Estratégias para Classificação Binária
Um estudo de caso com classificação de e-mails

Eduardo Elias Ribeiro Junior *

04 de julho de 2016

Resumo

Em Machine Learning têm-se em diversas situações o interesse em realizar predições a partir de algoritmos computacionais que independam da ação humana. Uma das mais comuns tarefas preditivas no campo aplicado é a de classificação. Neste trabalho apresentamos um rol de técnicas de classificação binária aplicadas a um conjunto de dados do repositório UCI Machine Learning que refere-se a classificação de e-mails em `spam` ou `não-spam`. As técnicas de classificação apresentadas e aplicadas permeiam os campos de Estatística Multivariada, Machine Learning e Inferência Paramétrica. Foram ao todo 11 técnicas de classificadas sob o qual a abordagem via Random Forest (árvores de decisão aleatórias) apresentou o melhor desempenho considerando resumos da curva ROC obtidos de classificações na base de teste e nas amostras de validação cruzada.

Palavras-chave: *Classificação, Análise Discriminante, Regressão Logística, Árvores de decisão, Random Forest, Bagging, Boosting, SVM.*

Sumário

1 Introdução **2**

2 Material e Métodos **2**

3 Resultados **6**

3.1 Discriminant Analysis-Based 6
3.2 Generalized Linear Model-Based 7
3.3 Classification Trees-Based 9
3.4 Support Vetor Machine-Based 10
3.5 Comparação das abordagens 11

4 Conclusões **13**

5 Referências **14**

*Universidade Federal do Paraná - DEST, edujrrib@gmail.com

1 Introdução

Em Estatística aplicada pode-se destacar dois principais interesses a cerca da análise de dados, são eles: i) Compreender o relacionamento entre variáveis de interesse e características de uma amostra e; ii) Realizar predições por meio de métodos estatísticos ajustados por dados de uma amostra.

Na área de Aprendizado de Máquina (Machine Learning), o segundo tópico citado é predominante. Nessa área têm-se interesse em obter algoritmos computacionais que independam da ação humana, ou seja, que permitam o computador aprender. Para tal finalidade diversas ferramentas foram propostas não se restringindo a modelagem estatísticas da forma convencional (adotando um modelo de probabilidades para a variável de interesse condicionada às covariáveis, cuja há uma relação funcional entre essas variáveis). BREIMAN (2001) discute a excessiva utilização de modelos estatísticos em contraste com as abordagens presentes no aprendizado de máquina.

Tanto em Aprendizado de Máquina quanto na Estatística em geral, no contexto univariado, as ferramentas para análise são específicas à classe da variável resposta. Para variáveis quantitativas (contínuas ou discretas) têm-se interesse em predizer o valor da variável para uma nova observação não presente na amostra, e.g. o preço de uma ação do mercado financeiro, já para variáveis qualitativas o interesse está na classificação de novas observações nas classes da variável em estudo, e.g. classificar o estado de uma doença.

Neste trabalho aborda-se estratégias para classificação no caso de uma variável qualitativa binária, que contém apenas duas classificações. Essas estratégias permeiam os campos de modelagem estatística da forma convencional e algoritmos da área de aprendizado de máquina. Essas abordagens são descritas na Seção 2. O conjunto de dados, sob o qual aplica-se os métodos para classificação, refere-se a e-mails recebidos por funcionários de uma empresa onde deseja-se classificá-los como spams ou não-spams, detalhes sobre o conjunto de dados são descritos na Seção 2.

Após a aplicação dos métodos é de interesse no trabalho avaliar o desempenho dos classificadores e os compará-los quanto ao poder preditivo. Para tal finalidade explora-se a curve ROC (*Receiver Operating Characteristic*), um popular gráfico de exibe simultaneamente os dois tipos de erros em todos os limites possíveis (JAMES et al., 2013). Resumos obtidos a partir desta curva são utilizados para comparação de classificadores, em geral o valor de AUC (*Area Under Curve*) é exaustivamente apresentado e utilizado como critério de avaliação.

O trabalho é organizado em cinco seções. Na Seção 1 contextualiza-se o problema de classificação e as abordagens usualmente aplicadas. Na seção 2 o conjunto de dados e os métodos para obtenção dos classificadores são apresentados. A seção 3 é destinada à exibição e comparação dos resultados obtidos dos classificadores ajustados, ainda nesta seção discuti-se particularidades nos resultados e identifica-se o melhor classificador para o conjunto de dados em análise. Na seção 4 são apresentados as conclusões obtidas no estudo e alguns possíveis tópicos que ainda podem ser abordados. Na seção 5 as referências bibliográficas que embasam o estudo são apresentadas.

2 Material e Métodos

Para aplicação e competição dos métodos de classificação, que são descritos adiante, considerou-se o conjunto de dados disponibilizado por George Forman no repositório UC Irvine Machine Learning (LICHMAN, 2013). Os dados referem-se a mensagens de e-mails recebidas por funcionários da empresa

Hewlett-Packard - HP[1], cujo principal objetivo é obter um bom classificador de mensagens para **spams**, e-mails indesejados (anúncios de sites de produtos, propostas para dinheiro rápido, correntes, pornografia, entre outros), ou **não-spams**.

No conjunto de dados há 4601 e-mails registrados e para cada e-mail têm-se a informação de sua classificação como spam ou não-spam. A proporção de spams é exibida na Figura 1, onde nota-se que a maioria dos e-mails são não-spam. Porém, o percentual de e-mails classificados como spams não é tão baixo, em números absolutos são 1813, o que viabiliza a aplicação de métodos para classificação.

Além da informação sobre a classificação do e-mail também são disponibilizadas outras informações com características do e-mail. Todas as variáveis presentes no conjunto de dados são descritas na Tabela 1. São 58 variáveis, sendo que 48 delas se referem ao percentual de ocorrência de uma palavra no e-mail, e.g. `make` representa o percentual de ocorrências da palavra "make". Para essas variáveis há um considerável excesso de zeros.

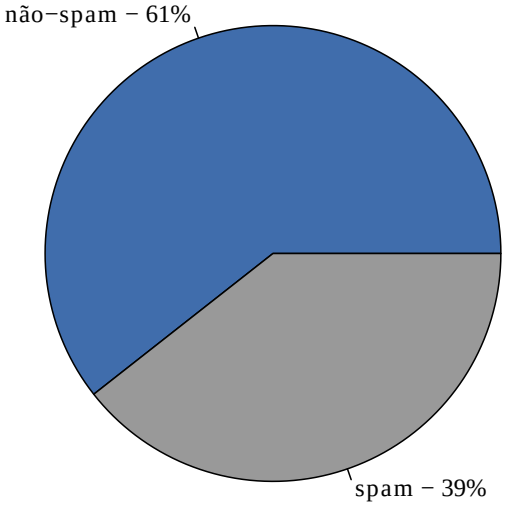

Figura 1: Proporção de e-mails classificados como spams e não-spams

Para evitar a escolha de classificadores que ajustem de forma demasiada à amostra sob a qual foram treinados (*overfit*), adotou-se a divisão aleatória do conjunto de dados. Duas bases foram constituídas, uma para ajuste dos classificadores com 70% dos e-mails (3220) e outra com 30% (1381 e-mails) para avaliar o desempenho dos classificadores ajustados. Como tentativa de preservar as características dos e-mails na base de treino, manteve-se o mesmo percentual de spams e não-spams nas partições do conjunto de dados.

Para obtenção dos classificadores são utilizados métodos de classificação que foram seccionados em quatro grandes grupos da área, os métodos fundamentados em: **Discriminant Analysis**; **Generalized Linear Model**; **Classification Trees**; e **Support Vetor Machine**. A seguir esses métodos são brevemente descritos.

Discriminant Analysis

A análise discriminante é uma técnica da estatística multivariada que surgiu das contribuições de Fisher à área. É uma das técnicas mais antigas e mais empregadas para classificação.

Seja $\Omega_1, \Omega_2, \ldots, \Omega_g$ populações assume-se que são normalmente distribuídas com vetores de média desconhecidos e mesma matriz de covariâncias. Ainda, considere X_j a matriz de dimensão $n_j \times p$, com as p covariáveis das n_j observações pertencentes à j-ésima amostra. A regra de classificação será

$$\text{alocar } x \text{ para } \Omega_j \text{ se } j = \underset{i \in (1,2,\ldots,g)}{\arg\max} \left\{ \log(\pi_i) - \frac{1}{2}(x - \bar{x})^t \Sigma^{-1} (x - \bar{x}) \right\}$$

[1] Website da empresa http://www.hp.com/

Tabela 1: Descrição das variáveis disponíveis no conjunto de dados

Informação provida	**Variáveis**	**Tipo**
Percentual de ocorrências da palavra no e-mail	`make`, `address`, `all`, `num3d`, `our`, `over`, `remove`, `internet`, `order`, `mail`, `receive`, `will`, `people`, `report`, `addresses`, ...	númerica
Percentual de ocorrências do caractere no e-mail	`charSemicolon`, `charRoundbracket`, `charSquarebracket`, `charExclamation`, `charDollar`, `charHash`	númerica
Comprimento médio das sequencias com letras maiúsculas	`capitalAve`	númerica
Comprimento da maior sequencia com letras maiúsculas	`capitalLong`	númerica
Número total de letras maiúsculas no e-mail	`capitalTotal`	númerica
Classificação do e-mail em spam	`type`	binária

em que π_i é uma probabilidade a priori de que a observação sob teste pertença a população Ω_i, neste trabalho adotaremos como probabilidade a priori a proporção de observações em cada população. Esse classificador é chamado de discriminante linear de Fisher.

Uma variação do discriminante linear ocorre quando não se considera que as matrizes de variância e covariância são iguais e assim a função que determina a regra de decisão ganha mais um termo flexibilizando a fronteira de decisão.

Além dos métodos de análise discriminante linear e quadrática também serão abordados alguns métodos relativamente mais recentes para obtenção de classificadores fundamentados em análise discriminante, são eles **Análise Discriminante Regularizada - RDA** (do inglês *Regularized Discriminante Analysis*) e **Análise Discriminante Penalizada - PDA** (do inglês *Penalized Dicriminant Analysis*). Para RDA adiciona-se dois parâmetros, que são arbitrariamente escolhidos, à função que determina a regra de decisão. Estes parâmetros ponderam essa função flexibilizando sua forma. Na metodologia PDA atribui-se penalidades aos vetores discriminantes de Fisher, ou seja, maximiza-se o núcleo da verossimilhança para cada grupo sujeito a uma restrição imposta arbitrariamente. Essa abordagem surgiu para problemas *"small n large p"*, portanto nas aplicações para esse conjunto de dados não espera-se grandes diferenças. Porém ressalta-se que este método é de grande valia, pois em casos $p > n$ as outras abordagens não funcionam.

Generalized Linear Model

Dos modelos pertencentes a classe dos modelos lineares generalizados (*Generalized Linear Models*) será utilizados somente o modelo denominado modelo logístico, cujo a distribuição considerada para a relação condicional $Y \mid X$ é Binomial(m, π) e função de ligação logito (que dá nome ao modelo). Assim o modelo pode ser escrito, juntamente com sua função de verossimilhança a ser maximizada, conforme abaixo:

$$Y \mid X_i \sim \text{Binomial}(m_i, \pi_i)$$
$$\log\left(\frac{\pi}{1-\pi}\right) = X\beta$$

$$\mathscr{L}(\beta; \underline{y}) = \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i}$$

sendo $\pi_i = \frac{e^{x_i\beta}}{e^{x_i\beta}+1}$. Assim obtendo as estimativas dos β's a partir da maximização de $\mathscr{L}$ podemos calcular π_i. A classificação do i-ésimo indivíduo seguirá a regra: se $\pi_i < p_c$, classifica no grupo 0 e se $\pi_i \geq p_c$ classifica no grupo 1. O valor de p_c é arbitrário, porém pode-se escolher o valor de p_c que confere a maior especificidade e sensibilidade.

Neste trabalho aborda-se também a estimação dos parâmetros utilizando a metodologia de *Gradiente Boosting* que, de forma resumida, reponderam iterativamente a amostra atribuindo maiores pesos às observações classificadas de forma incorreta na iteração anterior HOFNER et al. (2014).

Classification Trees

Os métodos de classificação fundamentos em árvores de decisão ganharam espaço no campo da Estatística aplicada, principalmente, a partir dos anos de 1990. Esse método é uma extensão dos modelos de regressão e não é restrito à classificação. Em síntese o método se baseia na estratificação binária das covariáveis que levam a decisões, conforme ilustrado na Figura 2.

Neste trabalho serão utilizados métodos de classificação que aprimoram as árvores de decisão. O primeiro deles é o procedimento **bagging** que consiste em reamostrar os dados de treino, obter um classificar para cada conjunto reamostrado e tomar como novo desfecho dos nós terminais as classes modais das reamostras classificadas. Com isso diminui-se a variância do classificador. Uma modificação no procedimento de bagging, fazendo com que cada árvore gerada pelas reamostras tenha preditores distintos, leva o nome de **Random Forest** que também serão aplicadas.

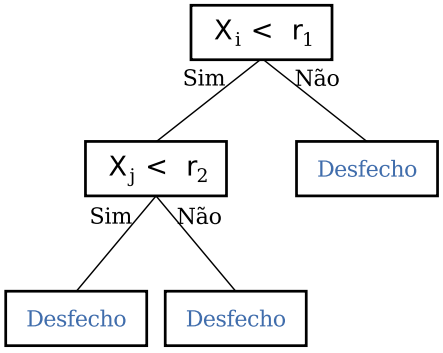

Figura 2: Ilustração de árvores de decisão

Support Vector Machines

Os métodos de classificação baseados em Support Vector Machines - SVM são construídos basicamente pela interpretação geométrica do problema. Dispondo as observações de um problema de classificação em um hiperplano de dimensão p, SVM procuram maximizar as margens do subespaço $p-1$ desse hiperplano que melhor separam as observações.

Neste trabalho serão aplicados os métodos chamados de Support Vector Classifier que permitem classificações incorretas com relação às margens por meio de um parâmetro adicional C que define a magnitude total deste erro. Além disso, também serão obtidos classificadores SVM com diferentes núcleos (*kernels*) conforme exibido abaixo:

- Linear: $K(x_i, x_k) = \langle x_i, x_k \rangle$
- Polinomial: $K(x_i, x_k) = (1 + \gamma \langle x_i, x_k \rangle)^d$
- Gaussiano: $K(x_i, x_k) = \exp(-\sigma \|x_i, x_k\|^2)$

Os parâmetros que definem as expansões kernel são arbitrários. Nas análises se faz a avaliações de classificadores com diferentes valores dos parâmetros para escolhê-los.

Para comparação dos classificadores obtidos com os diferentes métodos apresentados será feita a avaliação da curva ROC construída com os resultados das classificações na base de teste. Resumos da curva ROC, como a área abaixo da curva - AUC, acurácia, sensibilidade, especificidade, valor preditivo positivo - PPV (*Positive Predictive Values*) e negativo - NPV (*Negative Predictive Values*) são utilizados. Abaixo exibe-se os cálculos para cada um deles, conforme KUHN (2008).

Considere a seguinte matriz de classificação.

Observado	**Predito**	
	`não-spam`	`spam`
`não-spam`	A	C
`spam`	B	D

- `acurácia` = $\frac{A+D}{A+B+C+D}$
- `sensibilidade` = $\frac{A}{A+C}$
- `especificidade` = $\frac{D}{B+D}$
- `PPV` = $\frac{A}{A+B}$
- `NPV` = $\frac{D}{C+D}$

A `AUC` é calculada conforme utilizando o método de integração por trapézios.

No trabalho também se faz uso do procedimento de validação cruzada *10-fold*, ou seja, ainda na base de treinamento se divide a amostra em dez partes utilizando 9 para ajuste do classificador e uma para avaliação. Isso é feito considerando 10 vezes, considerando um conjunto de 9 amostras diferentes a cada vez. Ainda repete-se esse procedimento 3 vezes para minimizar o erro de escolha de um classificador sobreajustado. Assim têm-se 31 classificadores para cada técnica aplicada, 30 referentes as amostras da validação cruzada e 1 considerando toda a base de treinamento.

3 Resultados

Nesta seção são apresentados e discutidos os resultados provenientes dos classificadores, obtidos com os métodos citados na Seção 2, aplicados no conjunto de teste.

Primeiramente são apresentados e comparados os classificadores de mesma do mesmo grupo, o classificador que obteve o melhor desempenho na comparação dentro do grupo foi mantido para comparação posterior entre os grupos.

Todas as análises são realizadas com o pacote `caret` do R, que é um *wrapper* para outros pacotes do R. A facilidade de utilizar as funções deste pacote é que os resultados são padronizados facilitando a comparação.

3.1 Discriminant Analysis-Based

Os classificadores ajustados com base na metodologia de análise discriminante são:

- `LDA`: Linear Discriminant Analysis
- `QDA`: Quadratic Discriminant Analysis
- `RDA`: Regularized Discriminant Analysis
- `PDA`: Penalized Discriminant Analysis

Foram testados diferentes valores para os parâmetros λ e γ, na `RDA` e diferentes λ na `PDA`, a fim de se obter o conjunto de parâmetros ótimos, o chamado *tunning* em Aprendizado de Máquina.

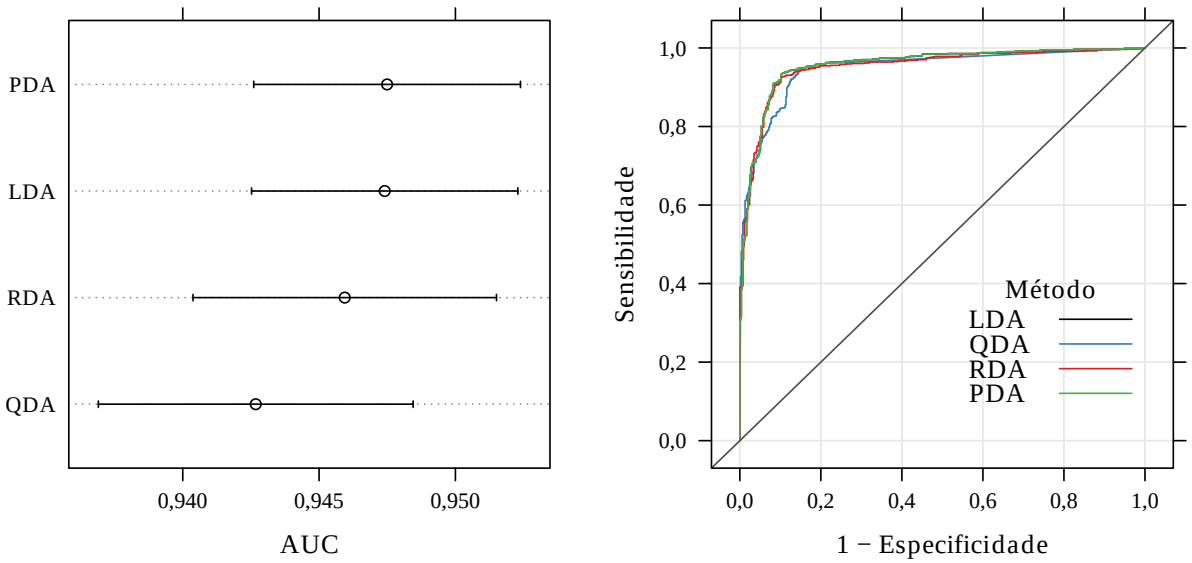

Figura 3: (Esquerda) Intervalos de confiança para a área abaixo da curva ROC baseados nas 3 repetições das 10 amostras de validação cruzada. (Direita) Curva ROC dos classificados aplicados à base de teste.

Observando a Figura 3, não há nenhum método que se destaca. Os intervalos de confiança construídos a partir das 30 classificações feita na validação cruzada se sobrepões e os valores médios de AUC são muito próximos. Considerando as curvas ROC dos classificadores com todas as observações da base de treinamento, também não há grandes diferenças, destaca-se apenas uma menor sensibilidade quando considerado o classificador `QDA`, conforme pode ser visto na Tabela 6.

Tabela 2: Resumos da curva ROC dos classificadores fundamentados em Análise Discriminante

Medida	**LDA**	**QDA**	**RDA**	**PDA**
AUC*	0,947 (0,943, 0,952)	0,943 (0,937, 0,948)	0,946 (0,94, 0,952)	0,947 (0,943, 0,952)
Acurácia*	0,899 (0,882, 0,915)	0,825 (0,804, 0,844)	0,906 (0,889, 0,921)	0,899 (0,882, 0,915)
Sensibilidade	0,955	0,744	0,941	0,955
Especificidade	0,814	0,949	0,851	0,814
PPV	0,888	0,957	0,907	0,888
NPV	0,921	0,707	0,904	0,921
AUC	0,956	0,948	0,952	0,956

*Valor com intervalo de confiança de 95% baseado nas 30 amostras de validação cruzada apresentado entre parênteses.

Outro resultado que chama a atenção é a similaridade dos resultados de `LDA` e `PDA`. Isso pode ser atribuído ao fato de que não há grande número de covariáveis na base ao ponto que as penalidades não afetam significativamente o ajuste do classificador.

Assim adotamos como classificador representante da abordagem por análise discriminante o `LDA`.

3.2 Generalized Linear Model-Based

Para os classificadores baseados em modelos lineares generalizados são tomados duas abordagens:

- `GLM-MLE`: Generalized Linear Models ajustado via maxima verossimilhança; e

- `GLM-Boost`: Gradiente Boosting aplicado em GLM

Para a abordagem Boosting também realizou-se o *tunning* para o parâmetro `mstop` que determina o número de iterações. O valor que forneceu o melhor desempenho foi de 100 iterações.

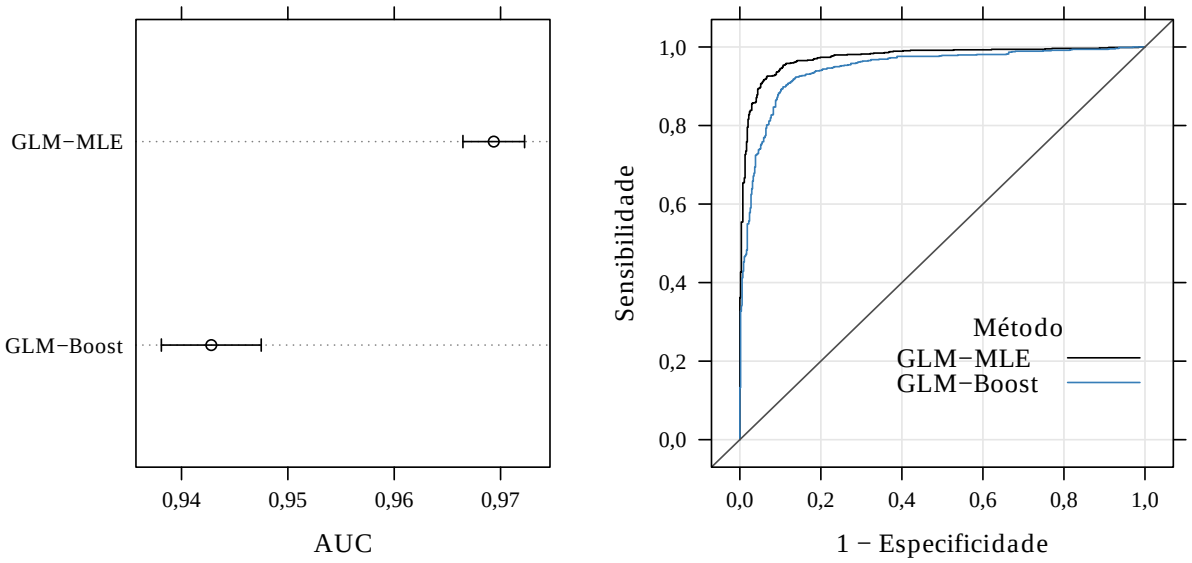

Figura 4: (Esquerda) Intervalos de confiança para a área abaixo da curva ROC baseados nas 3 repetições das 10 amostras de validação cruzada. (Direita) Curva ROC dos classificados aplicados à base de teste.

Na avaliação dos dois classificadores nota-se uma grande diferença de desempenho em favor do método via máxima verossimilhança tradicional. Isso pode ser observado tanto na Figura 4, onde apresenta-se os intervalos de confiança baseados nas classificações de validação cruzada para AUC e as curvas ROC, quanto na Tabela 5 que contém os resumos da curva ROC.

Tabela 3: Resumos da curva ROC dos classificadores fundamentados em Modelos Lineares Generalizados

Medida	**GLM-MLE**	**GLM-Boost**
AUC*	0,969 (0,966, 0,972)	0,943 (0,938, 0,947)
Acurácia*	0,927 (0,912, 0,94)	0,865 (0,845, 0,882)
Sensibilidade	0,949	0,955
Especificidade	0,893	0,726
PPV	0,932	0,843
NPV	0,919	0,912
AUC	0,974	0,945

*Valor com intervalo de confiança de 95% baseado nas 30 amostras de validação cruzada apresentado entre parênteses.

A superioridade da abordagem convencional de estimação dos parâmetros do modelo linear generalizado é particular desta análise. Em investigações do fato, pode-se atribuir esse melhor desempenho ao bom comportamento do conjunto de dados, eles são linearmente separáveis, desfavorecendo assim o método Boosting.

Portanto, para representar a abordagem via modelos lineares generalizados o classificador `GLM-MLE` é mantido.

3.3 Classification Trees-Based

Considerando agora os métodos baseados em árvores de decisão, são apresentados os resultados referentes as seguintes abordagens:

- `Tree-BAG`: Bagging Classification Trees
- `Rand-Forest`: Random Forest

Das abordagens tratadas apenas para em Random Forest realizou-se o *tunning*. Este foi feito para o parâmetro `mtry` que representa o número de variáveis aleatórias escolhidas em cada divisão da amostra, o valor de melhor desempenho foi de 29 variáveis. Não foi realizada a "poda", ou *prune* em inglês, das árvores.

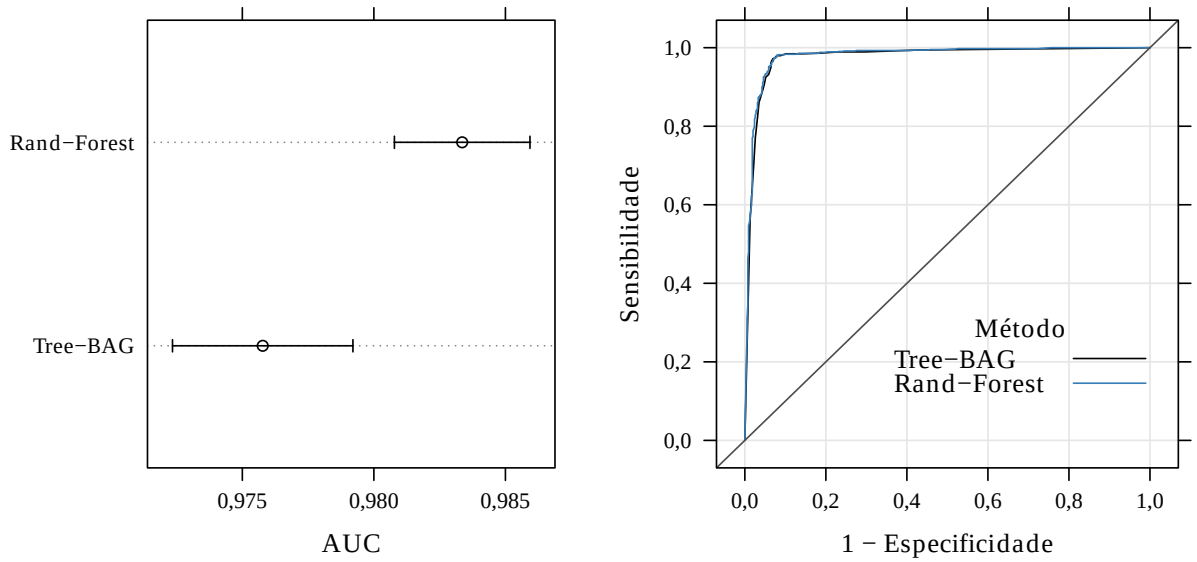

Figura 5: (Esquerda) Intervalos de confiança para a área abaixo da curva ROC baseados nas 3 repetições das 10 amostras de validação cruzada. (Direita) Curva ROC dos classificados aplicados à base de teste.

Para os resultados exibidos na Figura 5 nota-se uma dissimilaridade entre os intervalos de AUC baseados nas 30 classificações da validação cruzada (à esquerda) e a curva ROC de classificação da base de teste com o classificador ajustado com todos as observações da base de treinamento. O classificador `Rand-Forest` apresentou um melhor desempenho na validação cruzada, quando comparado com o `Tree-bag`, isso se reflete nos intervalos de confiança para AUC que não se sobrepõe. Porém quando utilizado toda a base de treinamento os resultados praticamente se equivalem, note a similaridade entre as curvas ROC à direita na Figura 5. Essa semelhança também é observada nos valores pontuais da curva ROC na Tabela **??**. Na Tabela os valores pontuais apresentam um ligeiro melhor desempenho para o classificador `Tree-BAG`, mas quando observado o desempenho na validação cruzada o classificador `Rand-Forest` se sobressai.

Assim, adotando o critério de melhor desempenho na validação cruzada mantemos o classificador `Rand-Forest` para comparação finais com os demais métodos.

Tabela 4: Resumos da curva ROC dos classificadores fundamentados em Árvores de decisão

Medida	**Tree-BAG**	**Random-Forest**
AUC*	0,976 (0,972, 0,979)	0,983 (0,981, 0,986)
Acurácia*	0,956 (0,944, 0,966)	0,951 (0,939, 0,962)
Sensibilidade	0,973	0,965
Especificidade	0,93	0,93
PPV	0,955	0,955
NPV	0,957	0,946
AUC	0,976	0,979

*Valor com intervalo de confiança de 95% baseado nas 30 amostras de validação cruzada apresentado entre parênteses.

3.4 Support Vetor Machine-Based

Finalmente no último grupo de métodos considerados no trabalho têm-se os resultados para os classificadores baseados em Support Vector Machines. Foram ajustados os classificadores considerando o kernel Linear e as expandindo as características das observações através dos kernels Polinomial e Gaussiano.

- `SVM-Linear`: Support Vector Classifier com kernel Linear;
- `SVM-Poly`: Support Vector Classifier com kernel Polinomial; e
- `SVM-Gauss`: Support Vector Classifier com kernel Gaussiano.

Para todos os cados realiza-se o *tunning* do parâmetro `C` que determina o custo de classificação incorreta. Quando considerada a expansão via kernel polinomial também se faz o tunning dos parâmetros `degree` (d) e `scale` (γ), conjuntamente com o `C`, os parâmetros foram fixados em `degree` = 2, `scale` = 0,01 e `C` = 1. Para a expansão via expansão Gaussiana o parâmetro `sigma` (σ) foi fixado em 0,0282 com `C` = 4. No kernel linear o `C` que proporcionou um melhor desempenho foi de 1,502.

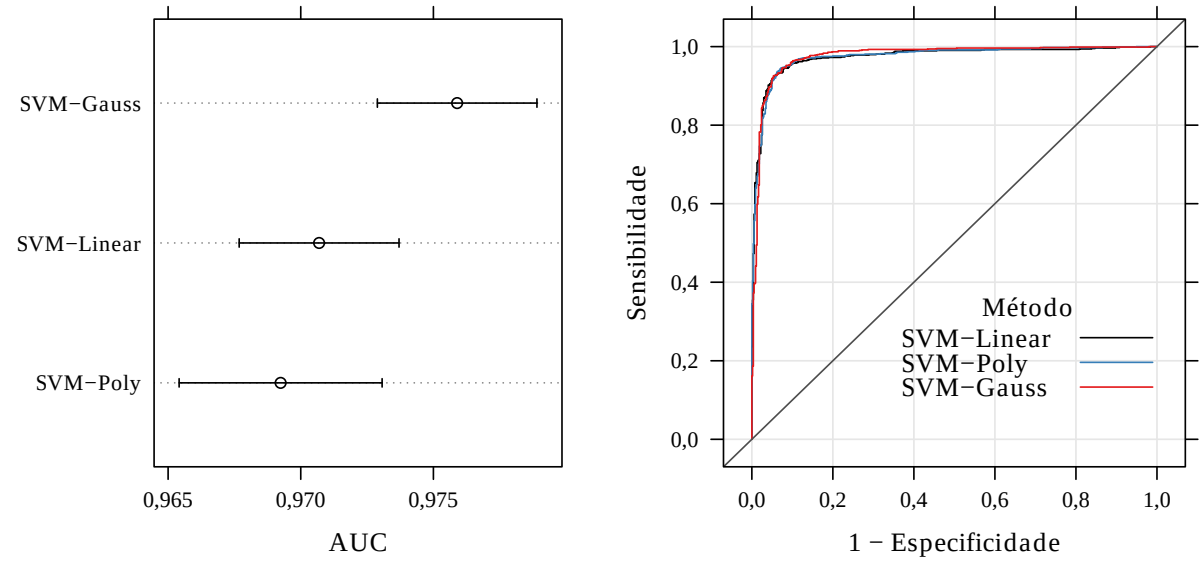

Figura 6: (Esquerda) Intervalos de confiança para a área abaixo da curva ROC baseados nas 3 repetições das 10 amostras de validação cruzada. (Direita) Curva ROC dos classificados aplicados à base de teste.

Nos resultados obtidos dos classificadores baseados em Support Vector Machines, temos algo similar ao apresentado na Seção 3.3. Os resultados baseados na validação cruzada, apresentados à esquerda na Figura 6, favorecem o classificador que utiliza a expansão de característica via kernel Gaussiano, embora os intervalos se estejam sobrepostos. Porém, nos resultados dos classificadores quando aplicados à base de teste são muito similares, à direita da Figura 6. Complementando a visualização gráfico os resultados pontuais são apresentados na Tabela **??**. Nota-se que os classificadores obtiveram resultados realmente muito parecidos, mesmo para os resultados na validação cruzada a diferença se dá somente com três casas decimais.

Tabela 5: Resumos da curva ROC dos classificadores fundamentados em Support Vector Machines

Medida	**SVM-Linear**	**SVM-Poly**	**SVM-Gauss**
AUC*	0,971 (0,968, 0,974)	0,969 (0,965, 0,973)	0,976 (0,973, 0,979)
Acurácia*	0,934 (0,92, 0,947)	0,935 (0,92, 0,947)	0,936 (0,921, 0,948)
Sensibilidade	0,958	0,963	0,956
Especificidade	0,897	0,892	0,904
PPV	0,935	0,932	0,939
NPV	0,933	0,94	0,93
AUC	0,973	0,974	0,976

*Valor com intervalo de confiança de 95% baseado nas 30 amostras de validação cruzada apresentado entre parênteses.

Mesmo que timidamente, nota-se que a expansão de características via kernel Gaussiano proporcionou melhores resultados. Assim manteve-se esse classificador no rol de classificadores elencados para comparação entre abordagens.

3.5 Comparação das abordagens

Nesta seção, os métodos que apresentaram melhor desempenho em cada abordagem são contrastados. Os métodos sob comparação são os denominados por `LDA`, `GLM-MLE`, `Rand-Forest` e `SVM-Gauss`. Na figura 7, à esquerda, são apresentadas as curvas ROC, provenientes da classificação da base de teste, para cada um dos classificadores. As curvas apresentam comportamentos razoavelmente similares, mas percebe-se que o classificador `LDA` apresentou um desempenho insatisfatório com relação dos demais. Isso também é observado nos resultados da validação cruzada, apresentados à direita da Figura 7. Cada ponto neste gráfico representa um valor de AUC de cada um dos classificadores na validação cruzada. Perceba os valores obtidos para `LDA` estão todos abaixo da linha pontilhada que representa a igualdade de valores, ressaltando seu mal desempenho em comparação com os demais. Nas outras dispersões nota-se um melhor desempenho para `Rand-Forest` e uma similaridade entre `GLM-MLE` e `SVM-Gauss`.

Na Figura 8 outra apresentamos outra forma de comparação dos classificadores. Nesta figura temos os intervalos de confiança para a `especificidade`, `sensibilidade` e `AUC`, à esquerda, e os valores de AUC para cada uma das amostra da validação cruzada de cada uma dos classificadores, à direita. Em ambos gráficos a mesma indicação observada na Figura 7 pode ser vista. Temos desempenhos melhores seguindo a ordem `Rand-Forest`, `SVM-Gauss`, `GLM-MLE` e por fim `LDA`.

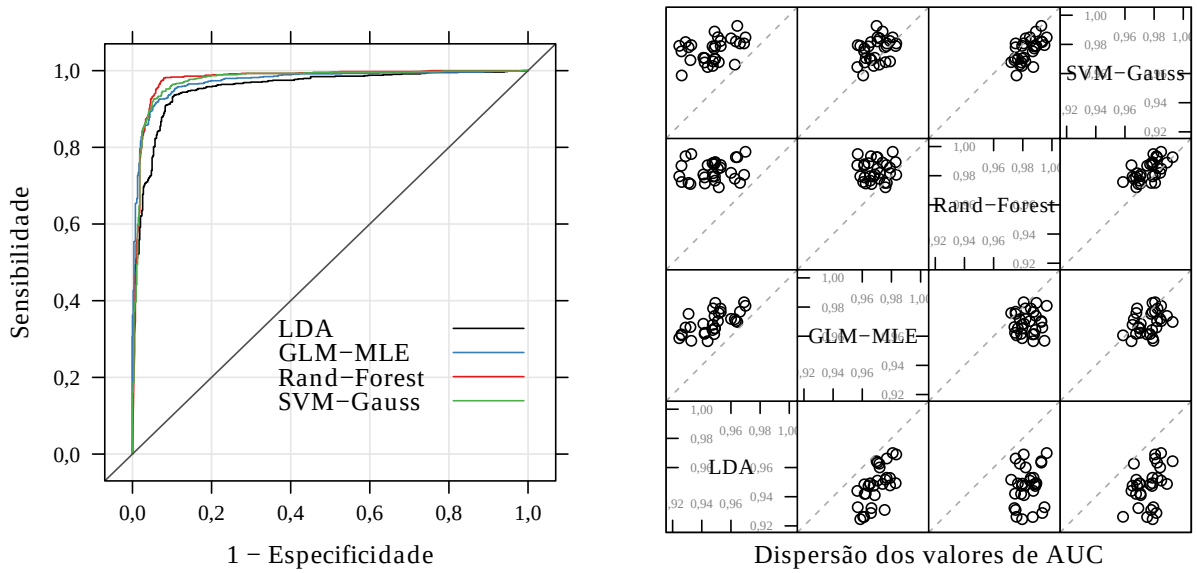

Figura 7: (Esquerda) Curva ROC dos classificadores aplicados à base de teste. (Direita) Gráficos de dispersão dos valores de AUC obtidos para cada uma das 30 amostras da validação cruzada.

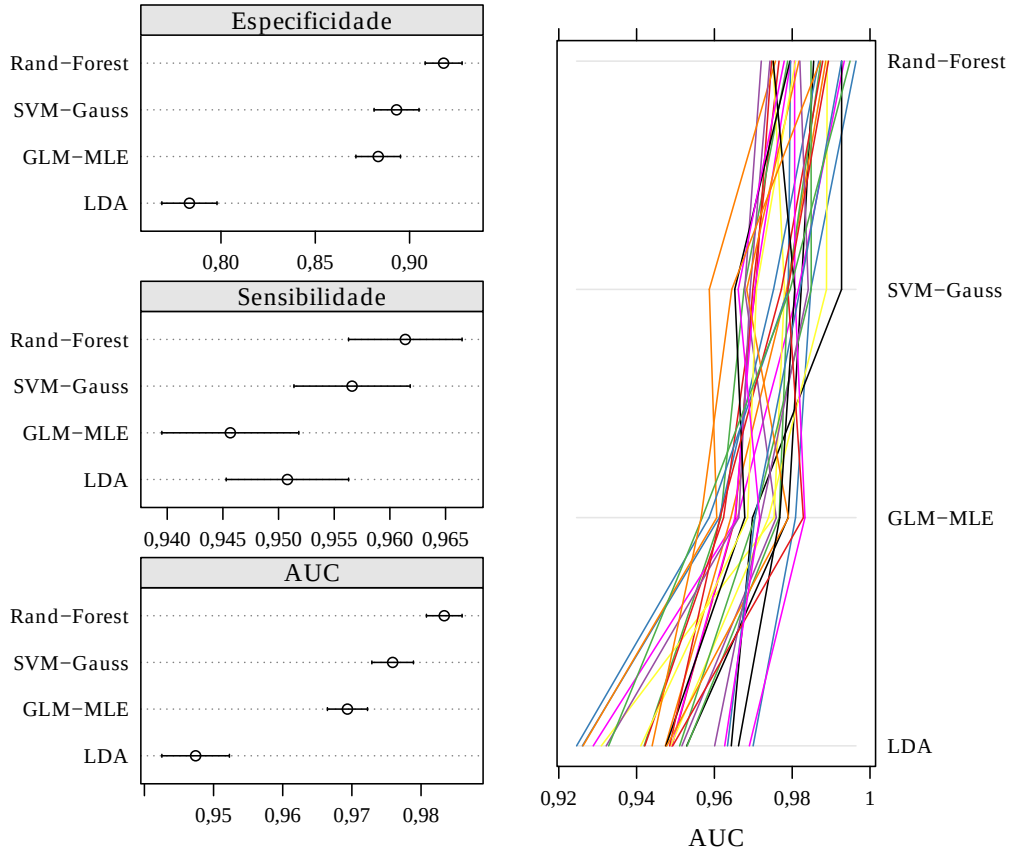

Figura 8: (Esquerda) Intervalos de confiança para especificidade, sensibilidade e área abaixo da curva ROC. (Direita) Valores de AUC. Ambos baseados nas 3 repetições das 10 amostras de validação cruzada

Tabela 6: Resumos da curva ROC dos classificadores com melhores desempenhos

Medida	**LDA**	**GLM-MLE**	**Rand-Forest**	**SVM-Gauss**
AUC*	0,947 (0,943, 0,952)	0,969 (0,966, 0,972)	0,983 (0,981, 0,986)	0,976 (0,973, 0,979)
Acurácia*	0,899 (0,882, 0,915)	0,927 (0,912, 0,94)	0,951 (0,939, 0,962)	0,936 (0,921, 0,948)
Sensibilidade	0,955	0,949	0,965	0,956
Especificidade	0,814	0,893	0,93	0,904
PPV	0,888	0,932	0,955	0,939
NPV	0,921	0,919	0,946	0,93
AUC	0,956	0,974	0,979	0,976

*Valor com intervalo de confiança de 95% baseado nas 30 amostras de validação cruzada apresentado entre parênteses.

Com os resultados apresentados anteriormente temos que o melhor desempenho para classificação se deu considerando a abordagem baseada em árvores de decisão, mas especificamente o classificador **Random Forest**. Para as abordagens via Support Vector Machines (considerando a expansão via kernel Gaussiano) e Modelos Lineares Generalizados (modelo logístico ajustado via máxima verossimilhança) observou-se resultados similares e satisfatórios. Já para os métodos baseados em Análise Discriminante não se obteve um desempenho em comparação com as demais técnicas.

Material Suplementar

Toda a análise foi realizada com o auxílio do software R e está disponível online no endereço <https://github.com/JrEduardo/ce064-ml/tree/master/finalWork>. Dúvidas, sugestões e críticas são sempre bem-vindas.

4 Conclusões

No desenvolvimento do trabalho foram apresentados onze técnicas para obtenção de classificadores, seccionadas em quatro grandes áreas com abordagens distintas de classificação (fundamentadas em análise discriminante, modelos lineares, generalizados, árvores de decisão e support vector machines). Na aplicações das técnicas de classificação observou-se resultados muito bons no que tange a predição, isso se deve ao fato do conjunto de dados em estudo apresentar covariáveis mensuradas que favoreceram a classificação.

Mesmo com todos os classificadores apresentando bons resultados de classificação pode-se compará-los através de resumos da curva ROC e dos resultados da validação cruzada e nessa comparação foram verificados melhores desempenhos dos classificadores baseados em árvores de decisão.

5 Referências

BREIMAN, L. Statistical Modeling: The Two Cultures. **Statistical Science**, v. 16, n. 3, p. 199–231, ago. 2001.

HOFNER, B. et al. Model-based boosting in R: A hands-on tutorial using the R package mboost. **Computational Statistics**, v. 29, p. 3–35, 2014.

JAMES, G. et al. **An introduction to statistical learning**. Traducao. [s.l.] Springer, 2013. v. 112

KUHN, M. Building Predictive Models in R Using the caret Package. **Journal Of Statistical Software**, v. 28, n. 5, p. 1–26, 2008.

LICHMAN, M. **UCI machine learning repository**University of California, Irvine, School of Information; Computer Sciences, 2013. Disponível em: <http://archive.ics.uci.edu/ml>