Experimentos em classificação binária com SVM e KNN

Leonardo Azeredo
Universidade Federal do Rio de Janeiro Email: azeredol@cos.ufrj.br

***Abstract*—Este trabalho explora experimentalmente o desempenho de diferentes configurações de hiper parâmetros para os modelos de SVM e KNN em classificação binária. Os experimentos foram realizados sobre uma das bases de dados do KEEL, conhecida como *banana data set*, uma base com dois atributos e duas classes.**

1. Introdução

Em problemas de aprendizado de máquina, uma tarefa sempre presente é a seleção de modelos de melhor desempenho. Um técnica conhecida como validação (e uma variação desta chamada e validação cruzada) é a técnica de escolha para executar essa tarefa.

Quando se fala em seleção de modelos, isso se refere ao conceito mais amplo possível de modelo, incluindo desde técnicas diferentes, como redes neuras, SVM, KNN, regressão, etc, mas também diferentes configurações dos hiper parâmetros de cada uma das técnicas diferentes.

Neste trabalho, a técnica de validação cruzada K-fold foi utilizada para avaliar diferentes configurações do modelo SVM e do modelo KNN para a classificação binaria de uma base de dados conhecida como KEEL e determinar que modelo possui o melhor desempenho em termos de erro fora da amostra estimado.

Setembro, 2018

2. Modelos de aprendizagem utilizados

Neste trabalho foram utilizados dois modelos distintos para classificação. Uma breve descrição destes encontra-se nas seções a seguir.

2.1. SVM

Máquinas de vetores de suporte é uma técnica de classificação desenvolvida para classificação binária por [5] em 1995 que visa encontrar o hiperplano ótimo separando duas classes através do conceito de maximização da margem entre os pontos mais próximos entre as classes onde os pontos do lado errado da margem são ponderados para reduzir sua influencia. Todo o problema pode ser formulado como um problema de programação quadrática [4].

Quando não é possível encontrar um separador linear os pontos são projetados para espaços de dimensões mais altas, através de técnicas conhecidas como *SVM Kernel*.

2.2. KNN

A técnica de K vizinhos mais próximos é um método que pode ser usado para classificação onde as entradas consistem do K exemplos da amostra mais próximos segundo o espaço de atributos e a saída é uma associação a uma das classe. Assim, os exemplos novos podem ser classificações de acordo com essa associação.

2.3. Validação cruzada *K-Fold*

A validação cruzada é uma técnica de validação de modelo usada para avaliar como os resultados de uma análise estatística vão generalizar para quando considerado um conjunto de dados qualquer. A técnica consiste em executar múltiplas instâncias distintas de um modelo em fragmentos diferentes de uma amostra, usando a parte que não participa o treinamento em cada rodada para estimar o erro fora da amostra. O erro de validação final retornado, então, é a média dos erros de todas as rodadas, enquanto o modelo retornado é treinado com todos os dados disponíveis [3].

A técnica de validação cruzada ideal consideraria $N-1$ (onde N é o tamanho da amostra) elementos da amostra por treinamento, mas em conjuntos de dados reais isso seria inviável. Assim, uma variação comum é conhecida como validação cruzada *K-fold*, onde os da dos são divididos em K conjuntos disjuntos e, a cada rodada, cada um deles é usado para determinar um erro de validação [3]. Em termos práticas, o valor de K deve ser tão grande quanto possível dentro das limitações de tempo.

3. Descrição dos experimentos e resultados

O objetivo dos experimentos foi avaliar a performance de diferentes configurações de parâmetros para o modelo de classificação usando SVM e KNN sobre a base de dados binária KEEL-dataset.

Adicionalmente, houve o interesse de analisar o impacto que diferentes valores para o parâmetro k na técnica de validação cruzada *K-fold* tem sobre os resultados obtidos com o uso do SVM. Para isso, para os experimentos com esse modelo foram utilizados três valores distintos: 3, 5 e 10.

Diante da análise dos resultados de diferentes kernels durante os experimentos com o SVM, notou-se que kernel

RBF produziu resultados consideravelmente melhores do que os outros. Por isso, diante na natureza deste kernel que leva em consideração a influencia de um ponto sobre os outros em função de sua distância, foi selecionado o modelo KNN para comparação, já que este modelo também tem inerentemente esta propriedade. Para a comparação com o modelo SVM, apenas o valor 10 para K na validação cruzada foi utilizado nos experimentos com o KNN.

3.1. KEEL-dataset

A base de dados utilizada possui 5300 pontos com apenas dois atributos de valores reais e duas classes e é conhecida como *banana data set* [2] em virtude da aparência dos dados quando plotados em um gráfico de dispersão tal como o da figura 1.

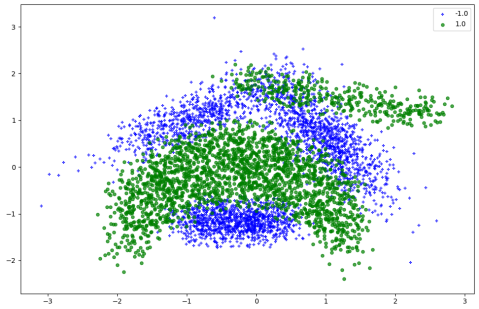

Figure 1. Dispersão do *banana data set* com suas classes demarcadas.

Avaliando o gráfico de dispersão dos dados é possível notar as duas características mais marcantes desta amostra. Primeiramente os dados são claramente não separáveis linearmente no espaço sem algum tipo de transformação não linear. A segunda característica é o fato e que, aparamente, existe uma grande quantidade de ruído ao menos em parte da amostra, como se pode notar na seção superior direita do gráfico onde existe uma grande mistura entre pontos de diferentes classes.

3.2. Estimativa artificial do erro fora da amostra

Apara tornar possível a avaliação do impacto de diferentes valores para o parâmetro K da validação cruzada na estimativa do erro de validação é necessário ter como referencial o uma estimativa sem viés do erro fora da amostra [1].

Para isso, o conjunto de dados da amostra foi dividido e 20% dos pontos (1060 pontos) foi selecionado aleatoriamente para servir ao objetivo de gerar um valor de referencia para o erro fora da amostra de cada modelo. Assim, todos os experimentos de validação cruzada descritos a seguir foram realizados sob apenas 80% do total de pontos (4240 pontos), apenas para permitir essa comparação teórica entre diferentes valores de K.

É evidente que se este não fosse parte dos objetivos, em uma situação real de seleção de modelos, 100% dos dados seriam usados para a validação cruzada/treinamento e o valore de K seria escolhido sempre como o maior possível dentro dos limites de tempo de uma aplicação prática.

3.3. Implementação

A implementação de deu utilizando a linguagem Python 3.5 e a biblioteca para aprendizado de máquina popular *scikit-learn*. Um componente dinâmico da biblioteca chamado *GridSearchCV*, que automaticamente realiza a validação cruzada de múltiplas combinações de parâmetros para um modelo com o número de *folds* especificado e reporta resultados detalhados, foi utilizada para facilitar os experimentos.

3.4. Seleção de hiper parâmetros para o SVM

Entre os diferentes hiper parâmetros selecionáveis no modelo SVM o principal é o tipo de kernel usado para representar a transformação não linear dos dados. Cada um destes por sua vez admitem uma série de parâmetros próprios e, para os experimentos realizados foram utilizadas as seguintes combinações: para o kernel sigmoide o parâmetro $\gamma \in [1, 0.5, 0.01]$, para o kernel linear não existem parâmetros adicionais, para o kernel polinomial o grau $\in [2, 3, 4, 5, 6, 7, 8, 9, 10]$ e para o kernal RDF o parâmetro γ foi mantido no *default* igual a $1/N$. Para todos os outros parâmetros possíveis foram mantidos os valores padrão das ferramentes usadas na implementação.

Além disso, para cada combinação, foi executada a técnica de validação cruzada e produzido as estimativa do erro de validação para valores de 3, 5 e 10 no *K-fold*. Os resultados estão detalhados nas tabelas 1, 2 e 3, onde os melhores e piores valores para o erro de validação estão sublinhados.

Os resultados não mostram nenhuma mudança significativa com relação ao erro de validação com diferentes valores de K, estando idênticos no melhor caso para os três valores usados e, no pior caso apresentando apenas uma pequena alteração quando K = 3.

K = 3	
Hiper parâmetros	**E_val**
{'gamma': 0.01, 'kernel': 'sigmoid'}	0.448
{'gamma': 0.5, 'kernel': 'sigmoid'}	0.700
{'gamma': 1, 'kernel': 'sigmoid'}	0.711
{'kernel': 'linear'}	0.448
{'kernel': 'poly', 'degree': 2}	0.332
{'kernel': 'poly', 'degree': 3}	0.359
{'kernel': 'poly', 'degree': 4}	0.369
{'kernel': 'poly', 'degree': 5}	0.421
{'kernel': 'poly', 'degree': 6}	0.366
{'kernel': 'poly', 'degree': 7}	0.423
{'kernel': 'poly', 'degree': 8}	0.373
{'kernel': 'poly', 'degree': 9}	0.420
{'kernel': 'poly', 'degree': 10}	0.371
{'kernel': 'rbf'}	0.100

TABLE 1. PARÂMETROS AVALIADOS COM K=3 E AS ESTIMATIVAS DE ERRO FORA DA AMOSTRA. O MELHOR E O PIOR VALOR PARA E_{val} ESTÃO MARCADOS.

K = 5	
Hiper parâmetros	**E_val**
{'gamma': 0.01, 'kernel': 'sigmoid'}	0.448
{'gamma': 0.5, 'kernel': 'sigmoid'}	0.699
{'gamma': 1, 'kernel': 'sigmoid'}	0.713
{'kernel': 'linear'}	0.448
{'kernel': 'poly', 'degree': 2}	0.333
{'kernel': 'poly', 'degree': 3}	0.360
{'kernel': 'poly', 'degree': 4}	0.367
{'kernel': 'poly', 'degree': 5}	0.421
{'kernel': 'poly', 'degree': 6}	0.365
{'kernel': 'poly', 'degree': 7}	0.423
{'kernel': 'poly', 'degree': 8}	0.375
{'kernel': 'poly', 'degree': 9}	0.421
{'kernel': 'poly', 'degree': 10}	0.371
{'kernel': 'rbf'}	0.100

TABLE 2. PARÂMETROS AVALIADOS COM K=5 E AS ESTIMATIVAS DE ERRO FORA DA AMOSTRA. O MELHOR E O PIOR VALOR PARA E_{val} ESTÃO MARCADOS.

K = 10	
Hiper parâmetros	**E_val**
{'gamma': 0.01, 'kernel': 'sigmoid'}	0.448
{'gamma': 0.5, 'kernel': 'sigmoid'}	0.701
{'gamma': 1, 'kernel': 'sigmoid'}	0.713
{'kernel': 'linear'}	0.448
{'kernel': 'poly', 'degree': 2}	0.332
{'kernel': 'poly', 'degree': 3}	0.358
{'kernel': 'poly', 'degree': 4}	0.366
{'kernel': 'poly', 'degree': 5}	0.421
{'kernel': 'poly', 'degree': 6}	0.367
{'kernel': 'poly', 'degree': 7}	0.422
{'kernel': 'poly', 'degree': 8}	0.376
{'kernel': 'poly', 'degree': 9}	0.421
{'kernel': 'poly', 'degree': 10}	0.372
{'kernel': 'rbf'}	0.100

TABLE 3. PARÂMETROS AVALIADOS COM K=10 E AS ESTIMATIVAS DE ERRO FORA DA AMOSTRA. O MELHOR E O PIOR VALOR PARA E_{val} ESTÃO MARCADOS.

As tabelas 4, 5 e 6 mostram os resultados da avaliação do erro fora da amostra do conjunto de teste do pior e do melhor modelo. É possível notar que independente do valor de K o desempenho real fora da amostra dos modelos selecionados é equivalente.

Além disso, quando observada a relação entre o erro de validação de cada versão do *K-fold* e o erro de teste, as três instâncias de melhor caso foram, também, equivalentes.

Quando consideramos o modelo de pior caso com cada valor de K, houve equivalência quando K = 5 e K = 10. Já quando K = 3 nota-se uma variação ligeiramente maior.

Tal como era de se esperar pela teoria, o erro de validação dos modelos de pior caso são ligeiramente otimistas em relação ao erro fora da amostra real. Também como esperado, a qualidade da estimativa do erro fora da amostra (quando medida em termos de diferença absoluta entre os valores da validação e do teste) é maior quando se aumenta o valor de K e, para K = 3, esta é de fato menor do que com os outros valores. Nestes casos não houve diferença visível entre tal qualidade entre os valores 5 e 10 para K.

Já nos melhores casos aconteceu uma anomalia onde o erro de validação apresentou resultados práticos pessimistas em comparação com o erro de teste. Acredita-se que isso possa ser explicado pela natureza estocástica de todo o processo, inclusive quando da criação do conjunto de teste.

K = 3	
Configuração	**E_out**
{'gamma': 1, 'kernel': 'sigmoid'}	0.73
{'kernel': 'rbf'}	0.091

TABLE 4. ESTIMATIVA DO ERRO FORA DA AMOSTRA QUANDO K = 3 PARA O PIOR E O MELHOR CASO ESTUDADO.

K = 5	
Configuração	**E_out**
{'gamma': 1, 'kernel': 'sigmoid'}	0.73
{'kernel': 'rbf'}	0.091

TABLE 5. ESTIMATIVA DO ERRO FORA DA AMOSTRA QUANDO K = 5 PARA O PIOR E O MELHOR CASO ESTUDADO.

K = 10	
Configuração	**E_out**
{'gamma': 1, 'kernel': 'sigmoid'}	0.73
{'kernel': 'rbf'}	0.091

TABLE 6. ESTIMATIVA DO ERRO FORA DA AMOSTRA QUANDO K = 10 PARA O PIOR E O MELHOR CASO ESTUDADO.

A figuras 2 e 3 apresentam os gráficos de dispersão dos dados usados no trenamento do pior e do melhor modelo (quando K = 10), com as regiões de decisão correspondentes aproximadas e os pontos correspondentes ao vetores de suporte encontrados marcados. Os gráficos permitem confirmar visualmente e qualidade superior de um modelo em relação ao outro, já indicada pelos resultados numéricos.

Quando observados o número de vetores de suporte encontrados por cada um dos modelos, observa-se a relação proporcional entre o desempenho do modelo e a razão $\#VS/(N-1)$ (0.714 no pior caso e 0.293 no melhor), onde $\#VS$ (3028 no pior caso e 1242 no melhor) corresponde ao número de vetores e N (4240 para o conjunto de treinamento) ao tamanho da amostra. O valor esperado de tal razão corresponde a um limite superior para o erro fora da amostra e, nestes experimentos, quando comparados com os valores de uma única rodada com os valores de teste não nota-se nenhuma discrepância significativa (no melhor caso o erro de teste foi bem menor que o limite e no pior o superou, mas com uma margem pequena).

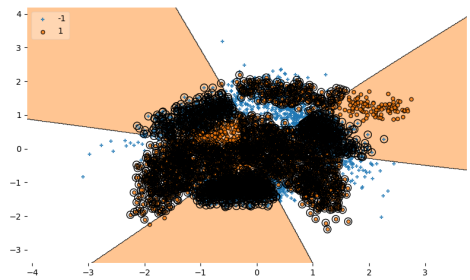

Figure 2. Gráfico com as regiões de decisão e os 3028 vetores de suporte marcados para a prior configuração do SVM.

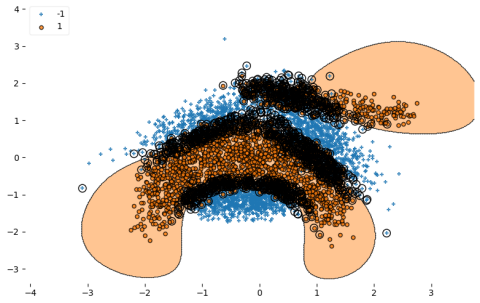

Figure 3. Gráfico com as regiões de decisão e os 1242 vetores de suporte marcados para a melhor configuração do SVM.

3.5. Seleção de hiper parâmetros para o KNN

Para os experimentos com o modelo KNN, foi realizada validação cruzada com K = 10 e entre os hiper parâmetros possíveis, foram-se mantidos os padrões da ferramente utilizada para todos com exceção de dois: o método de calculo da influência dos pontos entre si foi fixado para ponderar pela distância e o número de vizinhos a serem considerados foi determinado pelo conjunto $[5, 10, 15, 20, 25, 30, 35, 4045, 50]$. A tabela 7 mostra os resultados do erro de validação de cada modelo, onde foram sublinhados o pior e o melhor caso.

K = 10	
Número de vizinhos	**E_val**
5	0.109
10	0.107
15	0.106
20	0.104
25	0.102
30	0.101
35	0.098
40	0.099
45	0.101
50	0.102

TABLE 7. NÚMERO DE VIZINHOS AVALIADOS COM K=10 E AS ESTIMATIVAS DE ERRO FORA DA AMOSTRA. O MELHOR E O PIOR VALOR PARA E_{val} ESTÃO MARCADOS.

A figura 4 mostra as regiões de decisão correspondentes ao melhor modelo avaliado na validação cruzada (com 35 vizinhos considerados).

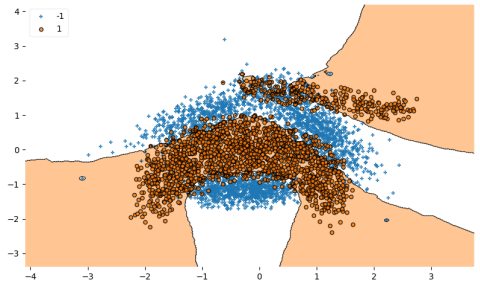

Figure 4. Gráfico com as regiões de decisão para a melhor configuração do KNN.

A tabela 8 mostra os valores da estimativa de erro fora da amostra representada pelo erro no conjunto de teste do pior e do melhor caso dos modelos KNN selecionados. De fato, o valor do erro de validação, também nestes experimentos, se mostrou muito próximo do erro de teste.

K = 10	
Número de vizinhos	**E_out**
5	0.109
35	0.096

TABLE 8. ERRO NO CONJUNTO DE TESTE PARA O MELHOR E O PIOR CASO DO MODELO KNN.

4. Resultados comparativos finais

Quando comparados os desempenhos do melhor modelo do tipo SVM selecionado e do melhor modelo do tipo KNN nota-se que quando utilizada a métrica de acurácia estes são muito próximos, com o Kernel SVM tendo um erro de teste $0,005$ menor mas um erro de validação $0,002$ maior do que o modelo KNN.

A valiando métricas comuns quando se lida com problemas de classificação, como a precisão (a fração de instâncias relevantes de uma classe dentre as recuperadas), cobertura (a fração de instâncias relevantes recuperadas sobre o total de instâncias relevantes) e *f1-score* (média geométrica entre as anteriores) não é notável também nenhuma diferença significativa entre os modelos. O modelo SVM possui uma precisão ligeiramente superior ao modelo KNN em relação à classe 1 e uma cobertura ligeiramente superior em relação à classe -1 e sendo assim, é possível dizer que seu desempenho é sutilmente melhor do que o concorrente, dependendo o interesse existente na correta classificação de cada classe. No entanto, como neste trabalho não se considerou nenhuma peculiaridade em relação à importância relativa entre as classes não é possível afirmar categoricamente que um modelo se sairá melhor que o outro em novas amostras.

SVM			
Classe	***precision***	***recall***	***f1-score***
-1	0.89	0.96	0.92
1	0.94	0.85	0.89
Média	0.91	0.91	0.91

TABLE 9. MÉTRICAS DIVERSAS DO MODELO SVM SELECIONADO EM RELAÇÃO AO CONJUNTO DE TESTE.

KNN			
Classe	***precision***	***recall***	***f1-score***
-1	0.89	0.94	0.92
1	0.92	0.85	0.89
Média	0.90	0.90	0.90

TABLE 10. MÉTRICAS DIVERSAS DO MODELO KNN SELECIONADO EM RELAÇÃO AO CONJUNTO DE TESTE.

As figuras 5 e 6 mostram as matrizes de confusão dos modelos SVM e KNN selecionados em relação ao conjunto de teste.

Quando comparadas suas matrizes de confusão, os dois modelos também não parecem significativamente diferentes,

apresentando a mesma taxa de acerto e erro quando a classe verdadeira é 1 e, quando esta é 0, apresentando apenas uma pequena diferença na qual a correta predição da classe é ligeiramente maior no modelo SVM.

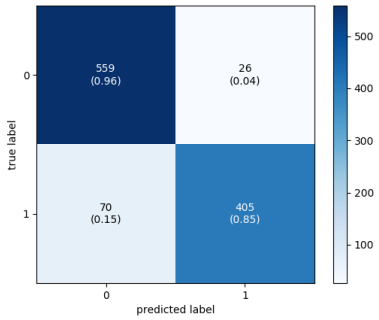

Figure 5. Tabela de confusão da melhor configuração do SVM quando avaliando o conjunto de teste.

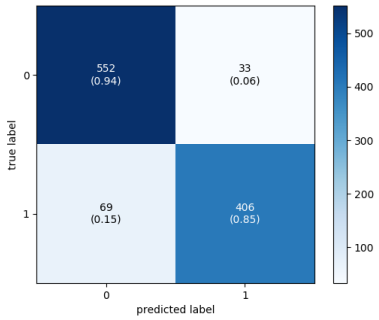

Figure 6. Tabela de confusão da melhor configuração do KNN quando avaliando o conjunto de teste.

5. Conclusão

Neste trabalho foi realizada uma comparação experimental entre diferente configurações de parâmetros para o modelo SVM e KNN em uma base de dados sintética de classificação binária, através da técnica de validação cruzada.

Os resultados mostraram que, para a base utilizada, os modelos SVM com kernel RBF e o modelo KNN funcionam melhor em relação a outras configurações do SVM sob a métrica da acurácia. Nota-se que ambos estes modelos se baseiam teoricamente na influência proporcional às distâncias de um ponto em relação a outros, o que parece ser uma característica marcante para atingir resultados interessantes nos experimentos aqui realizados.

References

[1] Y. S. Abu-Mostafa, *Learining From Data* ., 1999

[2] KEEL-dataset, http://sci2s.ugr.es/keel/dataset.php?cod=182

[3] https://work.caltech.edu/lectures.html]

[4] Cortes, C., Vapnik, V. (1995). Support vector machine. Machine learning, 20(3), 273-297.

[5] Cortes, C. Vapnik, V. (1995). Support-vector network. Machine Learning, 20, 1–25