E-fólio B, notas lançadas

◀ Notas dos Efolios não disponíveis no Cartão de Aprendizagem

Sessão síncrona para preparação para prova final, 7 de junho, 21h ▶

Lista de respostas a começar pela mais antiga

Configurações

Este tópico de discussão está bloqueado. Já não pode responder ao mesmo.

E-fólio B, notas lançadas
por José Coelho - Segunda, 5 Junho 2023, 10:31

Caros estudantes,

Estão lançadas as notas do e-fólio B. Houve 34 trabalhos entregues, podem ver no recurso do e-fólio B a nota e respetivo feedback individual. Existe a seguinte distribuição dos estudantes pelas notas:

- [3.5; 4] valores: 9 estudantes
- [2.5; 3.5[valores: 15 estudantes
- [1.5; 2.5[valores: 8 estudantes
- [0.5; 1.5[valores: 1 estudante
- [0; 0.5[valores: 1 estudante

Em termos de notas dos e-fólios A e B, existe a seguinte distribuição dos estudantes pelas notas:

- [7.5; 8] valores: 0 estudantes
- [6.5; 7.5[valores: 8 estudantes
- [5.5; 6.5[valores: 7 estudantes
- [4.5; 5.5[valores: 11 estudantes
- [3.5; 4.5[valores: 7 estudantes
- [0; 3.5[valores: 7 estudantes

Dos 34 trabalhos entregues no e-fólio B, 33 obtiveram aprovação na parte contínua, e podem ir ao e-fólio global. Os restantes estudantes em avaliação contínua, têm de aguardar pelo exame na época de recurso.

Houve excelentes trabalhos entregues, o que para um e-fólio aberto considero bem sucedido. No entanto vou procurar orientar no próximo ano, fornecendo talvez um exemplo resolvido passo a passo, e uma template de relatório.

Os critérios foram centrados nos três métodos e no relatório, tendo sido utilizadas as seguintes linhas orientadoras na atribuição das classificações:

Relatório:

- 0,25 Caracterização dos dados utilizados
  - variável dependente e independentes bem descritas, com as unidades de cada variável
  - número de observações
  - número de países
  - número de anos
  - natureza das variáveis (contínuas, discretas, categorias), bem como a distribuição dos valores
  - Scatterplot de cada variável independente sobre a variável dependente, para ter uma noção inicial se existie ou não algum tipo de relação (ou calculo da correlação das variáveis independentes com a variável dependente)
- 0,25 Caracterização dos testes efetuados
  - Treino
  - Validação (se existir otimização de um hiperparâmetro)
  - Teste
- 0,25 Qualidade dos dados
  - Utilização de anos e países seguidos, portanto, observações independentes

  - Categorias criadas em cada variável, para realizar classificações em vez de regressões (há poucas observações para se fazerem regressões de qualidade).

- 0,25 apresentação / organização do relatório
  - Realizadas figuras explicativas, ou tabelas com informação agregada, ou simplesmente foi realizado capturas de ecrãs
  - índice e organização do relatório (no caso de ser extenso)
  - introdução e conclusões, de modo a tornar o documento independente do enunciado

- 0,25 conclusões suportadas nos resultados obtidos

Árvores de decisão

- 0,25 apresentação de uma árvore de decisão
- 0,25 normalizações nos dados, realizadas para este método
- 0,25 teste de hiperparametros (número de árvores ou outro), sem especificar o conjunto de validação.
- 0,25 apresentação de um indicador de performance

K vizinhos mais próximos

- 0,25 teste de hiperparametros (valor de k, ou outros), sem especificar o conjunto de validação.
- 0,25 normalizações e distâncias utilizadas bem especificadas
- 0,25 apresentação de um indicador de performance

Redes neuronais

- 0,25 teste de hiperparametros (número de neurónios e outros), sem especificar o conjunto de validação.
- 0,25 normalizações nos dados, realizadas para este método
- 0,25 apresentação de um indicador de performance

Taxa de erro elevada no conjunto de treino, para qualquer método, pode evidenciar que há algum erro a ser realizado. Como justificar valores mal classificados, se foram utilizados para fazer a árvore de decisão? Claro que podem existir erros, devido à poda na árvore de decisão, nos k vizinhos mais próximos existirem azares de vizinhos perto uns dos outros com valores distintos (significa que as variáveis escolhidas, ou a medida de distância não são adequadas), ou mesmo a rede neuronal ter parado o treino cedo, e alguns dados não terem o valor final convergido. Mas normalmente a taxa de erro deverá poder ser tão baixa quando se queira, o problema é que se for demasiado baixa, depois a taxa de erro no conjunto de teste poderá ser superior, devido ao overfitting.

Em vez de regressão, teria sido preferivel com poucas observações passarem para uma classificação. Todas as variáveis deveriam ser passadas para 3 a 5 categorias. Para ter uma regressão em condições, ou é algo muito óbvio, ou é necessário muitas observações. Neste contexto temos uma observação por país e ano, o que nos coloca neste e-fólio B numa situação de poucas observações. A categorização apenas da variável independente, é outra possibilidade.

Poucos a utilizarem os termos do manual para as funções de perda, L0/1, L1 e L2, preferindo os nomes em inglês mais utilizados vulgarmente.

Houve quem tenha escolhido como variáveis independentes, indicadores de outros objetivos. Assim, caso esses indicadores consigam explicar o indicador alvo, significaria que o objetivo alvo é irrelevante. Bastará satisfazer os restantes objetivos e esse ficará resolvido. As variáveis independentes deveriam ser retiradas principalmente de aspetos da sociedade que possam ser alterados por ações do governo, ou por mudança de hábitos da sociedade.

Houve quem tenha utilizado o factor() sobre variáveis contínuas. Tem de ser definidas categorias, 3 a 5, e só depois utilizar o factor(). O resultado de utilizar o factor() sobre uma variável contínua, é indicar que todos os valores são distintos, e portanto temos uma categoria para cada observação. Assim, não dá para aprender nada, já que as observações não se relacionam.

Qualquer situação já sabem que podem perguntar.

Cumprimentos,
José Coelho

Hiperligação permanente

◄ Notas dos Efolios não disponíveis no Cartão de Aprendizagem

Sessão síncrona para preparação para prova final, 7 de junho, 21h ►