Exame de AA / AA Exam.

Pergunta 1; Question 1

Português

Esta pergunta faz parte da matéria do primeiro teste.

Atenção: a sua resposta deve ser concisa e bem estruturada. Se vir que não tem tempo para dar uma resposta estruturada que cubra tudo o que é pedido, foque apenas os itens mais importantes. Lembre-se que é proibido copiar texto do material de consulta; responda pelas suas palavras.

Tem um conjunto de dados com 1500 pontos e 7 atributos numéricos, dividido em 2 categorias. Quer treinar um classificador para prever a categoria de novos exemplos obtidos da mesma população da qual obteve este conjunto. Tem também vários classificadores que quer experimentar. Descreva como faria para escolher o melhor classificador, treiná-lo e estimar o erro que irá ter no futuro quando tentar prever a classe de novos exemplos.

Deve focar estes aspectos por ordem decrescente de importância:

- Como organiza os seus dados, explicando porquê.
- Como escolhe o melhor modelo para classificar estes dados.
- Como treina esse modelo para obter o seu classificador final treinado.
- Como estima o erro que vai ter em exemplos futuros.
- Qual a medida de erro que vai usar e porquê.

English:

This question is part of the subjects covered in the first test.

Attention: your answer should be concise and well structured. If you feel you do not have time to give a well structured answer covering all items below focus on the most important ones. Remember that you may not copy text from other sources; answer using your own words.

You have a data set with 1500 points and 7 numeric attributes, divided into 2 categories. You want to train a classifier to predict the category of new examples obtained from the same population from which you obtained this set. There are several classifiers that you

want to try. Describe how you would choose the best classifier, train it and estimate the error you will have in the future when trying to predict the class for new examples.

You should focus on these aspects in decreasing order of importance:

- How you organize your data, explaining why.
- How you choose the best model to classify this data.
- How you train this model to get your final classifier trained.
- How do you estimate the error you will have in future examples.
- What error measure will you use and why.

Pergunta 2; Question 2

Português

Esta pergunta faz parte da matéria do primeiro teste.

Atenção: a sua resposta deve ser concisa e bem estruturada. Se vir que não tem tempo para dar uma resposta estruturada que cubra tudo o que é pedido, foque apenas os itens mais importantes. Lembre-se que é proibido copiar texto do material de consulta; responda pelas suas palavras.

Tem numa tabela as coordenadas de avistamentos de lobos e pontos preferenciais de pastoreio de ovelhas numa região rural. Para minimizar conflitos entre lobos e pastores, quer-se pôr uma vedação a separar o melhor possível as zonas de pasto e o território dos lobos.

Por questões económicas, assuma que a rede deve ser colocada numa linha recta para minimizar o seu custo. Explique o algoritmo que utilizaria para colocar esta rede em linha recta o mais afastada possível dos pontos mais próximos (lobos ou zonas de pasto), assumindo que é possível separar estes conjuntos com uma linha recta.

O que deve fazer se quiser uma rede em linha recta mas não for possível separar por completo lobos e ovelhas, e tiver de encontrar a recta que separa o maior número?

E se o orçamento permitir uma vedação mais curva, que alternativa deve usar?

Deve focar, por ordem decrescente de importância, estes aspectos:

- O funcionamento do algoritmo que escolheu para o primeiro caso.
- Modificações ou algoritmo alternativo que usaria para o segundo caso.
- Finalmente, o que faria no terceiro caso.

English:

This question is part of the subjects covered in the first test.

Attention: your answer should be concise and well structured. If you feel you do not have time to give a well structured answer covering all items below focus on the most important ones. Remember that you may not copy text from other sources; answer using your own words.

You have a table with the coordinates of sightings of wolves and preferential points of sheep grazing in a rural region. To minimize conflicts between wolves and shepherds, we want to put a fence separating as much as possible the pastures and the wolf territory.

For economic reasons, assume that the fence must be placed in a straight line to minimize its cost. Explain the algorithm you would use to place this fence in a straight line as far as possible from the nearest points (wolves or pasture areas), assuming that it is possible to separate these sets with a straight line.

What should you do if you want a straight line but it is not possible to completely separate wolves and sheep, and you have to find the line that separates the largest number of points?

And if your budget allows for a more curved fence, what alternative should you use?

You should focus on these aspects in decreasing order of importance:

- How the algorithm you chose for the first case works.
- What modifications or alternative algorithm you would need for the second case.
- Finally, what would you do in the third case.

Pergunta 3; Question 3

Português

Esta pergunta faz parte da matéria do primeiro teste.

Atenção: a sua resposta deve ser concisa e bem estruturada. Se vir que não tem tempo para dar uma resposta estruturada que cubra tudo o que é pedido, foque apenas os itens mais importantes. Lembre-se que é proibido copiar texto do material de consulta; responda pelas suas palavras.

Quer usar um classificador Naïve Bayes para prever problemas oncológicos com base em características das pessoas, como peso, se é fumador, se tem cinzeiros em casa, se faz exercício físico, etc. As categorias são duas: saudável e doente (com cancro). Segundo este modelo, se sabemos que uma pessoa é saudável, então ser fumador aumenta a probabilidade de ter cinzeiros em casa, diminui essa probabilidade ou não faz qualquer diferença? Explique porquê e use este exemplo para explicar também a diferença entre este modelo e o classificador completo de Bayes, indicando as vantagens e desvantagens de cada um.

Deve focar estes aspectos por ordem decrescente de importância:

- Responder à pergunta justificando a sua resposta.
- Explicar a diferença entre o classificador Naïve Bayes e o classificador completo de Bayes.
- Explicar as vantagens e desvantagens relativas de cada um destes classificadores.

English:

This question is part of the subjects covered in the first test.

Attention: your answer should be concise and well structured. If you feel you do not have time to give a well structured answer covering all items below focus on the most important ones. Remember that you may not copy text from other sources; answer using your own words.

You want to use a Naïve Bayes classifier to predict cancer problems based on people's characteristics, such as weight, if they smoker, if they have ashtrays at home, if they exercise, etc. There are two categories: healthy and sick (with cancer). According to this model, if we know that a person is healthy, then being a smoker increases the probability of having ashtrays at home, decreases that probability or does it make no difference? Explain why and use this example to also explain the difference between this model and the complete Bayes classifier, indicating the advantages and disadvantages of each.

You should focus on these aspects in decreasing order of importance:

- Answer the question, justifying your answer.
- Explain the difference between the Naïve Bayes classifier and the complete Bayes classifier.
- Explain the relative advantages and disadvantages of each of these classifiers.

Pergunta 4; Question 4

Português

Esta pergunta faz parte da matéria do primeiro teste.

Atenção: a sua resposta deve ser concisa e bem estruturada. Se vir que não tem tempo para dar uma resposta estruturada que cubra tudo o que é pedido, foque apenas os itens mais importantes. Lembre-se que é proibido copiar texto do material de consulta; responda pelas suas palavras.

Treinou um modelo de classificação no seu conjunto de dados. O erro de treino é de 1%, medido em exemplos mal classificados. O erro de validação é de 23%. Explique como poderia treinar e usar várias instâncias deste modelo para tentar mitigar este problema.

E se o erro de treino fosse igual ao de validação, ambos de 23%? Usaria a mesma abordagem ou outra? Justifique a sua resposta explicando como faria nesse caso para usar um conjunto de instâncias do seu modelo de forma a mitigar o problema.

Deve focar estes aspectos por ordem decrescente de importância:

- Explicar a abordagem no primeiro caso em que o erro de treino é muito menor, indicando como treina e usa as instâncias do classificador.
- Explicar se o segundo caso é diferente ou se pode resolvido com a mesma abordagem, justificando.
- Caso seja necessário outra abordagem, explicá-la, indicando como treina e usa as várias instâncias do classificador.

English:

This question is part of the subjects covered in the first test.

Attention: your answer should be concise and well structured. If you feel you do not have time to give a well structured answer covering all items below focus on the most important ones. Remember that you may not copy text from other sources; answer using your own words.

You trained a classification model using your data set. The training error is 1%, measured in incorrectly classified examples. The validation error is 23%. Explain how you would train and use multiple instances of this model to try to mitigate this problem.

What if the training error was the same as the validation error, both 23%? Would you use the same approach or a different one? Justify your answer by explaining how you would train and use a set of instances of your model in order to mitigate the problem in this case.

You should focus on these aspects in decreasing order of importance:

- Explain the approach in the first case where the training error is much lower, indicating how you train and use the classifier instances.
- Explain if the second case is different or if it can be solved with the same approach, justifying your answer.
- If another approach is needed, explain it, indicating how you train and use the various instances of the classifier.

Pergunta 5; Question 5

Português

Esta pergunta faz parte da matéria do segundo teste.

Atenção: a sua resposta deve ser concisa e bem estruturada. Se vir que não tem tempo para dar uma resposta estruturada que cubra tudo o que é pedido, foque apenas os itens mais importantes. Lembre-se que é proibido copiar texto do material de consulta; responda pelas suas palavras.

Imagine que um colega, entusiasmado com a matéria dada em Aprendizagem Automática, lhe diz ter encontrado uma forma de melhorar o desempenho de qualquer classificador linear. O seu colega explica-lhe que a equação abaixo relaciona, com probabilidade 1 menos delta, o limite superior do erro verdadeiro da hipótese de menor erro empírico com o seu erro empírico e um termo adicional cuja magnitude é da ordem da expressão indicada, em função da dimensão Vapnik-Chervonenkis, delta e o tamanho do conjunto de treino:

.

Com base nesta equação, o seu colega propõe que, como o classificador é linear, uma expansão linear dos atributos vai aumentar a dimensão VC do classificador. E como a dimensão VC está relacionada na equação com o tamanho do conjunto de treino, isto vai permitir treinar o classificador com mais exemplos, melhorando assim o seu desempenho.

Comente a proposta do seu colega apontando os erros que encontrar ou, se não tiver erros, explicando como funcionaria.

Deve focar, por ordem decrescente de importância, estes aspectos:

- O efeito da expansão linear dos atributos num classificador linear.

- O que a equação permite prever acerca do tamanho do conjunto de treino.
- O resultado esperado, na prática, de fazer o que o colega propõe.

English:

This question is part of the subjects covered in the second test.

Attention: your answer should be concise and well structured. If you feel you do not have time to give a well structured answer covering all items below focus on the most important ones. Remember that you may not copy text from other sources; answer using your own words.

Imagine that a colleague, enthusiastic about the subjects covered in Machine Learning, tells you that he has found a way to improve the performance of any linear classifier. Your colleague explains that the equation below relates, with probability 1 minus delta, the upper limit of the true error of the hypothesis of least empirical error with its empirical error and an additional term whose magnitude is in the order of the indicated expression, depending on the Vapnik-Chervonenkis dimension, delta and the size of the training set:

.

Based on this equation, your colleague proposes that, since the classifier is linear, a linear expansion of the attributes will increase the VC dimension of the classifier. And as the VC dimension is related in the equation to the size of the training set, this will allow you to train the classifier with more examples, thus improving its performance.

Comment on your colleague's proposal by pointing out the errors you find or, if it has no errors, explaining how it would work.

You should focus, in decreasing order of importance, on these aspects:

- The effect of linear expansion of attributes in a linear classifier.
- What the equation allows to predict about the size of the training set.
- The expected result, in practice, of doing what the colleague proposes.

Pergunta 6; Question 6

Português

Esta pergunta faz parte da matéria do segundo teste.

Atenção: a sua resposta deve ser concisa e bem estruturada. Se vir que não tem tempo para dar uma resposta estruturada que cubra tudo o que é pedido, foque apenas os itens mais importantes. Lembre-se que é proibido copiar texto do material de consulta; responda pelas suas palavras.

Tem uma tabela com os registos de manutenção de 286 viaturas. Cada linha da tabela corresponde a uma viatura e nas colunas tem 38 atributos numéricos como a quilometragem total, número de reparações prévias, consumo, preço de compra, idade em anos, entre

outros. Noutra coluna tem a avaliação de um mecânico indicando se a viatura ainda pode ser usada ou se deve ir para abate. Descreva o que faria para encontrar os seis melhores atributos para prever se uma viatura que encontre mais tarde ainda está funcional ou se deve ir para abate.

Explique também o que faria se em vez de seleccionar 6 atributos quisesse transformar estes 38 atributos numéricos em 6 atributos que melhor representassem a dispersão de valores nesta tabela.

Deve focar, por ordem decrescente de importância, estes aspectos:

- Como processaria os dados.
- Como seleccionaria os 6 atributos melhores.
- Como funciona o algoritmo que iria usar para reduzir estes 38 atributos a 6 atributos, e porque é que utilizaria esse algoritmo.

English:

This question is part of the subjects covered in the second test.

Attention: your answer should be concise and well structured. If you feel you do not have time to give a well structured answer covering all items below focus on the most important ones. Remember that you may not copy text from other sources; answer using your own words.

You have a table with the maintenance records of 286 vehicles. Each row in the table corresponds to a vehicle and in the columns have 38 numerical attributes such as total mileage, number of previous repairs, purchase price, age in years, among others. Another column has the evaluation of a mechanic indicating whether the vehicle can still be used or whether it should be scrapped. Describe what you would do to find the six best attributes to predict, for future vehicles, if it is still functional or should be scrapped.

Also explain what you would do if, instead of selecting 6 attributes, you wanted to transform these 38 numeric attributes into 6 attributes that best represent the dispersion of values in this table.

You should focus, in decreasing order of importance, on these aspects:

- How you would process the data.
- How you would select the 6 best attributes.
- How does the algorithm work that you would use to reduce these 38 attributes to 6 attributes, and why use that algorithm.

Pergunta 7; Question 7

Português

Esta pergunta faz parte da matéria do segundo teste.

Atenção: a sua resposta deve ser concisa e bem estruturada. Se vir que não tem tempo para dar uma resposta estruturada que cubra tudo o que é pedido, foque apenas os itens mais importantes. Lembre-se que é proibido copiar texto do material de consulta; responda pelas suas palavras.

Biólogos querem estudar a diversidade genética da *Vespa velutina* (vulgo vespa asiática) numa região. Para isso determinaram as coordenadas de 183 ninhos deste vespa. Agora querem aglomerar esses ninhos de forma a encontrar ninhos representativos de diferentes aglomerados mas não querem ter de especificar à partida quantos aglomerados querem formar. Indique que algoritmo de aglomeração (*clustering*) propõe usar neste caso, descrevendo como funciona e justificando a sua adequação pelas propriedades do algoritmo.

Além disso, um dos biólogos diz que ouviu falar num algoritmo chamado K-Means e pergunta se esse seria bom para esta situação.

Deve focar, por ordem decrescente de importância, estes aspectos:

- O funcionamento do algoritmo que escolheu usar.
- Porque é útil neste caso, explicando como as suas propriedades ajudam a resolver o problema.
- O que acha da aplicação do K-Means a este problema, justificando a resposta.

English:

This question is part of the subjects covered in the second test.

Attention: your answer should be concise and well structured. If you feel you do not have time to give a well structured answer covering all items below focus on the most important ones. Remember that you may not copy text from other sources; answer using your own words.

Biologists want to study the genetic diversity of *Vespa velutina* (Asian wasp) in a region. To do this, they determined the coordinates of 183 nests of this wasp. Now they want to cluster these nests in order to find representative nests from different clusters but they do not want to specify at the outset how many clusters they want to form. Indicate which clustering algorithm you propose to use in this case, describing how it works and justifying its suitability by the properties of the algorithm.

In addition, one of the biologists says she heard about an algorithm called K-Means and asks if it would be good for this situation.

- How the algorithm you chose to use works.
- Why it is useful in this case, explaining how its properties help solve the problem.
- What you think of the application of K-Means to this problem, justifying your answer.

Pergunta 8; Question 8

Português

Esta pergunta faz parte da matéria do segundo teste.

Atenção: a sua resposta deve ser concisa e bem estruturada. Se vir que não tem tempo para dar uma resposta estruturada que cubra tudo o que é pedido, foque apenas os itens mais importantes. Lembre-se que é proibido copiar texto do material de consulta; responda pelas suas palavras.

Num mapa tem indicados 672 pontos onde foram detectados cardumes de sardinhas. Os pescadores estão interessados em saber quais as zonas em que é mais fácil encontrar sardinhas. Não sabem se será só uma grande zona contígua ou se há sardinhas em várias regiões, mas estão interessados principalmente em distinguir entre as zonas melhores de pesca e meros avistamentos pontuais que possam não ser relevantes. Indique que algoritmo de aglomeração (*clustering*) propõe usar neste caso, descrevendo como funciona e justificando a sua adequação pelas propriedades do algoritmo.

Além disso, um dos pescadores, que é um aficionado da aprendizagem automática, diz que ouviu falar num algoritmo chamado Mistura de Gaussianas (Gaussian Mixture Models, GMM) e pergunta se esse seria bom para esta situação.

Deve focar, por ordem decrescente de importância, estes aspectos:

- O funcionamento do algoritmo que escolheu usar.
- Porque é útil neste caso, explicando como as suas propriedades ajudam a resolver o problema.
- O que acha da aplicação do GMM a este problema, justificando a resposta.

English:

This question is part of the subjects covered in the second test.

Attention: your answer should be concise and well structured. If you feel you do not have time to give a well structured answer covering all items below focus on the most important ones. Remember that you may not copy text from other sources; answer using your own words.

On a map there are 672 points where schools of sardines were detected. Fishermen are interested in knowing which areas are best for fishing sardines. They do not know if it will be just a large contiguous area or if there are sardines in several regions, but they are mainly interested in distinguishing between the best fishing areas and mere occasional sightings that may not be relevant. Indicate which clustering algorithm you propose to use in this case, describing how it works and justifying its suitability by the properties of the algorithm.

In addition, one of the fishermen, who is an aficionado of machine learning, says he heard about an algorithm called Gaussian Mixture Models (GMM) and asks if that would be good for this situation.

You should focus, in decreasing order of importance, on these aspects:

- How the algorithm you chose to use works.
- Why it is useful in this case, explaining how its properties help to solve the problem.
- What you think of the application of GMM to this problem, justifying your answer.