INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA

ADEETC
Aprendizagem Auotmática
1º Semestre 2017/2018

Docente: Gonçalo Marques

Relatório 1º Trabalho Prático

Rui Santos nº 39286

26 de Novembro 2017

Conteúdo

Lista de Figuras

1 Introdução

Este trabalho consiste na classificação/identificação de faces. Para este efeito, será usada a base de dados "labeled faces in the wild". O trabalho está dividido em várias etapas, as quais se descrevem resumidamente de seguida.

1. Projetar e avaliar o desempenho do classificador dois K vizinhos mais próximos (kNN);
2. Estimar qual o número óptimo de vizinhos;
3. Repetir a avaliação do desempenho usando faces processadas com PCA;
4. Estimar qual o núm,ero óptimo de componentes principais;
5. Verificar se normalizar a variância dos dados transformados é benéfico;
6. Testar e avaliar outro classificador além do kNN, nos dados transdormados e comparar os resultados com os do kNN;
7. Projetar e avaliar o desempenho de um classificador à sua escolha para o problema de identifcação/verificação de uma das seguintes personalidades: George W. Bush ou Colin Powell ou Tony Blair. DE notar que este é um problema de classificação binária e deverá ter isso em conta na avaliação.

De forma a respeitar a ordem dos pontos do trabalho e a estrutura do relatório, os processos de desenvolvimento serão descritos e explicados de acordo com essa ordem no capitulo aseguir.

2 Desenvolvimento

2.1 Avaliação do desempenho do kNN

Como é descrito no enunciado do trabalho, existem várias personalidades na biblioteca fornecida, no entanto é de notar que existem demasiadas amostras para algumas personalidades, este factor pode induzir o classificador em erro na classificação de algumas amostras não dando resultados coerentes.
De forma a contornar este problema, limita-se o número de imagens/amostras para no máximo 50, sendo o resultado final um leque de 1410 imagens totais e na mesma as 6510 componentes e 34 classes, classes essas que correspodem individualmente a cada personalidade, ou seja, 34 nomes de personalidades.

O classificador que iremos utilizar nesta primeira parte é o *KNeighborsClassifier*
Este classificador é considerado dos mais básicos no tópico de machine learning, pois somente tem em conta valores locais em torno do valor atualmente a analisar.

Para instanciarmos um classificador deste tipo, há parâmetros que temos de atribuir valores/constantes, parâmetros esses que são os seguintes:
- Número de vizinhos: É preferivel que seja uma constante ímpar de forma a não haver empates quando este está a classificar a que classe pertence o ponto em análise, ou seja, se somente temos 4 vizinhos e dois deles são classificados como uma determinada classe e outros dois com outra, o resultado pode não ser verdadeiro pois não havia mais vizinhos para desempatar.
- Peso: Este parâmetro tem grande influência na classificação definindo qual o método usado para a mesma, se é passado um valor *'uniform'*, todos os vizinhos têm o mesmo tipo de peso a atribuir uma classe, se é passado o valor *'distance*, os vizinhos com o inverso da distância maior, têm mais peso atribuir uma classe.

Na continuação deste processo, temos de introduzir um modelo para testar as amostras de cada personalidade, para isso iremos utilizar Validação Cruzada *Kfold*. A validação cruzada tem como objectivo a divisão em subconjuntos de acordo com, uma constantes que indica o numero de conjuntos e o valor que é dividido, neste caso é um valor da amostra.
Para esta parte iremos utilizar o *StratifiedKFold* que tem como parâmetros o numero de divisões, se queremos aplicar um *shuffle* de forma a não obtermos resultados falaciosos, pois se as classes estão organizadas por ordem crescente por exemplo, e a aplicação dos kfolds é feita sequencialmente iremos confun-

dir a validação, para contornar este problema, este parâmetro tem o valor *true* e indicamos tambem que queremos sempre a mesma divisão *random state = 0* com o mesmo gerador de números sempre. Usamos este modelo por se adequar bastante à base de dados que estamos a utilizar, visto que há classes com bastantes mais ocorrências que outras.

2.2 Número óptimo de vizinhos

Posto isto, estamos prontos para apresentação de resultados.

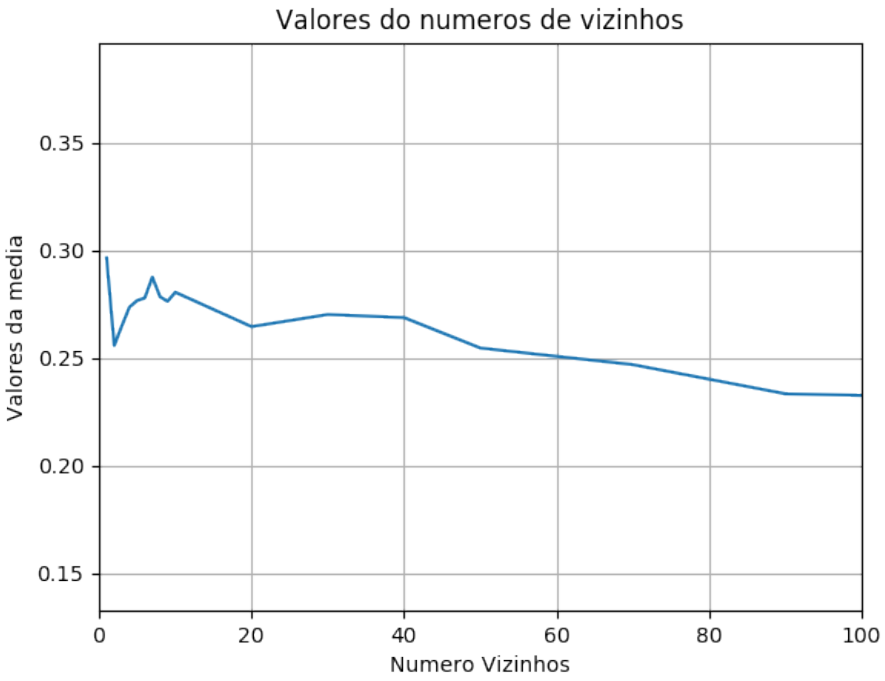

Figura 1: Descoberta do melhor número de vizinhos

Como podemos analisar na figura anterior, foi feito um teste para descobrir qual o melhor número de vizinhos, com o valores obtidos, podemos concluir que com o número de divisões igual a 3, o melhor número de vizinhos é 1, com uma média de acerto aproximadamente 30%, é de notar que não é de todo uma boa classificação.

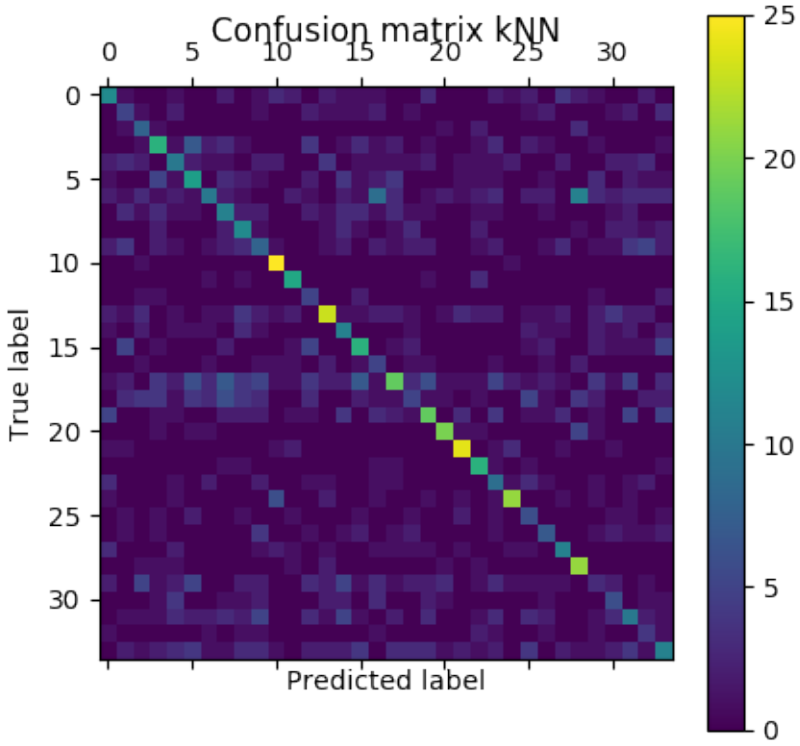

Figura 2: Matriz de confusão kNN - Nº Vizinhos 1

A matriz de confusão dá-nos uma perspectiva sobre os resultados para todas as imagens. Podemos observar que existem bastantes incongruências, e que apenas em algumas classes o classificar consegue realmente classificar corretamente.
Para uma melhor interpretação do gráfico, quanto mais colorida é a diagonal, maior é a percetagem de classificação correta.

2.3 Avaliação do desempenho com PCA

Nesta fase iremos utilizar o algoritmo de PCA, que visa extrair a componentes principais, conhecidas também como zonas de maior variância dos dados. O PCA projeta esses dados nessas mesmas zonas, no entanto é preferivel por nós especificar o numero de componentes que queremos guardar para extração. No entanto como não sabemos qual o melhor número para obter melhores resultados, iremos optar por uma abordagem similar à descoberta do melhor número de vizinhos.
Para este cáculo, estamos a usar o numero de vizinhos óptimo descoberta nas fases anteriores.

2.4 Número óptimo de componentes

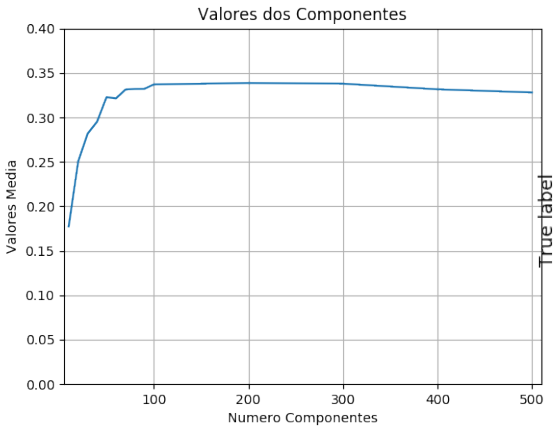

Figura 3: Descoberta do melhor número de componentes

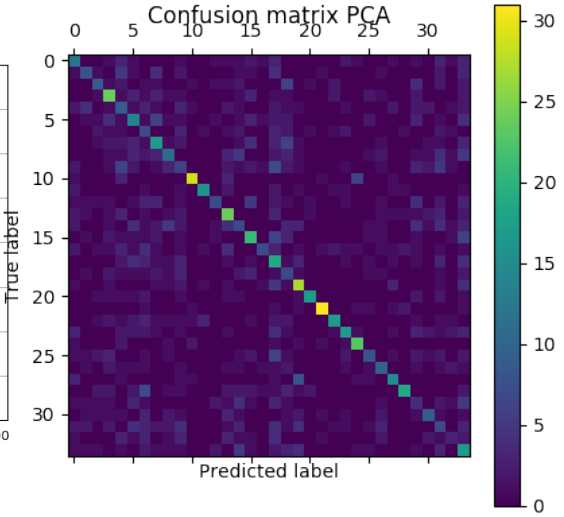

Figura 4: Matriz de confusão com PCA

Podemos concluir que o melhor número de componentes, está compreendido entre 200 e 300 de acordo com os valores que a curva apresenta.
Na figura à direita, podemos observar o efeito que o PCA tem na matriz de confusão quando usamos o mesmo classificado *KNeighborsClassifier*.
Anteriormente o classificador tinha uma precisão de aproximadamente 30%

e com a aplicação do PCA aumentou para 33.9%. É de facto vantajoso usar PCA, no entanto é de notar que a melhoria é minima.

2.5 Normalização dos dados

Agora iremos testar exatamente os mesmos passos que anteriormente tomamos mas com a normalização dos daods, isto é , aplicar uma variância nula, para isso apenas temos de indicar à função que calcula as zonas de variância através do parâmetro *whiten*.

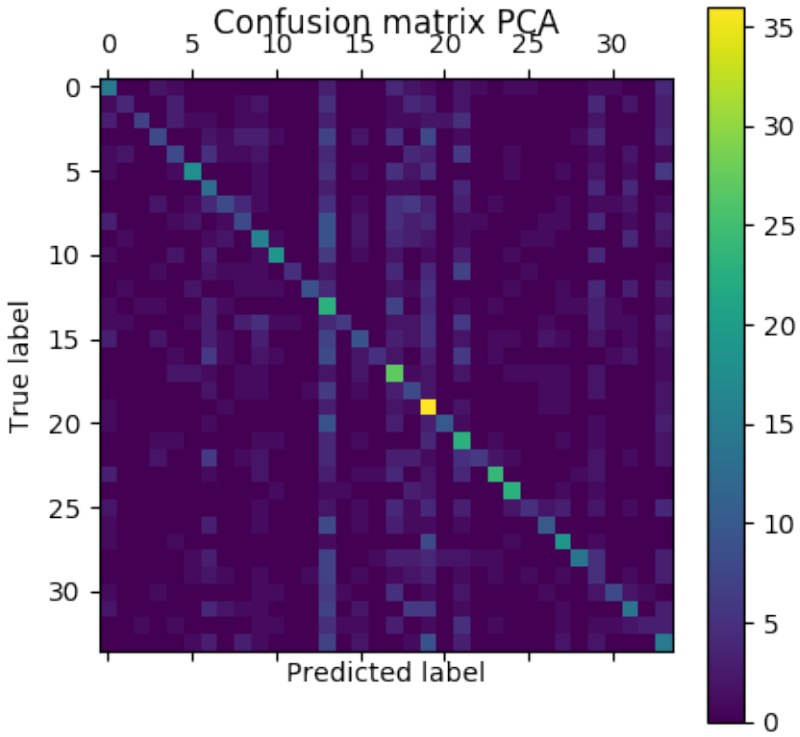

Figura 5: Matriz de confusão PCA normalizado

Ao observar a imagem podemos reparar que há uma nuvem entre as classes 15 e 20, que o classificador tem uma avalição bastante fora do que é suposto ser normal, e a sua média de acerto baixou novamente para os 30%, valor que so era tão baixo quando não usavamos o PCA.

2.6 Avaliação com outro classificador

Para esta fase, teriamos de escolher outro classificador que não o *KNeighborsClassifier* de forma a ter uma comparação e concluir se seria o melhor.
Escolhemos o *QuadraticDiscriminantAnalysis*, que aplica a cada classe uma densidade gaussiana e tem como metodo de analise uma *quadric surface* de cada objecto ou evento.

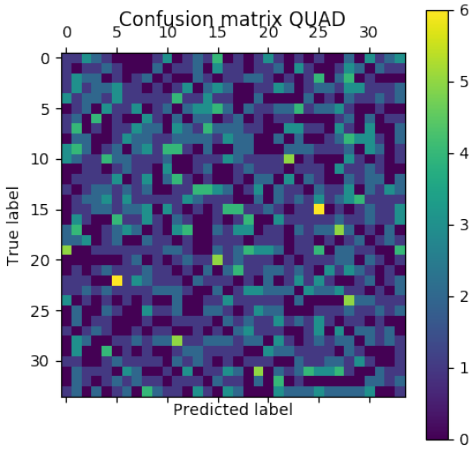

Figura 6: Matriz de confusão com QUAD- Normalizada

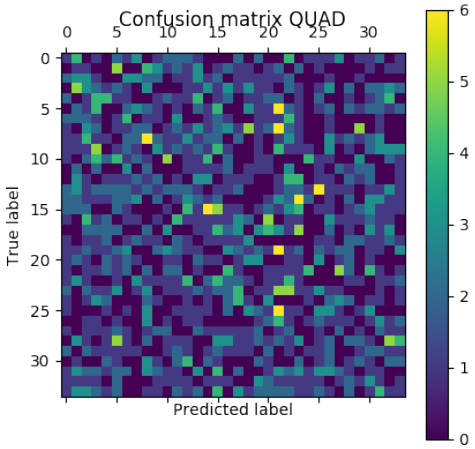

Figura 7: Matriz de confusão com QUAD

Podemos concluir que este classificador não é de todo indicado para este problema, pois tem uma taxa de sucesso na classificação muito baixa. A causa poderá ser, o facto de este classificador procurar por padrões similares a uma ellipse, e nem sempre as caras apresentam tal forma.
A analise das amostras quando normalizadas, apresentam uma taxa de sucesso de 0.031% e não normalizadas 0.046%, observamos novamente que a normalização dos dados não é benéfica e que em comparação ao *KNeighbours*, tem um desempenho eficiente muito menor.

2.7 Classificação Binária

Nesta fase iremos avaliar qual a eficácia do classificador *SVC* em avaliar as seguintes personalidades:

- Tony Blair
- Colin Powell
- George Bush

Para completar este ponto, decidimos utilizar um novo classificador, e tambem uma outra ferramenta de seleção que aplica uma pesquisa exaustiva de acordo com os parâmetros passados, neste caso, o classificador. Esta ferramenta é o *GridSearchCV*.

	precision	recall	f1-score	support
Alejandro Toledo	0.70	0.54	0.61	13
Alvaro Uribe	0.75	0.46	0.57	13
Andre Agassi	0.14	0.20	0.17	5
Ariel Sharon	0.44	0.50	0.47	8
Arnold Schwarzenegger	0.10	0.29	0.14	7
Colin Powell	0.55	0.50	0.52	12
David Beckham	0.50	0.67	0.57	9
Donald Rumsfeld	0.40	0.35	0.38	17
George W Bush	0.46	0.60	0.52	10
Gerhard Schroeder	0.39	0.58	0.47	12
Gloria Macapagal Arroyo	0.92	0.86	0.89	14
Guillermo Coria	1.00	0.50	0.67	10
Hans Blix	0.50	0.33	0.40	12
Hugo Chavez	0.64	0.56	0.60	16
Jacques Chirac	0.40	0.29	0.33	7
Jean Chretien	0.38	0.50	0.43	12
Jennifer Capriati	0.64	0.64	0.64	11
John Ashcroft	0.62	0.67	0.64	12
John Negroponte	0.50	0.50	0.50	8
Junichiro Koizumi	0.82	0.82	0.82	11
Kofi Annan	0.86	0.60	0.71	10
Laura Bush	1.00	0.83	0.91	12
Lleyton Hewitt	0.33	0.40	0.36	5
Luiz Inacio Lula da Silva	0.78	0.50	0.61	14
Megawati Sukarnoputri	0.86	0.75	0.80	8
Nestor Kirchner	1.00	0.38	0.55	8
Recep Tayyip Erdogan	0.40	0.29	0.33	7
Roh Moo-hyun	0.86	1.00	0.92	6
Serena Williams	0.64	0.82	0.72	11
Silvio Berlusconi	0.67	0.36	0.47	11
Tom Ridge	0.71	0.56	0.63	9
Tony Blair	0.32	0.64	0.42	11
Vicente Fox	0.89	0.57	0.70	14
Vladimir Putin	0.31	0.50	0.38	8
avg / total	0.62	0.56	0.57	353

Figura 8: Métrica Binária

Os parâmetros *precision* e*recall* significam respectivamente, classificações positivas corretas com influência dos falsos negativos e positivos classificados corretamentes com influência dos falsos positivos.
A precision e uma mediada de avaliaçao usada quando o objetivo e reduzir o número de falsos positivos. Por outro lado, o recall e usado quando o objetivo e reduzir os falsos negativos.
Sendo que o parâmetro *f-score* é a interceção dos dois parâmetros anteriores. Como pedido no ponto, podemos verificar que para as personalidades em questão o classificador tem um desempenho mediano com 52% para o Collin Powell, 42% para o Tony Blair e 52% para George W. Bush.

3 Conclusão

Em suma, conseguimos verificar que ambos os classificadores usados apresentam taxas de acerto bastante más, tal como se pode verificar nos resultados obtidos, no entanto um tem bastante melhor performance que o outro.

Podemos também concluir que o facto de não possuirmos um computador com elevada capacidade computacional, nos leva a utilizar também classificadores menos eficientes, visto os outros serem muito mais pesados para as nossas máquinas, apesar de o uso deste classificadores, quando a testar quais seriam os melhores vizinhos ou componentes, demorou uma quantidade de tempo consideravel a obter os resultados.

4 Bibliografia

precison recall

PCA

GridSearchCV

Eighen faces

QuadraticDiscriminantAnalysis

slides da cadeira