Instituto de Matemática e Estatística - USP

MAC0459/MAC5865 - Data and Engineering Science

General Test - 11/2019 - Partially based on Skiena's "The Data Science Design Manual" book.

By uploding this test I declare that I have worked on it alone and I know that I can be failed in the discipline if I break this circle of confidence.

Good luck !

Q1. This site bellow has data on flights and flights' operations in Brazil.

`https://www.anac.gov.br/dadosabertos/areas-de-atuacao/voos-e-operacoes-aereas`

Your task in this question is to check the dataset bellow and formulate three questions to the dataset and answer them using EDA methods and visualizations.

`https://www.anac.gov.br/dadosabertos/areas-de-atuacao/voos-e-operacoes-aereas/dados-estatisticos-do-transporte-aereo`

The expected answer is to be given in a notebook begining with a brief description of the dataset and your strategy to answer your questions.

Q2. Enrich your questions and answers with another one of the datasets:

- Domestic flights' fares.
- Percentage of delays and cancellations.

Q3. The L_k distance metrics seen in one of the classes implies equal weights for all space dimensions. One alternative to this idea is to weight the features differently:

$$d_k(p,q) = (\sum_{i=1}^{d} c_i |p_i - q_i|^k)^{\frac{1}{k}}$$

for all p, $q \in R^d$. This is called *dimension-weighted* L_p distance. Show that this definition satisfies the distance properties.

Q4. Perform k-means clustering manually on the following points, for $k = 2$:

$$S = \{(1,4),(1,3),(0,4),(5,1),(6,2),(4,0)\}$$

Present the points and the final clusters.

Q5. Answer these questions the more complete you can.

Suppose you build a classifier that answer *yes* on every possible input. What precision and recall will this classifier achieve?

Explain what precision and recall are. How do they relate to the ROC curve?

Is it better to have too many false positives, or too many false negatives? Explain.

What is cross-validation? How might we pick the right value of k for k-fold cross-validation?

Explain why we have training, test, and validation data sets and how they are used effectively?

Explain why we need so many performance measures for a classifier (precision, recall, F1, accuracy etc) and how they are used in practice?

Q6. Search your favorite News Websites until you find four (**six for Grad Students**) interesting charts/plots, ideally half good and half bad. For each, please critique along the following dimensions, using the vocabulary we have developed in this discipline:

Does it go a good job or a bad job presenting the data?

Does the presentation appear to be biased, either deliberately or accidentally?

Are the axes labeled in a clear and informative way?

Is the color used effectively?

How can we make the graphic better?

Q7. Read the paper: "How Algorithms Discriminate Based on Data They Lack: Challenges, Solutions, and Policy Implications" (paca.ime.usp.br) and write an abstract no longer than two pages. Your abstract should have a general idea of the paper, the main argument, methods and your critical analysis of the method.

Q8. For Grad Students Only Read the paper: "A survey on measuring indirect discrimination in machine learning" by (paca.ime.usp.br) and write an abstract no longer than four pages. Your abstract should have a general idea of the survey, the main results and the advantages and disadvantages for each method surveyed.

Q10. Auto-avaliação

1. Você assistiu a todas as aulas presenciais? Se não, o que a/o fez não assistir?
2. Como você avalia seu entendimento das aulas ao conteúdo?
3. Você participou das tarefas em grupo realizadas em sala de aulas? Se não, o que a/o fez não participar? Como você avalia a sua participação nas discussões do grupo?
4. Você procurou conhecer mais profundamente algum método exposto em aulas? Lembra quais? Qual a sua avaliação, em termos de entendimento, desse seu estudo?
5. Quantas horas por semana você tem para estudar extra-classe? Dessas horas, quantas você usou para acompanhar esta disciplina?
6. Você tem motivação para vir às aulas e participar das discussões? Se sim, o que aumentaria ainda mais sua motivação? Se não, o que você sugere para que as aulas sejam motivadoras?
7. A disciplina satisfez suas expectativas? Faça uma análise crítica dos objetivos pretendidos por você ao se matricular e dos objetivos que você alcançou.
8. Além do que foi dado, o que mais você gostaria de aprender? Existe algum tópico você gostaria que fosse aprofundado?
9. A quantidade e a dificuldade das tarefas práticas foi adequada e suficiente? Você sente-se seguro para continuar estudando o assunto sozinho? O que poderia ser melhorado nas tarefas? Avalie cada uma das tarefas.
10. Faça uma breve auto-avaliação de seu desempenho considerando sua facilidade com o conteúdo da disciplina, seu entendimento dos conceitos, sua assiduidade às aulas, sua participação em aulas e seu desempenho nas tarefas. Se você se sentir confortável para isso, atribua-se uma nota de 0 a 10 de acordo com sua auto-avaliação. Essa nota será considerada no cômputo final da sua nota, caso o professor entenda que você soube se auto-avaliar.