INF0615 – Aprendizado de Máquina Supervisionado

Trabalho 2 - Regressão Logistica

Rodolfo Dalla Costa Nicole Nogueira

9/11/2021

Introdução

O sistema imunológico humano é o sistema responsável por proteger o corpo de antígenos como vírus e bactérias. A produção de glóbulos brancos, nome dado às células que compoem o sistema imunologico, é originada por algumas cadeias proteicas presentes no antigeno. Desse modo, propoe-se utilizar um modelo de regressao logistica com o objetivo, a partir de determinadas caracteristicas de uma cadeia proteica, a mesma pode gerar uma resposta do sistema imunilogico.

Banco de dados

Tabela 1: Estatísticas sumárias do banco de dados

end_position	chou_fasman	emini
Min. : 6	Min. :0.534	Min. : 0.00
1st Qu.: 96	1st Qu.:0.912	1st Qu.: 0.25
Median : 202	Median :0.991	Median : 0.56
Mean : 308	Mean :0.996	Mean : 1.06
3rd Qu.: 391	3rd Qu.:1.074	3rd Qu.: 1.21
Max. :3033	Max. :1.546	Max. :25.14

kolaskar_tongaonkar	parker	isoelectric_point
Min. :0.849	Min. :-9.03	Min. : 3.69
1st Qu.:0.986	1st Qu.: 0.60	1st Qu.: 5.62
Median :1.020	Median : 1.79	Median : 6.52
Mean :1.021	Mean : 1.77	Mean : 7.07
3rd Qu.:1.055	3rd Qu.: 2.99	3rd Qu.: 8.68
Max. :1.255	Max. : 9.12	Max. :12.23

aromaticity	hydrophobicity	stability	target
Min. :0.0000	Min. :-1.971	Min. : 5.4	Min. :0.000
1st Qu.:0.0625	1st Qu.:-0.606	1st Qu.: 31.7	1st Qu.:0.000
Median :0.0749	Median :-0.331	Median : 42.3	Median :0.000
Mean :0.0757	Mean :-0.409	Mean : 43.8	Mean :0.271
3rd Qu.:0.0913	3rd Qu.:-0.190	3rd Qu.: 49.1	3rd Qu.:1.000
Max. :0.1823	Max. : 1.267	Max. :137.0	Max. :1.000

Na 1 pode ser observado os dados que foram utilizados para o desenvolvimento do trabalho. Nota-se que todos praticamente sao dados numericos, e a coluna target é a coluna resultado. A base foi dividida em 3

partes da base e uma outra base externa para testes, sendo portanto, 1 parte de treino, 1 de validaçao, 1 de teste e 1 de teste sobre o virus SARS. Cada uma contem respectivamente 9204, 2303, 2878 e 520 linhas e um total de 11 colunas (como observado acima).

Análise Descritiva

```
## [1] "Dados Faltantes no treino:  FALSE"
```

```
## [1] "Dados Faltantes na validação:  FALSE"
```

```
## [1] "Dados Faltantes no teste:  FALSE"
```

```
## [1] "Dados Faltantes no SARS:  FALSE"
```

Como pode ser observado, a base nao possui nenhum dado faltante para nenhuma das partes.

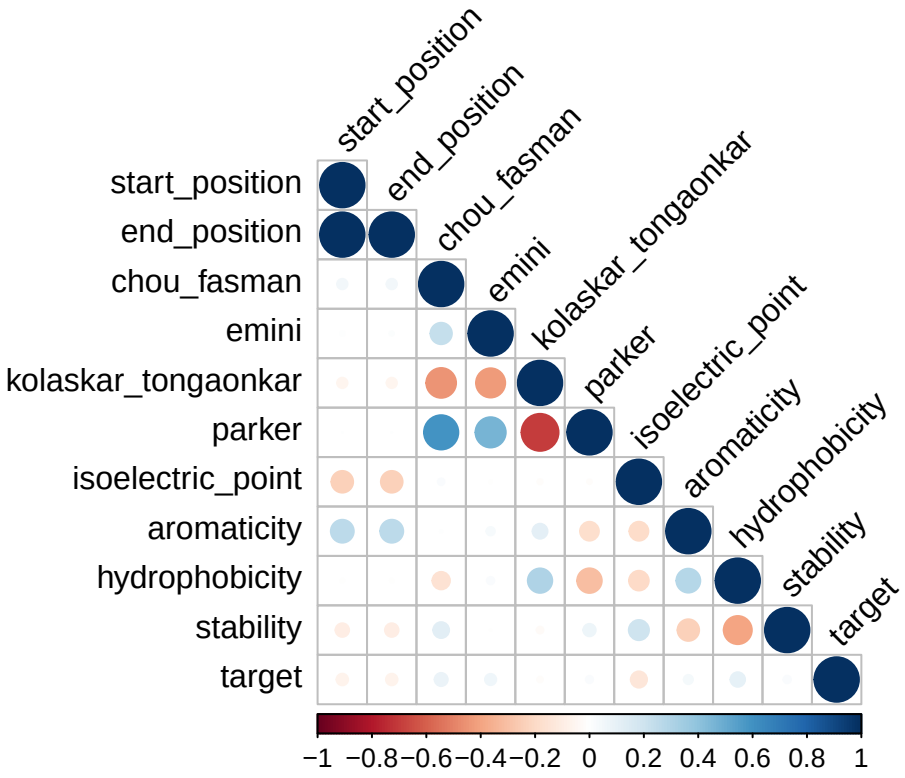

Figura 1: Correlações 2 a 2 das variáveis.

Dado o mapa de correlacoes da Figura 1, observa-se que a variavel `target` nao possui correlacoes com um valor evidentemente alto, sendo notavel uma correlacao inversa com a variavel `target` com a variavel `isoeletric_point` e uma correlacao inversa mais baixa com `start_position` e `end_position`. Em contrapartida nota-se um correlacao positiva, porem mais fraca, com as variaveis `chou_fasm`, `hydrophobicity`, `emini` e `aromaticity`.

Metodologia

Apos a etapda de inspecao dos dados, foi realizada a normalizacao utilizando o metodo Z-Norm. Em seguida, a partir dos dados nao balanceados foi gerada uma Baseline considerando todas as variaveis num polinomio de grau 1. Apos isso, foram aplicadas 3 tecnicas de balanceamento: SMOTE, undersample e a ponderada por pesos; dentre elas o balanceamento ponderado gerou melhores resultados. A partir disso, uma serie de hipoteses foram testadas para gerar o modelo: polinomios de 1 a 12, combinacoes considerando as variaveis de maior correlacao e a propria baseline.

Resultados e Conclusão