Universidade Estadual de Campinas
Instituto de Matemática, Estatística e Computação Científica
Departamento de Estatística

Avaliação 1

ME906 - Métodos em Aprendizado Supervisionado Máquina
Professor Dr. Ronaldo Dias

Nome	**RA**	**Contribuição**
Andre Saul Juarez Castro	272208	20%
Isabella Nascimento Peres da Silva	238015	20%
João Pedro de Campos Formigari	236144	20%
Mateus Lee Yu	236235	20%
Pedro Luis Rebollo	217460	20%

Campinas - SP

1 Discriminantes

A Análise Discriminante é uma ferramenta robusta para discernir padrões complexos e relações ocultas nas observações, é uma das técnicas de classificação frequentemente utilizadas. A Análise Discriminante Linear (ADL) assume que as classes em que as observações são classificadas possuem uma distribuição normal multivariada com dimensão $p \geq 1$, com médias distintas para cada classe porém com matrizes de covariância idênticas, já na Análise Discriminante Quadrática (ADQ) supomos que as matrizes de covariância são diferentes para cada classe. A função discriminante $\delta_k(x)$ na ADL e ADQ é uma equação que nos permite calcular a probabilidade de uma observação pertencer a uma classe específica, ela será utilizada como o classificador.

Suponha que temos a seguinte amostra $\{(\mathbf{X}^{(i)}_{p\times 1}, Y^{(i)})^n_{i=1}\}$, onde o vetor $\mathbf{X}^{(i)}_{p\times 1} = [X_1 \cdots X_p]^T$ representa as p características e $Y^{(i)}$ classifica qual das K classes a i-ésima observação pertence. O Classificador de Bayes é uma regra de decisão que atribui cada observação para a classe mais provável dado as informações $\mathbf{x_{p\times 1}}$, a fórmula do Classificador de Bayes é dada pela equação 1.

$$f(\mathbf{x}) = \arg\max_k P(\mathbf{X} = \mathbf{x}|Y = k)P(Y = k) \tag{1}$$

Na prática estimamos $P(Y = k)$ como a proporção de observações presentes na nossa amostra que pertence a classe k.

$$P(Y = k) = \frac{\sum_{i=1}^n \mathbb{1}_{\{y^{(i)}=k\}}(y^{(i)})}{n} = \pi_k$$

O termo $P(\mathbf{X} = \mathbf{x}|Y = k)$ também precisa ser estimado e a sua fórmula irá depender das suposições da ADL e ADQ.

2 Análise Discriminate Linear

2.1 Introdução

O LDA (*Linear Discriminant Analysis Multiclass*, em portugês, ADL ou Análise Discriminate Linear), é um método de classificação, também visto por alguns como técnica de redução de dimensionalidade, obtido a partir de uma generalização da técnica desenvolvida por Ronald A. Fisher em 1936 [2], chamada de Linear Discriminant or Fisher's Discriminant Analysis, método de classificação de duas classes (*two-class*). A generalização deste para uma técnica multicasse (*multiclass*) foi feita por C. R. Rao em 1948 [3], se tornando então esta o ADL Multiclasse que segue sendo amplamente usado até os dias de hoje em áreas como reconhecimento de imagem, estudos biomédicos, análises preditivas e muitos outros, o qual será apresentado nessa seção do presente trabalho.

2.2 Metodologia

Nesta seção será apresentada a metodologia empregada para a utilização do classificador LDA.

Em um cenário que temos apenas um preditor assumimos que X segue uma distribuição Normal, $\mathcal{N}(\mu_k, \sigma^2)$, onde μ_k é a média da k-ésima classe e σ^2 é a variância comum para todas as K classes. Ou seja,

temos que $f_k(x)$ é dado por

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(x-\mu_k)^2) \tag{2}$$

Utilizando essa densidade no Teorema de Bayes apresentado anteriormente tem-se a seguinte regra de decisão, aloca-se uma nova observação $X = x$ a classe em que foi obtido o maior $\delta_k(x)$, que vai possuir a forma:

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k}{2\sigma^2} + \log \pi_k \tag{3}$$

Porém como não conhecemos os valores dos parâmetros μ_k, π_k e σ^2 na prática utilizamos estimativas dos mesmos, essas são dadas a seguir.

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i \tag{4}$$

$$\hat{\pi}_k = n_k/n \tag{5}$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{k} \sum_{y_i=k} (x_i - \hat{\mu}_k)^2 \tag{6}$$

Colocamos então as estimativas na equação 3 dada anteriormente, resultando em:

$$\hat{\delta}_k(x) = x\frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}}{k} 2\hat{\sigma}^2 + \log \hat{\pi}_k \tag{7}$$

Vale apontar ao fato de que a regra de decisão depende de x somente através de uma combinação linear de elementos, e por isso o método LDA leva a palavra "linear".

Já no caso de múltiplos preditores para usar o LDA primeiro é necessário realizarmos a suposição de que $X = (X_1, X_2, ..., X_p)$ segue uma distribuição Normal multivariada (ou Gaussiana multivariada) com um vetor de médias de tamanho p específico por classe e uma martriz de covariância de tamanho $p \times p$ comum para as K classes, ou seja, $\mathcal{NM}_p(\mu_k, \mathbf{\Sigma})$. Consequentemente tem-se que uma observação pertencente a k-ésima classe tem distribuição $\mathcal{NM}_p(\mu_k, \mathbf{\Sigma})$, onde μ_k é o vetor de médias especificado por classe e $\mathbf{\Sigma}$ é a matriz de covariância comum a todas as classes. Ou seja, faremos o uso da densidade da Normal multivariada da k-ésima classe, dada por:

$$f_k(x) = \frac{1}{(2\pi)^{\frac{p}{2}} |\mathbf{\Sigma}|^{\frac{1}{2}}} \exp(-\frac{1}{2}(x-\mu_k)^T \mathbf{\Sigma}^-1(x-\mu_k)) \tag{8}$$

Coloca-se esta no Teorema de Bayes (equação 2) e após algumas manipulações algébricas teremos que o classificador de Bayes irá alocar uma observação $X = x$ a classe em que $\delta_k(x)$ for maior, sendo este dado por:

$$\delta_k(x) = x^T \mathbf{\Sigma}^{-1} \mu_k^T - \frac{1}{2} \mu_k^T \mathbf{\Sigma}^{-1} \mu_k + \log \pi_k \tag{9}$$

Porém, na prática não conhcemos os valores dos parâmetros $\mu_1, \mu_2, ..., \mu_K$, $\pi_1, \pi_2, ..., \pi_K$ e $\mathbf{\Sigma}$ então

precisamos estimá-los, para tal utilizaremos as seguintes fórmulas:

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i \tag{10}$$

$$\hat{\pi}_k = n_k/n \tag{11}$$

$$\hat{\mathbf{\Sigma}} = \sum_{k=1}^{K} \frac{1}{n-K} \sum_{i:y_i=k} (x_i - \hat{\mu_k})(x_i - \hat{\mu_k})^T \tag{12}$$

Depois colocamos essas estimativas na equação 9, resultando em:

$$\hat{\delta}_k(x) = x^T \hat{\mathbf{\Sigma}}^{-1} \hat{\mu}_k^T - \frac{1}{2} \hat{\mu}_k^T \hat{\mathbf{\Sigma}}^{-1} \hat{\mu}_k + log\hat{\pi}_k \tag{13}$$

Desta forma o LDA aloca uma nova observação $X = x$ à classe a qual $\hat{\delta}_k(x)$ é maior.

2.3 Vantagens

Este método possui suas vantagens, ou pontos positivos, como um destes podemos citar primeiramente que o ADL possui um algoritmo simples e portanto rápido de ser utilizado, considerando-se o esforço computacional necessário para sua execução.

Além disso, o mesmo ainda pode ser utilizado com diversas finalidades, de forma que por mais que seja a princípio um classificador pode ser utilizado para separação linear de observações, distinção de variáveis de um conjunto de dados, redução de dimensionalidade, entre outros, acarretando em seu possível uso em quaisquer áreas de interesse.

2.4 Desvantagens

Como ponto negativo desse classificador podemos considerar a sua suposição de normalidade multivariada e univariada das variáveis preditoras, isto porque, esta nem sempre está presente nos dados e por mais que alternativas sejam testadas, como por exemplo a transformação dessas variáveis, ou o uso do TLC (quando a amostra for grande suficiente) nem sempre essa poderá ser verificada, assim impedindo a execução correta do método.

2.5 Aplicação

Com o intuito de entender de forma mais prática sobre a Análise Discriminante Linear (LDA), o método foi aplicado a um banco de dados que é resultado de uma análise química de vinhos cultivados numa mesma região, mas derivados de três colheitas diferentes. A análise determinou as quantidades de 13 constituintes encontrados em cada um dos três tipos de vinhos.

Após uma análise prévia, foram escolhidas 4 variáveis que foram significativas ao modelo de classificação, sendo elas Álcool, Flavanoides, Intensidade de Cor e Prolina. Após essa análise, o banco foi separado em treino e teste e o método de LDA foi aplicado.

A Tabela 1 apresenta os resultados obtidos pelo modelo treinado no conjunto de teste. Nota-se que a precisão de acertos foi bem alta, isso se deve principalmente ao fato de que o conjunto de dados é "bem comportado".

Tabela 1: Matriz de Confusão - LDA

	1	2	3
1	48	1	0
2	2	55	9
3	0	0	42

Com o método de LDA, obtivemos uma acurácia de mais de 92% na classificação das colheitas em que os vinhos foram produzidos, tendo acertado a classificação de 145 dos 157 testes feitos.

3 Análise Discriminante Quadrática

3.1 Introdução

A Análise Discriminante Quadrática (ou ADQ, em inglês *Quadratic Discriminat Analysis* - QDA) foi desenvolvida com base no ADL, nessa temos suposições semelhantes em comparação com este último citado, porém supomos que as matrizes de covariância são diferentes para cada classe, isto é, $\mathbf{X}|Y = k \sim \mathcal{NM}_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$.

3.2 Metodologia

Substituindo a função de densidade de probabilidade e a estimativa probabilidade de cada classe na equação 1, e após algumas manipulações algébricas obtemos a função discriminante da QDA.

$$\begin{aligned} f(\mathbf{x}) &= \arg\max_k P(\mathbf{X} = \mathbf{x}|Y = k)P(Y = k) \\ &= \arg\max_k (2\pi)^{-\frac{p}{2}}|\boldsymbol{\Sigma}_k|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k))\right\}\pi_k \end{aligned}$$

Note que $(2\pi)^{-\frac{p}{2}}$ é uma constante.

$$f(\mathbf{x}) = \arg\max_k |\boldsymbol{\Sigma}_k|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k))\right\}\pi_k$$

Aplicando log em ambos os lados da equação.

$$\begin{aligned} \log f(\mathbf{x}) &= \arg\max_k \ -\frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}_k^{-1}(\mathbf{x} - \boldsymbol{\mu}_k) + \log\pi_k \\ &= \arg\max_k \ -\frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}_k^{-1}\mathbf{x} + \frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k + \frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k + \log\pi_k \end{aligned}$$

Note que $\boldsymbol{\Sigma}_k^{-1}$ é uma matriz simétrica, então temos que $\frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\mathbf{x} = \frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k$.

$$\log f(\mathbf{x}) = \arg\max_k \ -\frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}_k^{-1}\mathbf{x} + \mathbf{x}^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k + \log\pi_k$$

Portanto a função discriminante (quadrática em $\mathbf{x}$) possui a seguinte equação.

$$\delta_k(\mathbf{x}) = \frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}\mathbf{x}^T\boldsymbol{\Sigma}_k^{-1}\mathbf{x} + \mathbf{x}^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\mu}_k + \log\pi_k$$

3.3 Aplicação

Analogamente ao caso da LDA, foi usado o conjunto de dados de vinho para testar o métdo de QDA, utilizando as mesmas varíaveis e conjuntos de treino e teste que no caso anterior.

A Tabela 2 apresenta os resultados obtidos pelo modelo treinado de QDA no conjunto de teste. Nota-se que a precisão, assim como na LDA, foi bem alta, porém menor que a anterior.

Tabela 2: Matriz de Confusão - QDA

	1	2	3
1	49	0	0
2	17	43	6
3	2	0	40

Com o método de QDA, obtivemos uma acurácia de mais de 84% na classificação das colheitas em que os vinhos foram produzidos, tendo acertado a classificação de 132 dos 157 testes feitos.

4 Regressão Logística Penalizada

4.1 Introdução

A regressão logística é uma técnica poderosa utilizada para modelar a probabilidade de uma variável de resposta binária. No entanto, quando temos muitos preditores ou preditores altamente correlacionados, a estimativa padrão pode resultar em sobreajuste. A regressão logística penalizada, que introduz um termo de penalização, ajuda a abordar essas situações.

4.2 Modelo Básico de Regressão Logística

A regressão logística modela a probabilidade $P(Y = 1|X)$ como:

$$P(Y = 1|X) = \frac{\exp(\beta_0 + \beta_1 X_1 + ... + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + ... + \beta_p X_p)} \tag{14}$$

Isso também pode ser representado matricialmente como:

$$p = \frac{\exp(\mathbf{X}^T\boldsymbol{\beta})}{1 + \exp(\mathbf{X}^T\boldsymbol{\beta})} \tag{15}$$

4.3 Penalização

Na regressão logística penalizada, introduzimos um termo de penalização na função de verossimilhança para regularizar ou encolher os coeficientes. A função de verossimilhança é definida como:

$$\ell(\boldsymbol{\beta}) = \mathbf{y}^T\log\mathbf{p} + (\mathbf{1} - \mathbf{y})^T\log(\mathbf{1} - \mathbf{p}) \tag{16}$$

Com a penalização, ela se torna:

$$J(\beta) = -\sum_{i=1}^{n}[y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] + \lambda \sum_{j=1}^{p} |\beta_j|^q \tag{17}$$

Onde p_i é a probabilidade prevista para a i-ésima observação, λ é o parâmetro de penalização, e q define o tipo de penalização. Quando $q = 1$, temos a penalização Lasso; quando $q = 2$, temos a penalização Ridge.

4.4 Regressão Logística Lasso

A regressão logística Lasso é uma extensão da regressão logística que incorpora uma penalização L_1 nos coeficientes Tibshirani (1996). Esta penalização pode conduzir a alguns coeficientes para zero, tornando-a uma ferramenta útil não só para regularização, mas também para seleção de variáveis.

Formalmente, a função objetivo para a regressão logística Lasso é:

$$J_{\text{lasso}}(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \tag{18}$$

Onde λ é o parâmetro de penalização. Como na regressão Ridge, λ determina a força da penalização: valores maiores de λ resultarão em mais coeficientes sendo reduzidos para zero.

4.5 Regressão Logística Ridge

Hoerl & Kennard (1970) propuseram o método de regularização quadrática conhecido como regressão ridge, e le Cessie & van Houwelingen (1992) aplicaram a regressão logística ridge a um problema biomédico. A regressão ridge encolhe os coeficientes da regressão, restringindo as somas dos quadrados dos coeficientes. Assim, a verossimilhança ridge log é definida como:

$$J_{\text{ridge}}(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \frac{1}{2}\lambda \sum_{i=1}^{p} \beta_i^2 \tag{19}$$

Onde $\ell(\boldsymbol{\beta})$ é a verossimilhança não penalizada e λ determina o tamanho da penalização ridge. Com coeficientes $\boldsymbol{\beta} = (\beta_1, ...\beta_p)^T$, a verossimilhança log ridge pode ser reformulada como:

$$J_{\text{ridge}}(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \frac{1}{2}\lambda \|\boldsymbol{\beta}\|_2^2 \tag{20}$$

Apesar de a verossimilhança log ridge não ser realmente uma verossimilhança, neste trabalho, as verossimilhanças penalizadas das regressões serão referidas como verossimilhanças.

4.6 Diferenças

A principal diferença entre Lasso e Ridge é a forma da penalização. Enquanto Ridge penaliza com o quadrado dos coeficientes (penalização L_2), Lasso penaliza com o valor absoluto dos coeficientes (penalização L_1). Esta diferença torna o Lasso capaz de realizar seleção de variáveis ao forçar coeficientes para exatamente zero sob a penalização adequada.

Tibshirani (1996) introduziu o Lasso como um método para "encolher"ou regularizar coeficientes em modelos de regressão. Desde então, tem sido amplamente utilizado em muitas aplicações, especialmente quando o número de preditores é grande e se suspeita que muitos deles sejam irrelevantes ou reduntantes.

4.7 Aplicação

Analogamente aos dois casos anteriores (LDA e QDA), foi usado o conjunto de dados sobre vinhos com as mesmas características para testar o modelo de regressão logística penalizada. Foram ajustados dois modelos, um para a regressão logística Lasso e outra para a Ridge. As Figuras 1 e 2 mostram a comparação entre valores de lambda e o desvio do modelo, nota-se que valores baixos de lambda são superiores em ambos os casos.

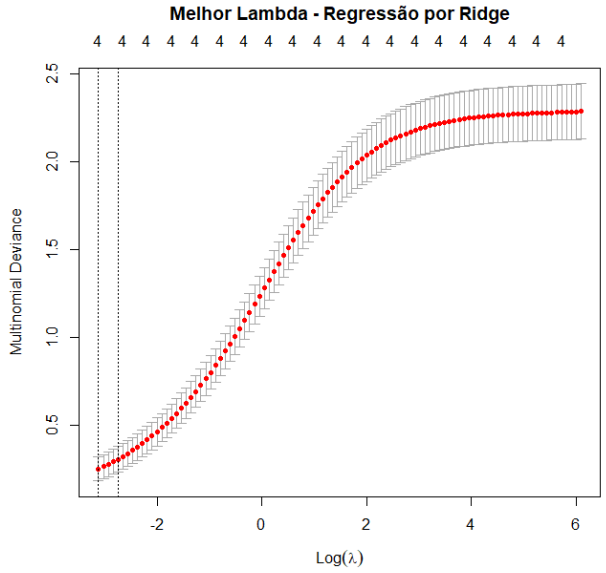

Figura 1: Comparação de Lambdas por desvio - Ridge

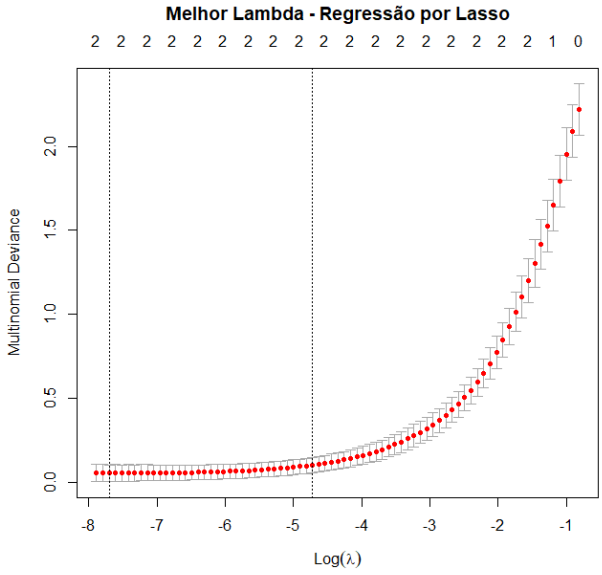

Figura 2: Comparação de Lambdas por desvio - Lasso

A Tabela 3 apresenta as matrizes de confusão de Ridge e Lasso respectivamente, nota-se uma grande precisão em ambos os casos, sendo que a Ridge acertou em 147 dos 157 testes e a de Lasso acertou em 142 de 157 testes.

Tabela 3: Matriz de Confusão - Ridge e Lasso

	1	2	3
1	49	0	0
2	4	56	6
3	0	0	42

	1	2	3
1	49	0	0
2	4	51	11
3	0	0	42

Dessa forma, vemos que a acurácia desses modelos são respectivamente 93,6% e 90,4%.

5 Avaliação das Técnicas de Classificação

A matriz de confusão é uma das formas para avaliarmos a performance de um classificador, após ajustarmos um modelo, realizamos as predições em um conjunto de teste de tamanho n e sumarizamos em uma tabela $K \times K$, onde a entrada na i-ésima linha e j-ésima coluna representa o número de observações pertencentes a classe j que foram classificadas na classe i. A matriz de confusão pode ser encontra na tabela 4.

	Classe Verdadeira		
Classe Predita	Classe 1	$\cdots$	Classe K
Classe 1	$n_{1,1}$	$\cdots$	$n_{1,K}$
$\vdots$	$\vdots$	$\ddots$	$\vdots$
Classe K	$n_{K,1}$	$\cdots$	$n_{K,K}$

Tabela 4: Matriz de Confusão

A acurácia das predições pode ser calculada somando os elementos da diagonal principal e dividindo pelo número de observações do conjunto de teste.

$$\text{Acurácia} = \frac{1}{n} \sum_{i=1}^{K} n_{i,i}$$

Referências

[1] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (1st ed.). Springer.

[2] Fisher, R.A. (1936), The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, 7: 179-188.

[3] Rao, C. R. (1948). The Utilization of Multiple Measurements in Problems of Biological Classification. Journal of the Royal Statistical Society. Series B (Methodological), 10(2), 159–203.

[4] Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.