto, foi utilizado o conjunto de amostras contidas no arquivo ”breast_cancer_wisconsin_data.txt”, presente no dataset solicitado. Tal dataset consiste em 699 amostras de diagnósticos de câncer de mama, ordenados cronologicamente, extraídos de relatórios de dados adicionados entre 1990 e 1992. Este dataset representa diferentes casos de câncer analisados e diagnosticados no Hospital da Universidade de Wisconsin.

As amostras contém dados de 2 classes (diagnóstico positivo e negativo), cada uma das amostras contendo 12 features, que representam características das células mamárias das pacientes submetidas aos testes. O objetivo deste trabalho é encontrar uma configuração do algoritmo Support Vector Machine que apresente melhor acurácia e menor taxa de erro dentre os sugeridos, utilizando-se estes dados.

Antes de realizar a classificação dos dados, notamos que algumas amostras possuíam features inexistentes. Na documentação dos dados aparecia que se tratavam de ”missing attribute values” (atributos inexistentes). Trocamos todos os seus valores por 0, para que não tivessem influência sobre os resultados.

Feito isso, definimos os valores de C em escala logarítmica na base 2, com range de 2^{-5} a 2^{15}, num vetor de 200 posições, para obter melhor resolução nos resultados.

Para validação cruzada, utilizamos a classe da biblioteca scikit-learn *StratifiedShuffleSplit* na configuração 70/30 (70% de dados de treinamento e 30% de dados de teste). A classe foi utilizada em conjunto com a classe *GridSearchCV*, que permite a comparação entre classificadores, com grande flexibilidade para alteração de parâmetros de cada classificador.

Utilizando este mecanismo, fizemos análise exaustiva de vários classificadores SVM, utilizando a implementação da biblioteca scikit-learn. Através do parâmetro *kernel* da classe *svm.SVC* selecionamos os tipos de funções não-lineares que seriam aplicadas ao nosso conjunto de dados. Seguem as configurações de kernel utilizadas para cada teste:

- *svm.LinearSVC()* - Classificador SVM linear;
- *svm.SVC(kernel=”rbf”)* - Classificador SVM com kernel de função de base radial (gaussiano)
- *svm.SVC(kernel=”poly”)* - Classificador SVM com kernel polinomial com coeficiente constante;
- *svm.SVC(kernel=”sigmoid”)* - Classificador SVM com kernel sigmóide.

Por fim, todos estes classificadores foram treinados para os 200 valores diferentes de margem C especificados no início desta seção. Plotamos os resultados usando o pyplot, da biblioteca matplotlib.

Todos os códigos utilizados neste trabalho foram feitos pelos integrantes do grupo e estão disponíveis em `https://github.com/arthurveiga/svm_mammography_dataset`

Uma grande ajuda foi obtida observando o código explicativo do próprio scikit-learn que pode ser encontrado em `http://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html`

III. RESULTADOS E ANÁLISE

Nesta seção vamos dividir por tipo de kernel e logo depois fazer um comparativo geral, colocando em escopo principal o erro, a acurácia e a precisão. Os seguintes kernels foram adotados:

- Linear;
- Função de Base Radial (Gaussiano);
- Função Polinomial;
- Função Sigmóide.

A. SVM Linear

Sabemos que o SVM Linear utiliza-se de um hiperplano para separar os dados na forma que se encontram no dataset. Os resultados obtidos foram muito bons, com acurácia e precisão de 96% e erro de apenas 6.9%, porém, com um C mais alto que os resultados seguintes. Isto indica que os dados devem estar dispostos de maneira razoavelmente separável linearmente.

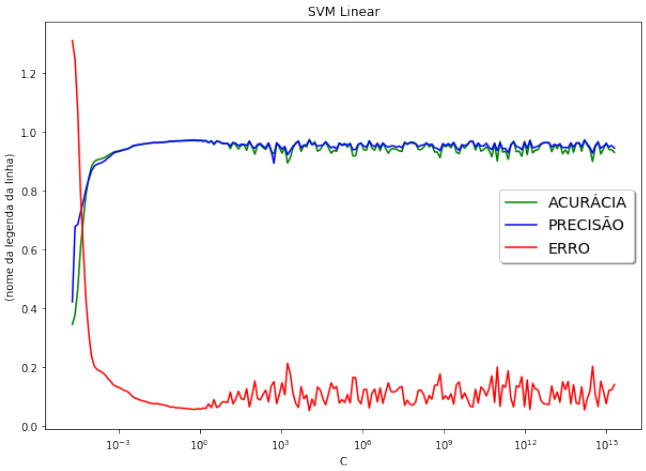

Fig. 2. SVM Linear: Gráfico de erro, acurácia e precisão.

- Maior acurácia: 0.965 referente ao C: 8.49
- Maior precisão: 0.961 referente ao C: 2.67
- Menor erro: 0.069 referente ao C: 8.49
- Tempo médio de fit: 6.89ms

B. SVM de Função de Base Radial (FBR)

Tendo o FBR como kernel do SVM, no caso da implementação do sklearn, funções do tipo Gaussiana são aplicadas às features do dataset, para que depois sejam classificadas. Isto faz com que a fronteira de classificação obtenha feições de acordo com a função gaussiana.

Os resultados obtidos também apresentaram resultados muito bons. Com acurácia e precisão acima de 93% e erro de 9%, com C próximo de zero. O tempo de fit também foi baixo. Isto indica que, como para C = 8 o classificador linear funcionou de maneira ligeiramente melhor, os dados, para uma classificação menos permissiva em relação à margem de separação (C próximo de zero), podem ser ajustados para

a função gaussiana, porém implicando em pequena perda de acurácia e precisão.

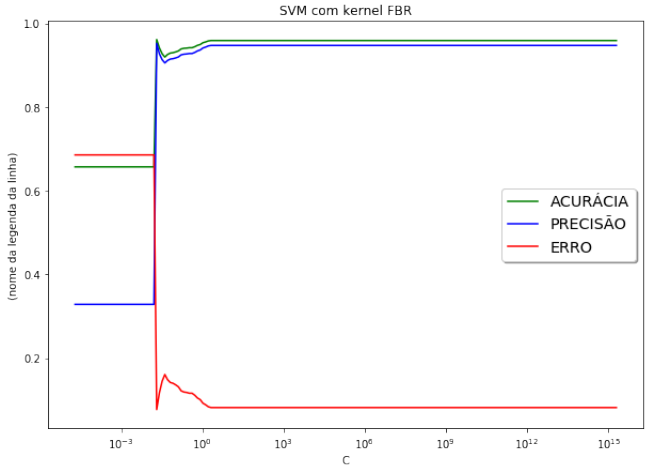

Fig. 3. SVM FBR: Gráfico de erro, acurácia e precisão.

- Maior acurácia: 0.950 referente ao C: 0.020
- Maior precisão: 0.938 referente ao C: 0.020
- Menor erro: 0.098 referente ao C: 0.020
- Tempo médio de fit: 4.79ms

C. SVM de Kernel Polinomial

No caso de kernel Polinomial, funções polinomiais são aplicadas às features do dataset. Aqui o coeficiente foi fixado em 2 (quadrático). Os resultados obtidos foram novamente muito bons. Com precisão e acurácia de 96%, e erro de 7%, para valores de C aidna mais próximos de ainda mais baixos que o Classificador FBR. Porém, o tempo médio de fit é bem mais alto.

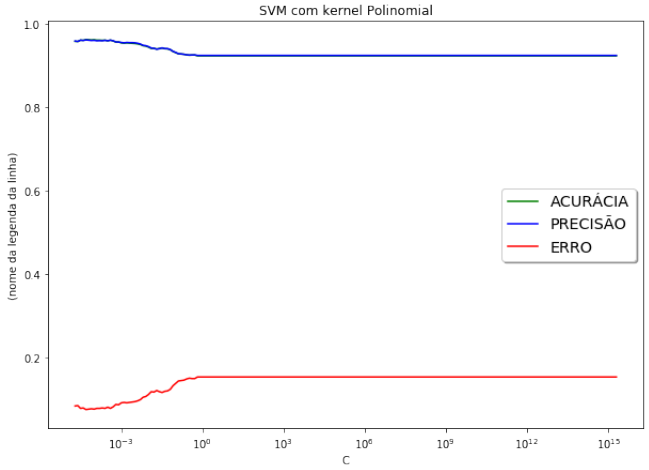

Fig. 4. SVM Polinomial: Gráfico de erro, acurácia e precisão relacionados.

- Maior acurácia: 0.962 referente ao C: 0.0002
- Maior precisão: 0.962 referente ao C: 0.0002
- Menor erro: 0.074 referente ao C: 0.0002
- Tempo médio de fit: 11.25ms

D. SVM de Função Sigmóide

Para o SVM de kernel Sigmóide, os espaço de features é aumentado aplicando-se funções sigmóides a ele. Aqui, os resultados foram de performance bem pior que os anteriores. Acurácia e precisão baixas, e erro alto. Os melhores valores encontrados foram referentes ao C: 2^{-5}, que é o menor valor de teste. Portanto, conforme o valor de C ia aumentando, os resultados foram ficando ainda piores. Isto indica um completo desajuste entre os dados e o classificador pela função sigmóide. Portanto, os resultados indicam que o classificador deve ser descartado.

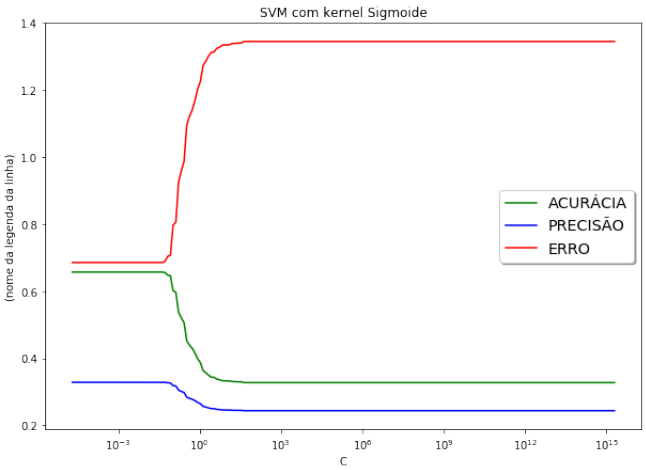

Fig. 5. SVM Sigmóide: Gráfico de erro, acurácia e precisão.

- Maior acurácia: 0.657 referente ao C: 2^{-5}
- Maior precisão: 0.328 referente ao C: 2^{-5}
- Menor erro: 0.68 referente ao C: 2^{-5}
- Tempo médio de fit: 6.91ms

E. Comparação entre os Kernels

TABLE I
COMPARAÇÃO ENTRE OS CLASSIFICADORES SVM

Classificador	**Acurácia**	**Precisão**	**Erro**
Linear	**96.5%**	96.1%	**6.9%**
RBF (Gaussiano)	95%	93.8%	9.8%
Polinomial	96.2%	**96.2%**	7.4%
Sigmóide	65.7%	32.8%	68%

Nota-se que os classificadores Gaussiano (RBF), Polinomial e Linear apresentaram desempenho parecido. Porém, de maneira geral, pode-se dizer que o Classificador Linear foi o que obteve melhores resultados. O valor de C para o qual o classificador linear obteve os melhores resultados, foi próximo de 8, enquanto para os outros classificadores, o melhor C encontrado foi próximo de zero. Isto é uma clara demonstração de como o uso desta margem pode adicionar robustez e eficiência ao algoritmo, visto que o aumento da margem para o classificador linear gerou os melhores resultados de todo o teste.

Um detalhe a ser salientado é que os classificadores não-lineares apresentaram comportamento anômalo para valores

de C muito baixos e/ou muito altos, como pode se ver nos gráficos, na forma de linhas retas no início ou final. Isto indica que a implementação do SVM utilizada possui certas restrições quanto ao tamanho da margem dependendo do kernel utilizado. Entretanto, os valores de maior precisão/acurácia e menor erro, para os kernels não-lineares, se apresentaram sempre para valores de C dentro da faixa em que o comportamento da função era normal. Porém, isto põe em dúvida os valores de tempo médio de fit, visto que, caso tenha havido algum problema no ajuste das funcões para valores ”extremos” de C, os tempos podem ter sido muito curtos ou muito longos, o que enviesaria o resultado do teste para o valor de tempo médio de fit.

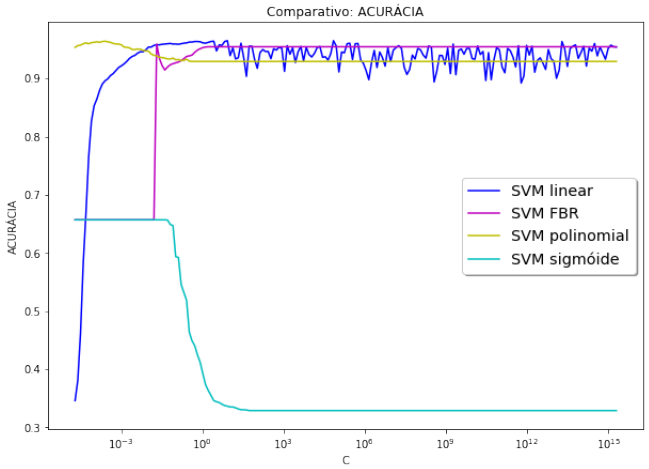

Fig. 6. Acurácia para todos os kernels.

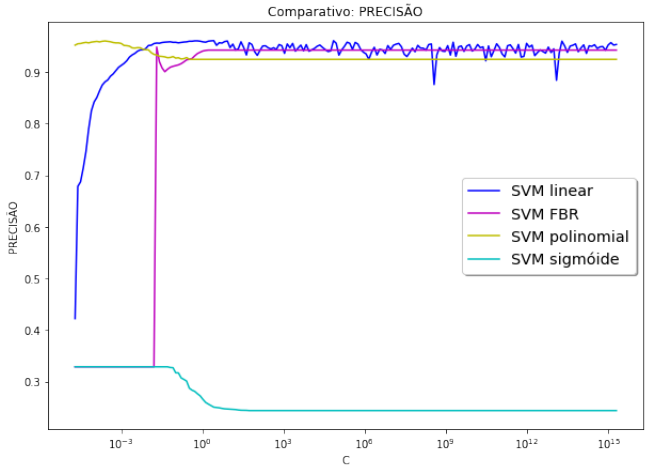

Fig. 7. Precisão para todos os kernels.

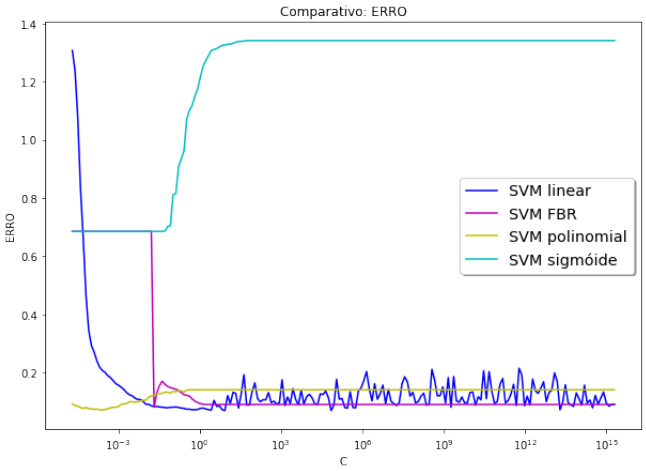

Fig. 8. Taxa de erro para todos os kernels.

IV. CONCLUSÃO

De forma geral, e motivados pelos resultados obtidos e comentados na seção anterior, concluimos que existe grande probabilidade de os dados serem melhor adequados à uma separação linear. Kernels não-lineares (com margens próximas de zero) também obtiveram resultados muito bons. Porém, como o melhor resultado foi obtido pelo kernel linear, mesmo que utilizando uma margem de tamanho consideravelmente maior, concluimos que o melhor classificador para tais dados é o Linear. Não só pelos melhores resultados, mas também por ser o mais simples e, intuitivamente, o mais rápido. Acreditamos que este trabalho foi uma ótima experiência que nos mostrou, na prática, a vantagem da utilização da margem C, e da robustez do algoritmo SVM. Devido a isso, concluímos que o SVM é um ótimo meio para efetiva classificação de dados, também por causa da capacidade de tratar os dados de forma linear e não-linear e da robustez que ela oferece.

REFERENCES

[1] DATASET - Breast Cancer Wisconsin (Diagnostic) Data Set: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

[2] An Introduction to Statistical Learning - Gareth James, Daniela Witten; Springer NYH, Dordrecht London

[3] http://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html