UNIVERSIDADE FEDERAL DO CEARÁ
CAMPUS SOBRAL
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E DE COMPUTAÇÃO

Reconhecimento de Padrões

Borges, C.D.B

MÉTODOS DE CLASSIFICAÇÃO: 1-NN, CENTROIDE MAIS PRÓXIMO, LDA E QDA

SOBRAL

2019

INTRODUÇÃO

No contexto de reconhecimento de padrões, classificação é o processo de decisão cujo objetivo é atribuir uma classe ou categoria a um determinado objeto. A classificação de padrões possui ampla importância no desenvolvimento da espécie humana, pois fundamenta diversas capacidades que auxiliam em nossa sobrevivência, socialização, cognição, linguagem, expressão artística e obtenção de conhecimento, aspectos habitualmente relacionados à inteligência. É razoável afirmar, portanto, que conceder essa capacidade à máquina é um passo importante na construção de sistemas computacionais inteligentes.

Quando se trata de sistemas computacionais, os objetos que devem ser classificados são tradicionalmente representados como vetores de múltiplas dimensões, nos quais cada dimensão representa uma característica do objeto. O espaço multidimensional definido por esses vetores é denominado espaço de características. Um método de classificação, por sua vez, realiza processamento sobre os vetores e o espaço de características com o propósito de definir a que classe pertence o objeto representado. Muitos métodos de classificação já foram propostos na literatura científica, com graus variados de complexidade e qualidade. Neste trabalho, alguns classificadores simples são avaliados e comparados. São eles: 1-NN, Centroide mais Próximo, LDA e QDA e suas versões ingênuas.

OBJETIVOS

O objetivo principal deste trabalho é avaliar as vantagens e desvantagens dos classificadores 1-NN, Centroide mais Próximo, LDA e QDA e suas versões ingênuas. Entre outros atributos, serão analisados:

(a) desempenho de classificação na base Iris;

(b) validação Leave One Out, 10-Fold e Holdout;

(c) tempo de processamento para treino e classificação.

MÉTODOS DE CLASSIFICAÇÃO

Nesta seção são apresentados brevemente os métodos experimentados e uma estimativa das complexidades associadas às suas fases de treinamento e classificação. Nesta análise, n representa o número de amostras na base de dados, c é o número de classes presentes e d é o número de dimensões que compõem os vetores de características.

1-NN

O método 1-NN é um caso especial do classificador K-NN, com K = 1, portanto, consiste em atribuir ao objeto em questão a classe de seu vizinho mais próximo no espaço

de características. Não há uma fase de treinamento anterior, nem construção de um modelo representativo da distribuição de dados no 1-NN. O processo de decisão do método é concentrado na fase de classificação, na qual são utilizados diretamente os elementos cujas classes são conhecidas. Tal processo consiste em calcular medidas de distância entre o objeto a ser classificado e cada um dos objetos conhecidos. Por fim, atribui-se ao objeto desconhecido a mesma classe do objeto cuja distância é mínima.

Centroide mais Próximo

A fase de treinamento do método do centroide mais próximo consiste em determinar os centroides das distribuições dos objetos de cada classe, dentro do espaço de características, e usá-los como protótipos para suas respectivas classes. Na fase de classificação, o objeto desconhecido é comparado apenas em relação aos centroides calculados anteriormente. Essa comparação é feita tradicionalmente usando uma medida de distância entre o objeto desconhecido e o centroide. A classe atribuída será, então, a mesma classe do centroide cuja distância ao objeto desconhecido é mínima.

QDA

O classificador QDA (do inglês, *Quadratic Discriminant Analysis*) consiste na modelagem dos vetores no espaço de características como uma distribuição normal multivariada e na aplicação do teorema de Bayes para obtenção de uma função discriminante capaz de determinar uma medida de distância entre o vetor de características a ser classificado e a distribuição de cada classe.

$$d(x, c_i) = \ln |\Sigma_i| + (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) \tag{1}$$

Essa modelagem resulta na Equação 1, função baseada na distribuição gaussiana multidimensional, em que $d(x, c_i)$ é o resultado da função discriminante entre o vetor x e a classe c_i. Além disso, μ_i, $|\Sigma_i|$ e Σ_i^{-1} são, respectivamente, o centroide da classe c_i, o determinante e a inversa da matriz de covariâncias formada pelos elementos de c_i.

A fase de treinamento de um modelo QDA consiste no cálculo dos valores μ_i, $|\Sigma_i|$ e Σ_i^{-1} para cada classe. A fase de classificação, por sua vez, consiste na aplicação da função discriminante da Equação 1 para cada classe e na obtenção do valor mínimo.

LDA

O classificador LDA (do inglês, *Linear Discriminant Analysis*) pode ser compreendido como uma simplificação do QDA, na qual a matriz de covariâncias utilizada é a

mesma para todas as classes e seu cálculo baseia-se simultaneamente em todas as amostras de treinamento. A Equação 1, portanto, é substituída pela Equação 2.

$$d(x, c_i) = (x - \mu_i)^T \Sigma^{-1} (x - \mu_i) \tag{2}$$

QDA ingênuo

O classificador QDA ingênuo, tradução do termo em inglês *naive*, é uma simplificação que assume a absoluta independência entre as dimensões dos vetores de características. Tal suposição implica que as covariâncias entre diferentes dimensões são iguais a zero. O QDA ingênuo, então, utiliza a mesma equação do QDA (Equação 1), no entanto, as matrizes de covariâncias de cada classe possuem termos não-nulos apenas na diagonal principal, ou seja, apenas as variâncias em cada dimensão são necessárias.

LDA ingênuo

Utilizando o mesmo princípio de total independência entre as dimensões das amostras, o LDA ingênuo simplifica o LDA através da suposição de que as covariâncias entre dimensões diferentes são nulas. O cálculo dos discriminantes para cada classe também é dado pela Equação 2, com o uso da matriz de variâncias em vez das covariâncias.

RESULTADOS

Nesta seção são apresentados os resultados de desempenho para classificação da base Iris e os tempos médios de treino e classificação para cada um dos métodos avaliados. Nas tabelas, as siglas SN e N representam, respectivamente, os resultados obtidos sem normalização e com normalização. O resultado da validação Holdout é a média das taxas de sucesso de 20 amostragens aleatórias e divisão de 70% para treino e 30% para teste.

	1-NN		CMP		LDA		QDA	
	SN	N	SN	N	Padrão	Ingênuo	Padrão	Ingênuo
Leave One Out	96.00	94.67	92.67	84.67	85.33	84.67	96.67	95.33
10-Fold	96.00	95.33	92.67	86.00	86.00	86.00	96.67	95.33
Holdout	95.53	94.11	92.56	85.56	85.22	86.44	95.78	95.00

Tabela 1 – Desempenho de classificação na base Iris em porcentagens.

A análise da Tabela 1 permite afirmar que os métodos que obtiveram melhor desempenho na base Iris foram QDA e 1-NN, cujos resultados encontram-se no intervalo entre 94% e 97%. Em comparação, LDA e o método do Centroide mais Próximo obtiveram taxas de classificação até 10% inferiores, sendo LDA o método menos adequado para este caso, pois sua faixa de desempenho encontra-se entre 84% e 87%.

Em relação ao processo de normalização, vê-se, também na Tabela 1, que seu resultado sobre a base Iris é uma redução de desempenho. Em todos os casos observados, os resultados são superiores com o uso dos dados não-normalizados, diferença que atinge até 7% quando se considera o método do Centroide mais Próximo.

Observa-se, novamente na Tabela 1, que são muito pequenas as diferenças de desempenho entre as versões padronizadas do LDA e QDA e suas versões ingênuas. As diferenças obtidas encontram-se na faixa de $\pm 1\%$. Isso indica que a suposição de independência entre as características das amostras na base Iris, embora não seja perfeita, é uma boa aproximação.

	1-NN		CMP		LDA		QDA	
	SN	N	SN	N	Padrão	Ingênuo	Padrão	Ingênuo
Leave One Out	0	0	111	92	297	262	645	511
10-Fold	0	0	98	101	319	280	665	602
Holdout	0	0	102	100	315	282	630	543

Tabela 2 – Tempos médios de treinamento em microssegundos.

	1-NN		CMP		LDA		QDA	
	SN	N	SN	N	Padrão	Ingênuo	Padrão	Ingênuo
Leave One Out	23	24	20	20	43	42	60	40
10-Fold	21	17	13	13	30	30	44	42
Holdout	16	15	12	12	38	29	40	58

Tabela 3 – Tempos médios de classificação em microssegundos.

A análise das Tabelas 2 e 3 permite afirmar que o método mais caro em termos de tempo de processamento é o QDA, tanto em termos de treino, quanto de classificação. O segundo mais dispendioso, em ambos os aspectos, é o LDA. Vê-se que o método do Centroide mais Próximo apresenta vantagem em relação ao 1-NN no quesito tempo de classificação. Essa vantagem, no entanto, surge com o custo de um determinado tempo de treinamento, elemento não exigido no 1-NN.

Se o principal quesito for o desempenho de classificação, o QDA padrão pode ser escolhido como o vencedor desta análise. No entanto, se for feito um ponderamento em relação ao desempenho de classificação e tempo de processamento, o método mais adequado é, por conta de sua simplicidade, o 1-NN. Essas conclusões são válidas para a base Iris, que possui relativamente poucas classes, poucas amostras e baixa dimensionalidade. A depender de combinações dessas variáveis, ambos os desempenhos de classificação e tempo podem ser alterados drasticamente.