UNIVERSIDADE DE SÃO PAULO - ICMC
Departamento de Ciências de Computação
SCC-0275 – Introdução à ciência de dados – Graduação - 2° Sem /2018
PROFa: Roseli Aparecida Francelin Romero
Estagiário PAE: Daniel Moreira Cestari

Aula 08 - Exploração de dados - Parte VI

Escolha apenas UM dataset de CLASSIFACAÇÃO para realizar os exercícios abaixo. Caso você não tenha utilizado um dataset desse tipo anteriormente escolha um novo para essa prática.

Se o seu dataset tem mais de duas classes transforme ele num problema binário.

Após a análise dos dados e um pré-processamento vem a etapa de modelagem dos experimentos. Essa etapa pode requerer voltar no pré-processamento caso perceba-se que algo precisa ser feito. A modelagem visa determinar as etapas da execução dos experimentos. No nosso cenário, experimento é a utilização de algoritmos de classificação, regressão ou agrupamento. Para tanto, é preciso definir, com ajuda da análise dos dados, o tipo do problema (classificação, regressão, ...), os atributos/features a serem utilizados e o processo de avaliação.

Essa prática foca mais no processo de avaliação. Para a avaliação é preciso definir qual a função de custo/erro adequada, e qual o estimador para o desempenho.

Utilizaremos medidas de desempenho para classificação binária baseadas na matriz de confusão (TFP, TFN, TVP, TVN).

Nas aplicações reais, o cliente dita qual a medida de desempenho deve ser utilizada, e muitas vezes não é uma das clássicas. E como essa medida, em geral, tem um impacto grande no treinamento do algoritmo de classificação, muitas vezes o algoritmo precisa ser adaptado e isso não é uma tarefa fácil.

Após a definição do tipo do problema e da medida de avaliação, é preciso definir como será estimado o desempenho final.

Esse processo está ligado á escolha do algoritmo de classificação bem como a escolha de alguns hiperparâmetros. Uma abordagem muito comum na área é a utilização do 10-fold Cross-Validation. Esse procedimento pode ser utilizado para estimar o desempenho do classificador final, bem como, na escolha de alguns poucos hiperparâmetros.

UNIVERSIDADE DE SÃO PAULO - ICMC
Departamento de Ciências de Computação
SCC-0275 – Introdução à ciência de dados – Graduação - 2° Sem /2018
PROFa: Roseli Aparecida Francelin Romero
Estagiário PAE: Daniel Moreira Cestari

1- Dada a introdução acima, já definimos que o tipo do problema é classificação. Defina quais os atributos você utilizará, e a medida de avaliação você acha adequada e explique o porquê dessas escolhas.

Lembre-se que o objetivo da classificação é fazer predições para dados não visto, ou seja, quando o algoritmo for colocado em produção ele classificará corretamente amostras não vistas.

2- Uma boa prática é escolher modelos mais simples, dados dois modelos com desempenho similar a escolha do mais simples é indicada pois com isso há algumas garantias de melhor generalização. Generalização, de maneira geral, é a propriedade que garante que o classificador terá desempenho parecido ao reportado no teste.

A definição da complexidade de um modelo nem sempre é uma tarefa fácil. Uma maneira de tentar mensurar isso é através do número de parâmetros do modelo, do tipo de função que ele implementa (linear ou não linear, cortes ortogonais no espaço, ...), ou da chamada dimensão VC (Vapnik-Chervonenkis) do classificador. A dimensão VC é um tópico mais avançado e faz parte da chamada teoria do aprendizado estatístico, ela é citada aqui apenas como curiosidade não é esperado que saibem sobre isso.

Execute a função *classificacao* definida no notebook com a medida de desempenho que você definiu (caso necessário implemente ela, tem um exmeplo no código de como fazer isso). Diga qual o modelo tem o melhor desempenho e explique porque você acha isso.

3- Implementar os procedimentos de amostragem para estimação do desempenho:

- 10-fold Cross Validation;
- Leave-one-out;
- Boostrap (1000 amostras de boostrap).

Faça sub e superamostragem do seu dataset para gerar dois outros datasets, um com ~200 amostras e outro com ~2000 amostras. Tente fazer essa amostragem mantendo a distribuição das classes.

Para os dois datasets criados executar as três maneiras de estimar o desempenho. Avalie a diferença na variância entre essas abordagens.

Para o 10-fold Cross Validation e o leave-one-out você pode utilizar a função *classificacao* já disponível apenas ajustando o parâmetro *folds*. No boostrap utilize 80% do dataset para treino e 20% para teste.

Essas execuções podem demorar um pouco, então tenham paciência.

No notebook tem exemplos de como fazer isso.

UNIVERSIDADE DE SÃO PAULO - ICMC
Departamento de Ciências de Computação
SCC-0275 – Introdução à ciência de dados – Graduação - 2° Sem /2018
PROFa: Roseli Aparecida Francelin Romero
Estagiário PAE: Daniel Moreira Cestari

4- Escolher um dos procedimentos da questão anterior (10-fold cross validation, leave-one-out ou bootstrap) e calcular as medidas de avaliação baseadas na matriz de confusão (TFP, TFN, TVN, TVP). Como em meio a tantas medidas de avaliação, comparar os classificadores? Como escolher o melhor?

Entrega

A entrega da prática deve ser feita em formato de notebook do jupyter. Crie um notebook seu, nas primeiras linhas identifique os membros da dupla, salve o notebook com o nome pratica_2_nUSP1_nUSP2. Submeta no tidia apenas o arquivo do notebook .ipynb, não crie pastas para separar as práticas, apenas faça o upload do arquivo .ipynb Não precisa fazer o upload dos datasets utilizados. A resolução desta prática deve seguir a mesma maneira da anterior. No corpo da prática primeiro identifique qual o conjunto de dados escolhido e começe a responder as perguntas no corpo do notebook. Procure utilizar as caixas de texto para a discussão (o notebook tem caixas de texto e de código).

Os dois alunos precisam submeter a prática no seu respectivo escaninho

O prazo de entrega é até 12/10 às 23:59

Funções utilizadas na prática

http://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics

- confusion_matrix(y_true, y_pred, labels=None, sample_weight=None)