Orientações sobre o Projeto Final

Prof. Danilo Silva
UFSC

EEL7513/EEL7514/EEL410250 – Aprendizado de Máquina

1 Introdução

O projeto final tem como objetivo aplicar os conceitos da disciplina, de forma livre e ampla, em algum problema prático de interesse do aluno. Este documento fornece orientações gerais sobre a realização do projeto e descreve as expectativas sobre seus entregáveis. Recomenda-se uma leitura completa e com atenção.

2 Preparação

2.1 Definição da equipe

O projeto pode ser realizado individualmente ou em dupla. Naturalmente, para um projeto em dupla, será esperado um desenvolvimento mais aprofundado do que para um projeto individual, o que terá impacto na avaliação.

Há muitas vantagens em realizar o projeto em dupla, no entanto, é importante certificar-se de que ambos os membros possuem genuíno interesse no tema e suficiente entrosamento e confiança para trabalhar em equipe, de forma a evitar potenciais surpresas e frustrações.

Espera-se que a dedicação ao projeto seja a mesma para ambos os membros da equipe; caso contrário, ao final do projeto, espera-se honestidade da equipe em informar o percentual aproximado de dedicação de cada membro, o que terá impacto na avaliação.

2.2 Definição do tema

O tema do projeto deve estar dentro do escopo da disciplina e envolver a aplicação prática de métodos de aprendizado de máquina em algum conjunto de dados público. Dado que o foco principal da disciplina é o aprendizado supervisionado, a grande maioria dos projetos aborda esse tipo de aprendizado como tarefa final, com algumas exceções envolvendo modelos generativos (ex: GANs).

Quaisquer técnicas abordadas na disciplina podem ser utilizadas, bem como técnicas não abordadas, sejam de aprendizado de máquina convencional ("clássico") ou de aprendizado profundo. É interessante que o trabalho apresente uma comparação entre múltiplas possíveis técnicas ou, caso haja uma única técnica dominante para o problema em questão, que sejam feitas comparações entre múltiplas variações da técnica (ex: diferentes métodos de pré-processamento, extração de atributos, regularização, modelos pré-treinados, etc, dependendo do que for aplicável). Também é possível se aprofundar em alguma técnica não vista na disciplina e adotar um enfoque parcialmente teórico, mas com ainda alguma aplicação prática (ao menos de forma a demonstrar a técnica).

É muito importante que o tema seja de real interesse da equipe, uma vez que a motivação é fator crucial para o sucesso do projeto.

2.3 Definição do conjunto de dados

Não há restrição quanto ao tipo dos dados, sejam dados tabulares, séries temporais, sinais, imagens, etc. No entanto, é um requisito inegociável que o projeto seja desenvolvido majoritariamente

baseando-se em dados disponíveis publicamente[1] na Internet. Opcionalmente, uma vez desenvolvido um modelo de aprendizado, é permitido aplicá-lo a dados privados (obtidos ou produzidos pela própria equipe), como forma de demonstrar o funcionamento do modelo ou avaliar seu desempenho em uma aplicação mais próxima da realidade, mas isso não é necessário para o projeto.

Outro requisito é que o conjunto de dados, além de público, já tenha sido previamente utilizado em trabalhos relacionados, preferencialmente publicados em conferências ou periódicos científicos. Por exemplo, um conjunto de dados disponível no Kaggle mas que só foi utilizado publicamente em notebooks do próprio Kaggle ou em postagens de blogs **não** é aceitável.

Tais requisitos visam proporcionar independência à equipe no desenvolvimento do projeto: por um lado, elimina-se a hipótese de que um eventual desempenho insatisfatório do modelo seja devido a falhas ou limitações no conjunto de dados; por outro lado, a equipe terá à disposição trabalhos confiáveis como referência para se inspirar e aprender, sem precisar partir do zero na investigação do problema. Dessa forma, uma primeira meta típica é conseguir reproduzir aproximadamente os resultados de algum trabalho da literatura. (Às vezes, dependendo do problema, só essa meta já é bastante desafiadora e será o objetivo final do trabalho; ou pode ser tão simples quanto rodar um código pronto. Em todo caso, tem-se uma direção clara por onde começar.)

É importante avaliar criticamente se o conjunto de dados proporciona as informações necessárias para a resolução da tarefa preditiva que se deseja resolver (caso seja uma tarefa preditiva). Frequentemente isso envolve uma exploração inicial dos dados. Uma alternativa mais rápida pode ser analisar artigos relacionados que abordam este mesmo conjunto de dados, os quais já fornecerão uma descrição dos dados e da tarefa de interesse. Caso os dados não sejam adequados, deve-se buscar outro conjunto, ou procurar outro problema; portanto, a pesquisa pelo problema a ser resolvido, dados adequados e trabalhos relacionados deve ocorrer de forma integrada.

Caso múltiplos conjuntos de dados possivelmente interessantes sejam encontrados, num primeiro momento todos podem ser considerados e descritos na proposta de projeto, para posterior seleção sob orientação do professor.

2.4 Revisão da literatura

É essencial realizar uma pesquisa prévia por trabalhos na literatura que resolvem o mesmo problema ou problemas semelhantes; em particular, tentar determinar qual é o estado da arte. Caso isso seja difícil num primeiro momento devido a um número muito grande de artigos relacionados, deve-se ao menos determinar um subconjunto de artigos mais importantes que serão estudados num momento posterior.

Para pesquisar artigos, uma sugestão é utilizar o Google Acadêmico (scholar.google.com) e buscar por palavras-chave em inglês relacionadas ao tema, se necessário experimentando diversas variações. Por exemplo, em um trabalho sobre reconhecimento de peças de xadrez em imagens, pode-se buscar por: chess piece recognition, chess piece detection cnn, chessboard image dataset, etc. Outra opção é utilizar como palavra-chave o nome do conjunto de dados escolhido, caso este possua um nome consolidado (normalmente cunhado por quem disponibilizou o conjunto). Por exemplo, a busca por "CIFAR-100" encontrará todos os artigos que citam este conjunto de dados.

Uma vez encontrado ao menos um artigo verdadeiramente relevante (que trata do mesmo problema usando o mesmo conjunto de dados), este pode servir de ponto de partida para encontrar outros, os quais normalmente serão citados na introdução ou numa seção de trabalhos relacionados. Dessa forma, encontram-se trabalhos anteriores ao artigo em questão. Outra abordagem é buscar por trabalhos posteriores, que citam o artigo em questão. Isso pode ser feito, por exemplo, encontrando o artigo no Google Acadêmico e clicando em "Citado por ...". Caso necessário, pode-se em seguida refinar a busca selecionando a opção "Pesquisar nos artigos de citação".

Em muitos casos (mas nem sempre), é possível que os próprios autores disponibilizem publicamente o código utilizado para produzir os resultados do artigo, ou que outras pessoas tenham feito sua própria implementação e disponibilizado o código. Isso pode ser muito interessante como ponto de

[1] Não significa que os dados precisem estar disponíveis diretamente para *download*; podem ser dados que exijam o preenchimento de um formulário de autorização ou inscrição em uma competição.

partida para um projeto (ou não—às vezes o código é tão complicado e difícil de modificar que mais atrapalha do que ajuda). Frequentemente tais implementações são disponibilizadas em um repositório no GitHub (`github.com`) ou podem ser encontradas através do site `paperswithcode.com`.

Parte da definição do problema envolve a definição da métrica de avaliação do modelo. Até mesmo a determinação de qual artigo representa o estado da arte pode depender da métrica considerada. Às vezes, a métrica mais consolidada na literatura para um determinado problema pode ser uma métrica convencional e de fácil compreensão, como acurácia ou RMSE, mas às vezes pode ser bem mais complexa (como mAP na detecção de objetos). Ao pesquisar por trabalhos relacionados, é útil estar atento a quais métricas são as mais comuns para avaliar e comparar modelos no problema em questão.

2.5 Recursos computacionais

Um último aspecto a ser considerado na definição do tema do projeto são os recursos computacionais necessários para seu desenvolvimento, especialmente no caso de modelos de aprendizado profundo, que para treinamento em tempo razoável tipicamente exigem uma boa GPU. Ao pesquisar por artigos de referência, é útil estar atento a quais recursos computacionais foram utilizados, em particular se foi usada uma única GPU, e ao tempo gasto com treinamento, caso estas informações estejam disponíveis. Se as plataformas Google Colab e Kaggle Code (as quais fornecem GPUs gratuitamente com algumas restrições) se mostrarem insuficientes, a equipe também pode contratar serviços pagos (por exemplo, o Google Colab Pro), caso esteja disposta a investir—ou optar por redimensionar o escopo do projeto.

3 Proposta de projeto

A proposta deve conter no mínimo as seguintes informações (não necessariamente nessa ordem ou seguindo essa estrutura):

- Descrição da tarefa. O que exatamente você está tentando prever (caso seja uma tarefa preditiva) e a partir de quais dados? Trata-se de regressão, classificação binária, classificação multi-classe, classificação multi-rótulo, ou ainda alguma tarefa mais complexa (como detecção de objetos)?
- Motivação. Por que é interessante resolver essa tarefa de aprendizado? Já existe ou você vislumbra alguma aplicação? Há justificativa ou intuição para acreditar que a tarefa é viável de ser resolvida,[2] i.e., que a variável de saída de fato pode ser predita a partir dos dados de entrada?
- Descrição dos dados. Qual a faixa de possíveis valores da(s) variável(is) de saída? Quantas e quais são as variáveis de entrada e suas características? É possível mostrar exemplos?
- Métricas. Qual ou quais métricas serão usadas para avaliar o desempenho dos modelos?
- Trabalhos relacionados. Que outros trabalhos existem na literatura que resolvem o mesmo problema ou problemas semelhantes, em particular usando o mesmo conjunto de dados? Já existem implementações disponíveis na Internet?
- Objetivos. Quais os objetivos do projeto? Ou, mais precisamente, qual esforço a equipe se propõe a realizar que satisfaria as expectativas para um projeto final e justificaria sua nota? Em particular, o projeto precisa se diferenciar de simplesmente baixar um código pronto e rodar num conjunto de dados padrão. Possibilidades incluem: investigar diferentes modelos e/ou variações de arquitetura e/ou novas escolhas de hiperparâmetros e/ou abordagens de pré-processamento e/ou treinamento; aplicar a um conjunto de dados diferente ou em condições diferentes do que já foi feito na literatura; reproduzir um artigo para o qual não há implementação pronta; estudar e demonstrar conceitos ou técnicas não abordados diretamente na disciplina; etc.
- Metodologia. Caso não já esteja contemplado no item anterior, descreva brevemente qual metodologia você pretende seguir, em particular, quais abordagens você pretende investigar e

[2] Por exemplo, predição de preço de ações em bolsa de valores a parte de sua série histórica não é um tema aceitável.

comparar.

- Transparência. O projeto será usado como parte de um projeto de outra disciplina, estágio ou trabalho de conclusão de graduação ou pós-graduação? Nesse caso, indique a disciplina/orientador/empresa, etc. Mais alguém está envolvido no projeto que não faz parte da equipe da disciplina? Nesse caso, delimite as partes do projeto pelas quais a equipe ficará ou não responsável.
- Referências. As referências bibliográficas mais importantes que você conseguiu pesquisar até o momento, em particular as referências dos trabalhos relacionados citados previamente.

Uma proposta bem desenvolvida já funciona como uma versão preliminar do relatório final, permitindo aproveitar ao menos parte do esforço. Por outro lado, é comum que surjam dúvidas a serem sanadas com o professor e a proposta seja então aperfeiçoada e ressubmetida.

4 Relatório final

O relatório final deve conter obrigatoriamente as seguintes seções (ou variações):

1. Introdução: deve explicar, no mínimo, o contexto e motivação para o projeto, e seus objetivos.
    - Trabalhos relacionados: descrever brevemente o que já foi feito sobre o tema, indicando as principais abordagens utilizadas, com foco nos trabalhos que mais se aproximam dos objetivos do projeto. Pode ser uma subseção da introdução ou uma seção separada.
2. Metodologia (ou "Material e Métodos" ou "Métodos"): deve descrever o conjunto de dados (tipicamente como uma subseção), os modelos e abordagens utilizadas, incluindo eventual pré- ou pós-processamento dos dados, bem como as métricas de avaliação.
3. Resultados (ou "Experimentos" ou "Experimentos e Resultados" ou "Resultados e Discussão"): deve apresentar os resultados obtidos e discuti-los. Também pode ser organizada na forma de experimentos, caso os experimentos possuam variações de metodologia. A discussão também pode ser apresentada em uma seção separada.
4. Conclusão

No contexto desta disciplina de aprendizado de máquina, espera-se ver (entre outras coisas) os seguintes itens:

- Faixas de valores de hiperparâmetros avaliados, ou uma descrição de hiperparâmetros que foram experimentados mas não resultaram em melhoria de desempenho, bem como a especificação dos hiperparâmetros dos modelos finais.
- Curvas de treinamento e validação ao longo das épocas, tabelas comparando desempenho de treinamento e de validação, ou qualquer evidência de que o modelo foi apropriadamente regularizado para evitar overfitting.
- Análise de erros e/ou de importância de atributos, sempre que for viável.

5 Código utilizado

Todo o código necessário para reproduzir os resultados deve ser organizado em um arquivo ZIP e submetido via Moodle. Capriche na organização e garanta que os principais resultados estão salvos no notebook. Não é necessário incluir o conjunto de dados (pelo contrário, é preferível não incluir), sendo suficiente informar o link de acesso.

6 Apresentação

Prepare uma apresentação de cerca de 20 minutos explicando o problema, metodologia e resultados. Não é necessário revisar conceitos básicos, mas é importante demonstrar domínio do conteúdo e fundamentação por trás das decisões tomadas. As apresentações ocorrerão em slots de 60 minutos, de

forma que haverá tempo para perguntas do professor e, quando necessário, para que a equipe possa mostrar partes do código utilizado. É esperado que cada membro da equipe participe igualmente da apresentação e das respostas às perguntas, demonstrando claramente a sua contribuição ao projeto.

7 Dúvidas?

Qualquer dúvida, certifique-se de pedir esclarecimento ao professor com antecedência.