II Workshop de Matemática, Estatística e Computação Aplicadas à Indústria

Utilizando testes estatísticos para comparar a performance de modelos de aprendizado de máquina

Luiz Guilherme Giordani[1]
Afonso Paiva[2]
ICMC-USP

1 Introdução

O processo de criação de um modelo de aprendizado de máquina é um estudo estatístico por natureza. Com um modelo treinado no conjunto de treino, ao aplica-lo nas bases de dados de validação e teste, estamos estimando a sua performance em novos dados.

Porém, contrário a outras metodologias estatísticas, no processo de criação e avaliação de modelos de aprendizado de máquina, não existe o rigor de estudar métricas de variabilidade dos modelos, como a variância e o desvio padrão. Consequentemente, a seleção de um melhor modelo, tanto na indústria como na academia, acaba sendo por aquele que mostra o melhor desempenho em uma medida de avaliação sumarizada em todo o conjunto de dados.

Por outro lado, existem metodologias robustas para estimar a performance desses modelos em um contexto onde a variância dos modelos, aplicados aos dados, são levados em consideração. Essas metodologias buscam utilizar de conceitos bem fundamentados na área de aprendizado de máquina, como a validação cruzada e a reamostragem, para formular testes de hipóteses onde o tomador de decisão, de forma implícita, utilizará de mais medidas estatísticas para definir qual o melhor modelo.

Assim, o presente trabalho tem como objetivo revisitar essas metodologias, primeiro em um contexto de simulação, para revisitar a robustez de cada método, e após isso explorar casos de usos práticos em conjuntos de dados reais.

2 Metodologia

Entre os métodos estudados, se encontram dois testes de hipóteses e duas técnicas de reamostragem. Os testes são o teste T pareado 5×2cv [2] e o teste F 5×2cv [1]. As técnicas de reamostragem são o *bootstrap* [3] e o teste de permutação pareado [4].

[1]luiz.giordani@usp.br
[2]apneto@icmc.usp.br

2.1 Teste T pareado 5×2cv

O teste propõe realizar 5 repetições de validação cruzada com duas partições, e a estatística de teste é calculada da seguinte forma:

$$\widetilde{t} = \frac{p_1^{(1)}}{\sqrt{\frac{1}{5}\sum_{i=1}^{5} s_i^2}} \tag{1}$$

Onde:

- i é a repetição $i \in [1..5]$ da validação cruzada,
- $p_1^{(1)}$ é a diferença entre as métricas de performance de dois modelos, treinados na primeira partição da primeira repetição da validação cruzada,
- s_i^2 é a variância da métrica de performance na repetição i,
- $\widetilde{t}$ é a estatística de teste t computada

A estatística de teste $\widetilde{t}$ computada em (1) deve ser comparada com a distribuição t-Student com 5 graus de liberdade.

2.2 Teste F 5×2cv

Similar ao teste T pareado 5×2cv, porém ao invés de utilizar somente a estatística de avaliação da primeira partição, o teste F 5×2cv utiliza-se todas as 10 geradas durante o estudo da seguinte forma:

$$f = \frac{\sum_{i=1}^{5}\sum_{j=1}^{2}(p_i^{(j)})^2}{2\sum_{i=1}^{5} s_i^2} \tag{2}$$

Onde:

- j é a partição $j \in [1, 2]$ da validação cruzada,
- $p_i^{(j)}$ é a diferença entre as métricas de performance de dois modelos treinados na partição i e repetição j da validação cruzada,
- f é a estatística de teste f computada

Como a Equação (2) resulta em uma razão entre duas distribuições qui-quadrado com 10 e 5 graus de liberdade respectivamente, naturalmente temos então uma distribuição F com 10 e 5 graus de liberdade, de onde podemos comparar a estatística de teste f.

2.3 Bootstrap

É um método de reamostragem que consiste em selecionar *m* amostras com reposição de tamanho *n*. Onde *n* é o número de observações no seu conjunto de dados.

Um caso de uso do *bootstrap* é a aproximação de distribuições de probabilidade para estatísticas onde a sua forma é de difícil derivação matemática, como o caso da mediana. Nesse caso, em um conjunto de dados, aplicamos a técnica, calculando a mediana em cada subamostra. Feito isso, podemos inferir sobre a distribuição dessa estatística.

Similarmente, em um estudo de aprendizado de máquina, podemos querer comparar a diferença de performance entre dois modelos. Dessa forma, seria calculada a diferença entre as métricas de performance de cada modelo em cada subamostra. Do conjunto dessas diferenças, podemos criar inferências sobre a distribuição das diferenças entre os dois modelos.

2.4 Teste de permutação pareado

Um outro método de reamostragem que consiste em trocar, aleatoriamente, as predições de dois modelos aplicados à mesma população. O algoritmo segue da seguinte forma:

1. Calcular a diferença entre a métrica de avaliação de cada modelo avaliado p_i
2. Realizar a troca de um subconjunto aleatório observações
3. Calcular a métrica de avaliação de cada modelo após a troca
4. Repetir 2 e 3 k vezes (k geralmente é entre 1000 a 2000)
5. Calcular o valor d a partir da proporção de casos que $p_k > p_i$

2.5 Estudo de simulação

Aplicar cada uma das técnicas em conjuntos de dados simulados, ou em situações que aproximam casos de uso de aprendizado de máquina. Esses casos serão criados para que não existam diferenças entre qualquer dois modelos ou métricas calculadas sobre os mesmos. A robustez de cada técnica será medida pela taxa de erro do tipo I, ou seja, detectar uma diferença estatística quando ela não existe.

2.6 Estudo prático

Aplicar as diferentes técnicas em situações reais de modelagem como:

- Comparação entre duas técnicas de modelagem
- Seleção de *features*
- Comparar um novo modelo com um sistema atual

Para isso utilizaremos o conjunto de dados *Credit Card Fraud Detection* [5] disponível na plataforma *Kaggle*.

3 Conclusão e Trabalho Futuro

O trabalho pretende estreitar os laços entre metodologia estatística e aplicações práticas de aprendizado de máquina. Realizando os estudos de simulação, pretende-se mostrar os pontos fortes e fracos de cada técnica. Em um segundo momento, no estudo prático, deseja-se demonstrar circunstâncias onde as técnicas podem ser aplicadas tanto na indústria como no meio acadêmico.

Com os resultados e a apresentação do trabalho, esperamos contribuir de forma significativa para um maior rigor na avaliação de modelos e estudos na área de aprendizado de máquina.

Referências

[1] Alpaydin, E. *Combined 5×2 cv F test for comparing supervised classification learning algorithms*. Neural computation, 11(8), 1885-1892, 1999. DOI 10.1162/089976699300016007

[2] Dietterich TG, *Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms*. Neural Comput 10:1895–1923, 1998. DOI 10.1162/089976698300017197

[3] Efron, B. and Tibshirani, R. J. *An Introduction to the Bootstrap*. Boca Raton, FL: CRC Press, 1994. DOI 10.1201/9780429246593

[4] J. Menke and T. R. Martinez, *Using permutations instead of student's t distribution for p-values in paired-difference algorithm comparisons*, 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541), 2004, pp. 1331-1335 vol.2, DOI 10.1109/IJCNN.2004.1380138.

[5] Pozzolo, A.D., Caelen, O., Borgne, Y.L., Waterschoot, S., Bontempi, G. *Learned lessons in credit card fraud detection from a practitioner perspective*. Expert Syst. Appl., 41, 4915-4928, 2014. DOI 10.1016/j.eswa.2014.02.026