Nanodegree Engenheiro de Machine Learning

Proposta de projeto final

Marco Aurélio Moura Suriani
09 de abril de 2019

Previsão da necessidade de reforço escolar para alunos de matemática de acordo com características demogaficas, sociais e acadêmicas

Histórico do assunto

Como professor, carreira que exerço há sete anos, um dos meus grandes interesses é tentar prever o desempenho de alunos baseando-se em suas principais características. Neste sentido, destaco o estudo de (Cortez e Silva, 2008), que faz parte de um esforço do governo de Portugal em melhorar os índices educacionais do país. Tal estudo buscou prever o desempenho (aprovado/reprovado) de estudantes na faixa de 15 a 22 anos em duas escolas a partir de características demográficas (como sexo e região), sociais (como escolaridade dos pais) e acadêmicas (como tempo semanal dedicado ao estudo e suporte externo), de modo a determinar se o aluno precisará de alguma forma de intervenção para melhorar seu desempenho acadêmico.

Já o estudo de (Kotsiantis; Pierrakeas e Pintelas, 2004) aplicou técnicas de Mineração de Dados para prever o desempenho de alunos universitários com base tanto em aspectos demográficos (como sexo e estado civil) quanto em avaliações passadas. A previsão foi realizada usando técnicas como Naive Bayes (que se mostrou a mais eficiente), Regressão Logística e k-Vizinhos mais Próximos (k-NN). Usando tais metodologias, os autores foram capazes de prever se um aluno seria aprovado ou reprovado com uma acurácia de 74%. Os resultados previram alunos que teriam fraco desempenho escolar na modalidade de ensino superior à distância, permitindo que os tutores tomem precauções e estejam mais bem preparados para lidar com casos do tipo.

Descrição do problema

O problema proposto é utilizar os mesmo dados usados no trabalho de (Cortez e Silva, 2008), que estão disponíveis no Conjunto de Dados de Desempenho de Estudantes do Repositório de Aprendizado de Máquinas da Universidade da Califórnia em Irvine, para tentar prever quais alunos precisarão de aulas de reforço na disciplina de Matemática com a maior antecedência possível. Desta forma, serão usados apenas os atributos conhecidos no início do aluno letivo, ou seja, não se levará em consideração as notas parciais obtidas antes da nota final, e nem a quantidade de faltas. A diferença entre o presente trabalho e o original, é que aqui será dado foco à previsão de necessidade de reforço e não à previsão de desempenho final (aprovado/reprovado). A variável target será a necessidade de reforço, que é quantificável (o aluno precisa de reforço ou não) e mensurável (através de seu desempenho final).

Conjuntos de dados e entradas

O conjunto de dados do repositório contém 33 variáveis para alunos de Matemática (395 instâncias) e para alunos de Português (254 instâncias), sendo elas: **1- school** - escola (binária: 'GP' - Gabriel Pereira ou 'MS' - Mousinho da Silveira), **2- sex** - sexo do estudante (binária: 'F' - feminino ou 'M' - masculino), **3- age** - idade

(numérica: 15 a 22), **4- address** - tipo de residência (binária: 'U' - urbana ou 'R' - rural), **5- famsize** - tamanho da família (binária: 'LE3' - até 3 ou 'GT3' - mais que 3), **6- Pstatus** - cohabitação com os pais (binária: 'T' - moram juntos ou 'A' - separados), **7- Medu** - educação da mãe (numérica: 0 - nenhuma, 1 - até 4º ano, 2 - até 9º ano, 3 - ensino médio ou 4 - ensino superior), **8- Fedu** - eduação do pai (numérica: idem anterior), **9- Mjob** - emprego da mãe (nominal: 'teacher', 'health', 'services', 'at_home' ou 'other'), **10- Fjob** - emprego do pai (nominal: idem anterior), **11- reason** - motivo de escolha da escola (nominal: 'home' - próximo de casa, 'reputation' - reputação da escola, 'course' - preferência pelo curso, ou 'other'), **12- guardian** - guarda do aluno (nominal: 'mother', 'father' ou 'other'), **13- traveltime** - tempo de percurso até a escola (numérica: 1 - 1 hora), **14- studytime** - tempo de estudo semanal (numérica: 1 - 10 horas), **15- failures** - reprovações passadas (numérica: n se 1<=n<3, caso contrário 4), **16- schoolsup** - suporte extra (binária: 'yes' ou 'no'), **17- famsup** - suporte familiar (binária: 'yes' ou 'no'), **18- paid** - aulas particulares pagas (binária: 'yes' ou 'no'), **19- activities** - atividades extra-curriculares (binária: 'yes' ou 'no'), **20- nursery** - cursou a pré-escola (binária: 'yes' ou 'no'), **21- higher** - deseja cursar ensino superior (binária: 'yes' ou 'no'), **22- internet** - acesso à Internet (binária: 'yes' ou 'no'), **23- romantic** - em relacionamento amoroso (binária: 'yes' ou 'no'), **24- famrel** - qualidade das relações familiares (numérica: de 1 - muito ruim a 5 - excelente), **25- freetime** - tempo livre após a escola (numérica: de 1 - muito baixo a 5 - muito alto), **26- goout** - sai com amigos (numérica: de 1 - muito baixo a 5 - muito alto), **27- Dalc** - consumo de álcool durante a semana (numérica: de 1 - muito baixo a 5 - muito alto), **28- Walc** - consumo de álcool durante o fim-de-semana (numérica: de 1 - muito baixo a 5 - muito alto), **29- health** - estado de saúde (numérica: de 1 - muito baixo a 5 - muito alto), **30- absences** - altas à escola (numérica: de 0 a 93), **31- G1** - nota no primeiro período (numérica: de 0 a 20), **32- G2** - nota no segundo período (numérica: de 0 a 20), **33- G3** - nota final (numérica: de 0 a 20, target).

Neste trabalho, serão usados apenas os dados dos alunos de Matemática e suas 395 instâncias. Os atributos **30- absences**, **31- G1**, e **32- G2** não serão usados pois, como se deseja uma previsão com a maior antecedência possível, não adianta usar atributos cujos valores só são conhecidos após decorrido certo tempo do início das aulas (de nada adianta uma boa previsão de necessidade de reforço, se é necessário esperar por 2/3 do ano letivo para obtê-la). Assim, a variável **33- G3** será transformada no target (necessidade de reforço) e as variáveis **1 a 29** serão os atributos. Ressalta-se que alguas destas variáveis demandarão pré-tratamento antes de serem usadas no treino/teste dos modelos.

Descrição da solução

A solução para este problema será uma previsão da necessidade de cada aluno de receber aulas de reforço a partir de um modelo de Aprendizado de Máquinas capaz de usar os 29 atributos descritos na seção anterior. Um aluno que precisa de reforço é aquele do qual se espera que não seja capaz de atingir nota mínima para aprovação no fim do ano letivo. Desta forma, a solução pode ser mensurada através da quantidade de previsões de necessidade de reforço. Por fim, a solução pode ser replicada quantas ezes for necessário a partir do modelo que será desenvolvido.

Modelo de referência (benchmark)

O trabalho de em (Cortez e Silva, 2008), que é a fontedos dados usados neste trabalho, determinou três configurações de entrada para os modelos de previsão: (A) aquele que levava em consideração as notas parciais G1 e G2 para prever se o aluno obteria uma G3 abaixo de 10 (reprovado) ou a partir de 10 (aprovado), (B) aquele que levava em consideração apenas G1 e (C) aquele que levava em consderação apenas os demais atributos (sem considerar as notas parciais). *O presente trabalho usará os resultados da configuração (C) como base para comparações.*

Para cada configuração de entrada, a previsão foi realizada através das seguintes técnicas de Mineração de Dados: Árvores de Decisão, Florestas Aleatórias, Redes Neurais e Máquinas de Vetores de Suporte. A avaliação dos modelos se deu por sua acurácia (#previsões corretas/#previsões). Os resultados mostram que as acurácias obtidas dependem menos do tipo de modelo do que da configuração de entrada, uma vez que a

configuração (A), que leva em consideração as notas dos alunos ao longo do ano, sempre obteve melhor acurácia do que a configuração (C), que não leva em conta o desempenho passado. Mais especificamente, observaram-se na configuração (C) acurácias entre 65,3% (Árvore de Decisão) e 70,6% (Máquinas de Vetores de Suporte). Vale ressaltar que na configuração (C), um preditor ingênuo que classifica todos os estudantes como aprovados obtém acurácia de 67,1%. Além disso, nas configurações (A) e (B), nenhum modelo superou o previsor ingênuo.

Métricas de avaliação

Primeiramente, conforme observado a seção anterior, o trabalho de (Cortez e Silva, 2008) leva em consideração apenas a acurácia para avaliação dos modelos, mas para o propósito deste trabalho, apenas tal métrica não será o bastante. Também deverão ser levadas em conta métricas como a precisão e a revocação (recall) para evitar que alunos sem reforço acabem reprovados e que alunos que acabariam sendo aprovados sejam submetidos desnecessariamente a aulas de reforço. A precisão é igual à razão (#número de previsões corretas de alunos que precisam de recuperação/#número de previsões de alunos que precisam de recuperação), enquanto que a revocação é igual à razão (#número de previsões corretas de alunos que precisam de recuperação/#número de alunos que precisam de recuperação). Por fim, também será usado o F1 score, que é a média harmônica entre precisão e revocação.

Além disto, o desempenho dos modelos será comparado ao desempenho de um preditor ingênuo, que classifica todos os estudantes como pertencentes à categoria com maior frequência. Desta forma, pode-se avaliar se os modelos estão de fato extraindo conhecimento dos dados.

Design do projeto

Etapas do projeto:

1) Pré-processamento dos atributos para que possam ser utilizados nos modelos propostos. Alguns atributos são binários, devendo ser codificados em 0/1, enquanto outros são nominais, devendo ser codificados usando uma técnica como one-hot.
2) Os atributos pre-processados, bem como o target, alimentarão algoritmos de classificação e de regressão de Aprendizado de Máquinas para prever a necessidade de reforço escolar aos estudantes. Entre os algoritmos que se planeja usar, estão a Regressão Logística, o Naive Bayes, as Árvores de Decisão, as Máquinas de Vetores de Suporte e os k-Vizinhos mais Próximos (k-NN), bem como a Regressão Linear. Os modelos serão treinados usando 66% dos dados e testados usando os 33% restantes (validação cruzada).
3) Os resultados serão comparados entre si através das métricas descritas (acurácia, precisão e revocação), e ttambém com um "previsor ingênuo", conforme descirtos.
4) O melhor modelo será usado para apontar a solução para o problema.

Referências Bibliográfias

01 Cortez, P. e Silva, A. **Using Data Mining to Predict Secondary School Student Performance.** In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7. [WEB link] (http://www3.dsi.uminho.pt/pcortez/student.pdf)

02 Kotsiantis, S.; Pierrakeas, C. e Pintelas, P. **Predicting Students' Performance in Distance Learning Using Machine Learning Techniques.** Applied Artificial Intelligence (AAI), 18, no. 5, 2004, 411–426. [WEB link] (https://pdfs.semanticscholar.org/ca32/7f56f4290809d7c243b57c97bb2eb0916ff1.pdf)

03 UC Irvine Machine Learning Repository **Student Performance Data Set.** [WEB link] (https://archive.ics.uci.edu/ml/datasets/Student+Performance)