Trabalho 2

INF-0615 – Aprendizado de Máquina Supervisionado I

Grupo:
Evandro Santos Rocha
Laíssa Pacheco de Oliveira
Rafael Dantas de Moura

O objetivo deste trabalho é treinar regressões logísticas utilizando uma base de dados com diversas informações sobre cadeias de proteínas para predizer se ela estimula ou não a produção de anticorpos no organismo humano.

1. Inspecionem os dados. Quantos exemplos vocês tem?

Com o comando `dim` aplicado aos conjuntos de treino e validação, obtivemos:

Conjunto	Número de exemplos (linhas)	Número de colunas
Treino	9204	11
Validação	2303	11

Já com o comando `summary`, podemos observar cada atributo. De imediato, podemos observar que nenhum dos conjuntos possui NA para nenhuma coluna. Observamos também que os dados da coluna “target” estão como contínuos. Como foi informado que este campo é binário, ele será transformado para fator logo no início do tratamento dos dados.

Adicionalmente, verificamos que os conjuntos de treinamento, de validação e (mais tarde) de teste são disjuntos, ou seja, não há registro (exemplo) repetido entre os conjuntos. Isso foi feito com o comando `merge`.

Há exemplos com features sem anotações? Como vocês lidariam com isso?

Fizemos a verificação de features sem anotações e não foi encontrada nenhuma. O próprio `summary` já tinha indicado isso, mas o comando `any(is.na(<dataset>))` retornou FALSE para os conjuntos, indicando que não existe NA em nenhum atributo.

Caso houvesse, o tratamento dependerá da quantidade de atributos, por linha, com valor igual a NA. Caso toda a linha não possuísse nenhum valor, a linha seria excluída do conjunto. Por outro lado, se apenas alguns valores da linha estivessem omitidos, faríamos um tratamento para decidir substituir o NA pela média, mediana, moda, etc. A escolha dependerá do contexto.

2. Inspecionem a frequência de cada classe. A base de dados está balanceada? Se não, como vocês lidarão com o desbalanceamento?

O atributo “target” possui duas classes, onde a classe 0 é bem mais frequente que a classe 1 (cerca de 2,7 vezes). Isso mostra um desbalanceamento entre elas (quadro abaixo). Faremos o balanceamento pela ponderação da função de erro, aplicando um peso relativo ao complemento da frequência.

Classe	Frequência	Frequência relativa	Peso aplicado
0	6709	0.7289222	0.2710778

1	2495	0.2710778	0.7289222

3. Apliquem alguma técnica de normalização (...)

Aplicamos a normalização **Z-norma** para todas as colunas, exceto a target.

4. Como baseline, treinem uma regressão logística com todas as features para predizer se haverá ou não a produção de anticorpos. Reportem a matriz de confusão relativa, o TPR, o TNR e a acurácia balanceada nas bases de treinamento, validação e teste (apenas arquivo proteins_teste_set.csv).

Acurácia balanceada, TPR e TNR (baseline):

Conjunto	TPR Taxa de Verdadeiros Positivos	TNR Taxa de Verdadeiros Negativos	Acurácia balanceada
Treinamento	0,60	0,64	0.62
Validação	0,59	0,64	0,615
Teste	0,60	0,64	0,62

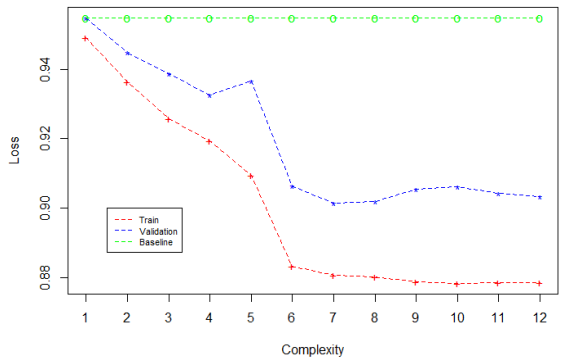
Matrizes de confusão balanceada (baseline):

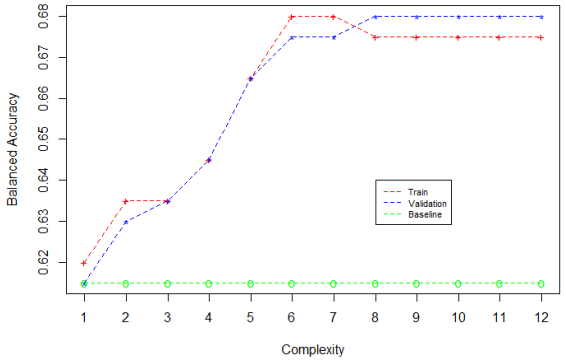
		Treinamento		Validação		Teste	
		0	1	0	1	0	1
Label	0	0,60	0,40	0,59	0,41	0,60	0,40
	1	0,36	0,64	0,36	0,64	0,36	0,64

5. Implementem soluções alternativas baseadas em regressão logística através da combinação das features ou modelos polinomiais para melhorar o resultado do baseline.

Primeiramente, criamos soluções baseadas em regressão logística através de modelos polinomiais e, depois, através da combinação de features.

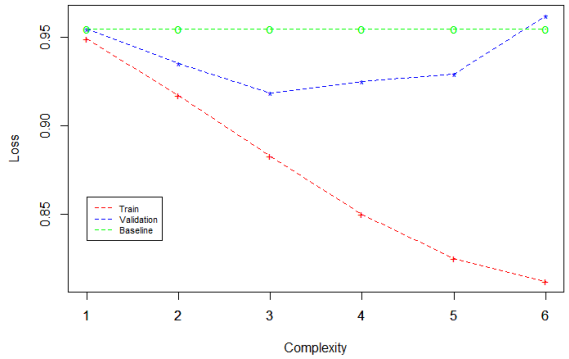
Por modelos polinomiais, criamos 12, onde o de grau 1 é a própria baseline. Para determinar a melhor solução foram analisados os gráficos de "viés e variância" e de acurácia balanceada (abaixo). Pelos gráficos, a **região ótima** está entre as regressões de grau 7 e 8: a perda é mínima em 7, porém, mesmo perdendo mais um pouco, ganha em acurácia na 8a. Podemos também dizer que o modelo está em **overfitting** entre as regressões 9 e 12 e, em **underfitting** entre 1 e 6. Então, entre as regressões logísticas polinomiais, assumimos a 8a como a melhor, onde a acurácia balanceada foi de **0,68** e perda de 0,9019757.

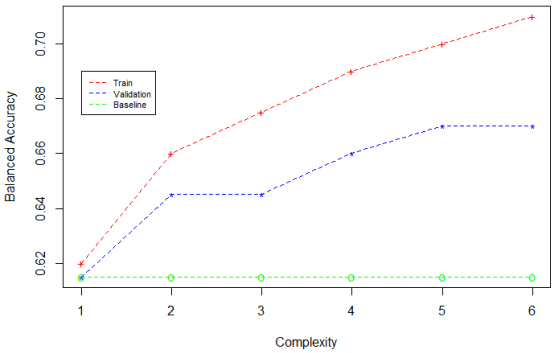




Também criamos regressões logísticas por combinação de features, desde 1 a 1 até 6 a 6, onde também analisamos os gráficos de "viés e variância" e acurácia balanceada (abaixo).

Pelos gráficos, o **ponto ótimo** está entre a 3ª e 5ª regressões: a menor perda está na 3ª e a maior acurácia balanceada está na 5ª. Certamente na combinação 6 a 6, o modelo entra em **overfitting** e, entre a 1ª e a 2ª regressão, o modelo está está em **underfitting**. Então, entre as regressões da região ótima, podemos assumir que a 4ª foi a melhor, onde a acurácia foi de **0,66** e perda de 0,9249632.





No final, considerando a maior acurácia balanceada e a menor perda, a **melhor** de todas as regressões foi a **polinomial de grau 8**.

Comparem suas soluções reportando a matriz de confusão relativa e a acurácia balanceada no conjunto de validação.

Para as regressões logísticas polinomiais (de 1 a 8, pois de 9 a 12 são iguais a 8ª):

	Grau 1		Grau 2		Grau 3		Grau 4		Grau 5		Grau 6		Grau 7		**Grau 8**	
	0	1	0	1	0	1	0	1	0	1	0	1	0	1	**0**	**1**
Label 0	0.59	0.41	0.61	0.39	0.59	0.41	0.61	0.39	0.63	0.37	0.64	0.36	0.64	0.36	**0.64**	**0.36**
Label 1	0.36	0.64	0.35	0.65	0.32	0.68	0.32	0.68	0.30	0.70	0.29	0.71	0.29	0.71	**0.28**	**0.72**
Acc Bal	0.615		0.63		0.635		0.645		0.665		0.675		0.675		**0.68**	

Para as regressões logísticas com combinação de features:

	Grau 1		Grau 2		Grau 3		**Grau 4**		Grau 5		Grau 6	
	0	1	0	1	0	1	**0**	**1**	0	1	0	1
Label 0	0.59	0.41	0.62	0.38	0.65	0.35	**0.66**	**0.34**	0.67	0.33	0.67	0.33
Label 1	0.36	0.64	0.33	0.67	0.36	0.64	**0.34**	**0.66**	0.33	0.67	0.33	0.67
Acc Bal	0.615		0.645		0.645		**0.66**		0.67		0.67	

Tomem apenas a melhor solução, baseada na acurácia balanceada no conjunto de validação, e reportem a matriz de confusão relativa, TPR, TNR e acurácia balanceada no conjunto de teste (apenas arquivo proteins_teste_set.csv).

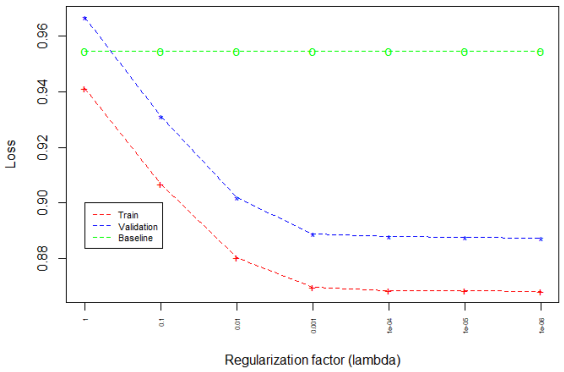
Informações da melhor solução, regressão logística de grau 8:

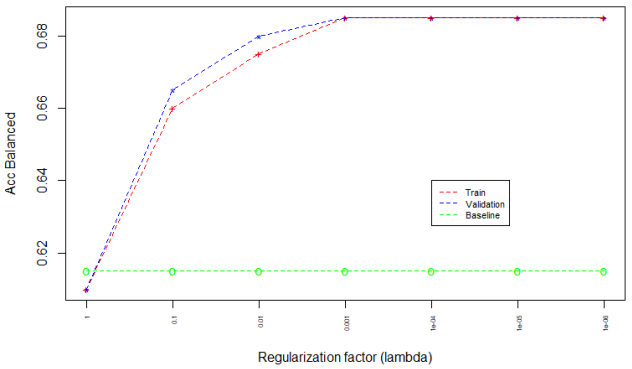
Matriz de confusão	**Prediction**	
	0	**1**
Label 0	0.65	0.35
Label 1	0.30	0.70

TNR	TPR	Acurácia balanceada
0,65	0,70	0,675

6. Tomem um dos modelos do item anterior e varie o fator de regularização (). Criem a curva viés/variância colocando os diferentes valores de no eixo das das abscissas. Identifiquem as regiões de underfitting, ponto ótimo e overfitting, e então tomem o modelo com o melhor fator de regularização e reporte a matriz de confusão relativa, o TPR, o TNR e a acurácia balanceada no conjunto de teste (apenas arquivo proteins_teste_set.csv).

Tomamos o modelo da **melhor solução, regressão logística de grau 8**, e variamos o lambda entre 1.0 e 10e-6 para obter os gráficos de "viés e variância" e acurácia balanceada (abaixo). Pelos gráficos, o **ponto ótimo** está entre os lambdas 10e-3 e 10e-4. Para os lambdas 10e-5 e 10e-6, podemos dizer que o modelo está em **overfitting**, e, para os lambdas entre 1.0 e 10e-2, o modelo está em **underfitting**. Então, entre as regressões da região ótima, podemos assumir que a de **lambda 10e-3** é a melhor, pois já obteve acurácia máxima de 0,685 e a perda não foi tão maior que a do lambda seguinte, 10e-4.





Informações da regressão logística de grau 8 com lambda igual a 10e-3:

Matriz de confusão	**Prediction**	
	0	**1**
Label 0	0.66	0.34
Label 1	0.31	0.69

TNR	TPR	Acurácia balanceada
0,66	0,69	0,675

6. Escrevam um relatório de no máximo 5 páginas:
(a) Descrevam o que foi feito, bem como as diferenças entre o seu melhor modelo e o seu baseline;

Alguns detalhes estão melhor explicados nas questões acima.

Inicialmente, questão 4, foi realizada uma regressão logística, chamada de baseline, com os dados de treino. Com essa **baseline**, foi feita uma predição, onde a acurácia balanceada foi de **0,615** no conjunto de validação.

Em seguida, questão 5, foram realizadas regressões logísticas através de modelos polinomiais e de combinação de features. Para as de modelos polinomiais, variamos do graus 1 ao 12, onde a de grau 8 foi a melhor com acurácia balanceada de **0,680**. Para as de combinação de features, combinamos de 1 a 1 até 6 a 6, onde a de 4 a 4 se mostrou a melhor com acurácia balanceada de **0,660**. Como conclusão, a melhor regressão logística foi a **polinomial de grau 8**.

Na questão 6, pegamos a melhor regressão e variamos o lambda, obtendo **10e-3 como o melhor lambda** e obtendo uma acurácia balanceada de **0,675**.

(a) Após desenvolverem todos os modelos, tomem o melhor modelo de todos (melhor performance no conjunto de validação). Reportem a matriz de confusão relativa e acurácia balanceada nos conjuntos de teste proteins_teste_set.csv e SARS_test_csv. Há uma diferença significativa entre eles ? Se sim, qual explicação você daria para essa diferença ?

Informações para o conjunto de teste:

Matriz de confusão	**Prediction**			TNR	TPR	Acurácia balanceada
	0	**1**		0,66	0,69	0,675
Label 0	0.66	0.34				
Label 1	0.31	0.69				

Informações para o conjunto de SARS:

Matriz de confusão	**Prediction**			TNR	TPR	Acurácia balanceada
	0	**1**		0,25	0,82	0,535
Label 0	0,25	0,75				
Label 1	0,18	0,82				

É possível notar que a acurácia balanceada foi bem menor no conjunto de SARS. Isso ocorreu, provavelmente, por se tratar de uma predição em cima de um modelo gerado por outra base de dados. Por outro lado, ocorreu algo interessante: a TPR (taxa de verdadeiros positivos) no conjunto SARS está bem maior, **0,82**, que a realizada no conjunto de teste, **0,69**, ou seja, acertou muito os casos positivos. Talvez algum atributo esteja contribuindo mais para isso.

(c) Uma Seção de conclusão do relatório explicando a diferença entre os modelos e o porquê que estas diferenças levaram a resultados piores ou melhores.

Como conclusão, ao realizar regressões logísticas através da combinação de features e pelo método polinomial, conseguimos melhores resultados, ou seja, menores erros e maiores acurácias balanceadas. Além disso, também vimos que podemos melhorar o modelo através de variar o lambda.

Também é essencial plotar os gráficos de “viés e variância” e de acurácia balanceada para, ao observar o comportamento das perdas e acurácia dos modelos, identificar o ponto ótimo e as regiões de underfitting e overfitting.