Instituto Superior Técnico

MEEC

Aprendizagem Automática

Lab 5

Evaluation and Generalization

Grupo 9

Manuel Diniz, 84125
Alexandre Rodrigues, 90002

Turno: 4ªf 11h00

Contents

Chapter 1

Classificação

1.1 Introdução

É pedido que se classifique um conjunto de dados relativo ao cancro. Foram escolhidos os seguintes classificadores:

- SVM Polinomial
- SVM Linear
- Naive Bayes
- Decision Tree

O grau escolhido para o SVM Polinomial é 3, o Naive Bayes utiliza o método Gaussiano e por fim, a Decision Tree foi apenas criada e treinada.

Quanto aos parâmetros utilizados para a avaliação da *performance*, estes foram os seguintes:

- Accuracy
- F-Measure
- Confusion Matrix

A *accuracy* mede a precisão do modelo, enquanto que o *F-Measure* devolve uma melhor medida dos casos classificados incorretamente dos mesmos. A *Confusion Matrix* faz um resumo dos resultados previstos.

1.2 *SVM*

Para ambos os casos, linear e polinomial, criou-se o classificador, treinou-se o modelo e calcularam-se as previsões do conjunto de dados para teste. Os resultados estão apresentados na tabela em baixo.

	Accuracy	F-Measure
SVM Linear	83.33%	81.81%
SVM Polinomial	62.5%	0%

As *Confusion Matrices* obtidas respetivamente para o caso linear e polinomial foram as seguintes:

$$\begin{bmatrix} 9 & 0 \\ 4 & 11 \end{bmatrix}; \begin{bmatrix} 0 & 9 \\ 0 & 15 \end{bmatrix}$$

1.3 *Naive Bayes*

Tal como efetuado anteriormente, repetiu-se o mesmo processo para a aplicação do classificador *Gaussiano de Naive Bayes*. Os resultados estão apresentados na tabela seguinte.

	Accuracy	F-Measure
Naive Bayes	62,5%	66.66%

E a sua *Confusion Matrix* foi a seguinte:

$$\begin{bmatrix} 9 & 0 \\ 9 & 6 \end{bmatrix}$$

1.4 *Decision Tree*

Por fim, criou-se uma árvore de decisão, calcularam-se os mesmos parâmetros avaliados anteriormente e reproduziu-se uma imagem da árvore.

Os parâmetros obtidos estão representados em baixo:

	Accuracy	F-Measure
Decision Tree	58.33%	50%

$$\begin{bmatrix} 5 & 4 \\ 6 & 9 \end{bmatrix}$$

De seguida apresenta-se a árvore de decisão gerada para o conjunto de dados testado sobre o cancro.

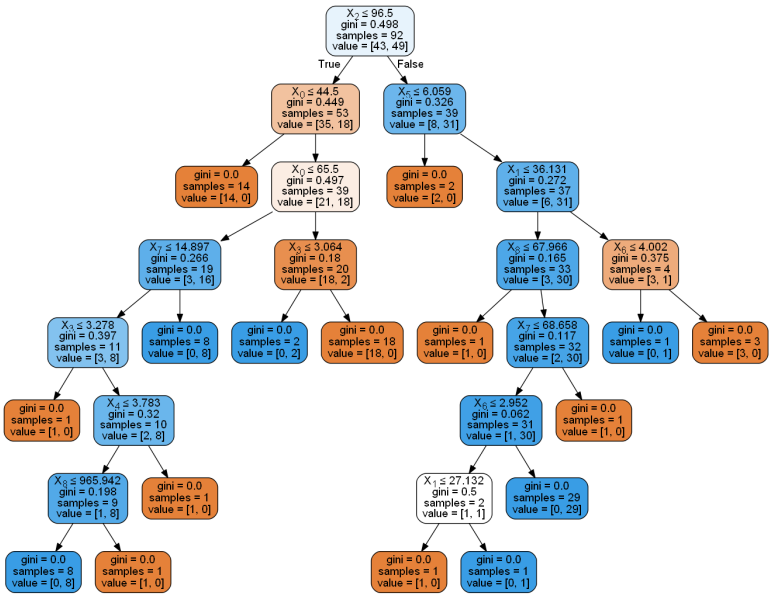

Figure 1.1: *Decision Tree*

1.5 Comparação de resultados

Procedeu-se de seguida à comparação e avaliação dos resultados obtidos ao longo dos treinos e testes do *dataset* fornecido.

	Accuracy	F-Measure
SVM Linear	83.33%	81.81%
SVM Polinomial	62.5%	0%
Naive Bayes	62.5%	66.66%
Decision Tree	58.33%	50%

Por análise direta da tabela, verifica-se que o classificador linear SVM é o mais preciso quer em termos de *accuracy*, quer em termos de *F-Measure*, registando uma *performance* de 83.33% e de 81.81% de *accuracy*, respetivamente.

De todas as *Confusion Matrices* é possível dizer que o classificador SVM é melhor a identificar quer valores positivos, quer valores negativos do que os outros classificadores.

Concluindo assim que o SVM deve ser o classificador a usar, uma vez que apresenta um conjunto total de melhores resultados. No entanto, o SVM pode requerer uma procura exaustiva dos hiperârmetros, pelo que uma função de optimização para realizar esta procura deverá ser implementada para reduzir o custo do processo.

Chapter 2

Regressão

2.1 Introdução

É fornecido um conjunto de dados relativo a imobiliário, com 13 *features* não específicadas e um resultado, o preço.

O conjunto de dados fornecido tem uma dimensão relativamente reduzida, com cerca de 400 amostras. Para compensar, ao treinar o modelo podemo-nos sentir tentados a treinar por um número de épocas elevado de modo a melhorar o desempenho. Com este método há o risco de do modelo se tornar *overfit*.

De modo a evitar isto, faz-se uso de um conjunto de validação. Após cada época, o modelo é avaliado com este conjunto, e a sua *performance* neste é usada para decidir se o treino deve ser parado ou não. O treino pára se a sua *loss* para este conjunto não decrescer após 50 épocas, e os melhores coeficientes do modelo são restaurados.

Já na regressão é importante escolher uma ordem que aproxime bem os dados, mas suficientemente reduzida de modo a que não ocorra *overfitting*.

2.2 Treino da rede neuronal

De modo a realizar a previsão de preços, estabelece-se, após alguma tentativa e erro, um modelo com quatro camadas, todas elas com ativação *relu*, e com 32, 32, 16 e 1 neurónios, por essa ordem. Normalmente a última camada tem ativação línear, mas neste caso é indiferente, visto que não se esperam resultados negativos para o preço.

Como referido anteriormente, o treino é parado antecipadamente após 50 épocas sem melhorias nos resultados. Como métricas são usados o erro absoluto percentual e erro absoluto médios, sendo o segundo também utilizado como *loss*.

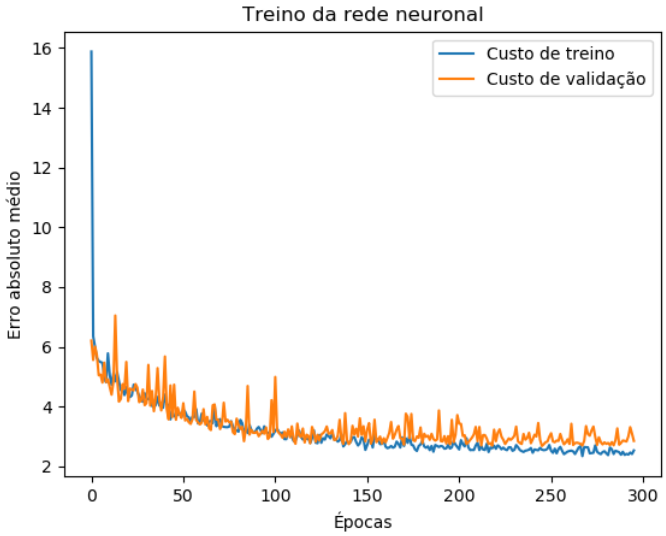

Figure 2.1: Evolução da *performance* da rede neuronal

Como se observar na figura 2.1, o treino pára ao fim de cerca de 300 épocas, quando o custo de validação estabiliza.

2.3 Regressão línear

O outro método de previsão utilizado é a regressão línear, em que o preço previsto é o resultado da soma das multiplicações de cada *feature* pelo coeficiente respetivo, e um *offset*. Sendo que existem 13 *features*, obtém-se 14 coeficientes.

O "treino" neste caso é instantâneo, os coeficientes obtidos através da equação normal. Os resultados estão abaixo.

2.4 Comparação de resultados

Os resultados das regressões estão na tabela seguinte.

	Erro absoluto percentual médio	Erro absoluto médio
Rede neuronal	13.7894%	2.6816
Regressão línear	17.3136%	3.3005

Devido ao conjunto de dados reduzido, e ao problema em geral, que não ilustra uma ciência exata, obtém-se um erro não insignificante na regressão. No entanto, é bem aproximado o suficiente para generalizar e obter aproximações. Não existe uma diferença muito significativa entre a *performance* dos dois métodos, sendo que a regressão línear já é bem aproximada. Isto sugere uma regressão polínomial provavelmente estaria a par da rede neuronal, com um custo computacional inferior, sendo que seria talvez uma boa escolha.

Mais uma vez, na hipótese de se usar uma regressão polinomial há que ter o cuidado de selecionar uma ordem não excessivamente elevada. Tendo em conta o tipo de dados, uma ordem entre 3 e 5

seria certamente suficiente.