Machine Learning
CPS 863
Terceiro Trimestre de 2024

Professor: Edmundo de Souza e Silva

Lista de Exercícios 1

ATENÇÃO!

- Faça as listas de forma que TODAS AS RESPOSTAS sejam DEVIDAMENTE COMENTADAS (passos para se chegar a resposta).
- A entrega da lista deve ser feita em UM ÚNICO arquivo PDF. Não envie vários pedaços separadamente!
- ATENÇÃO! Faça as listas de forma que TODAS AS RESPOSTAS sejam DEVIDAMENTE COMENTADAS (passos para se chegar a resposta).
  Não procure a solução na Internet ou em livros ou no chatGPT, pois o objetivo é que você mesmo avalie o que sabe. Obviamente, caso você já tenha conhecimento do problema, não leia a resposta (mesmo que já conheça o resultado final) e tente fazer sozinho. Só assim você poderá ter uma ideia melhor dos tópicos que você ainda não domina com desenvoltura.
- Anote as dúvidas encontradas para resolver **sozinho**. Em classe gostaria de saber quais as dúvidas que cada um teve para resolver o problema sem olhar a resposta.
- Qualquer referência a código é MUITO menos importante do que a EXPLICAÇÃO DOS PASSOS que foram realizados. O que mais importa é a explicação de como se chegou na solução.
- Para facilitar escrever a lista de forma clara, é possível traduzir equações a mão para LaTex: https://mathpix.com/, ver também https://www.overleaf.com/learn/latex/Questions/Are_there_any_tools_to_help_transcribe_mathematical_formulae_into_LaTeX%3F

Questão 1

Recordação

Uma caixa contém três moedas: duas são normais e uma moeda falsa com duas caras (P(Ca)=1). Se você pegar uma moeda da caixa e jogá-la, qual a probabilidade de sair cara? Se você pegar uma moeda da caixa e jogá-la, e sair cara, qual a probabilidade de ser a moeda falsa?

Questão 2

Material introdutório

Uma urna U_A tem $N = 1000$ bolas sendo 25% delas azuis e o restante pretas. Uma outra urna U_B também contém $N = 1000$ bolas, mas apenas 10% delas são azuis (e o restante pretas).

As urnas são idênticas externamente, exceto por uma marcação, U_A, U_B, que permite a identificação de cada uma. Entretanto, essa identificação está na parte inferior das urnas, de forma que não é possível visualizar o rótulo, exceto se a urna for levantada.

- João tira (de olhos vendados) 2 bolas azuis de uma das urnas. Você vai ter que adivinhar a urna escolhida. Se a probabilidade de João escolher uma das urnas for a mesma, qual a aposta que você fará? Note que, para fazer a aposta, você precisa determinar qual a probabilidade das bolas serem provenientes da urna U_A.
  Você tem confiança na sua aposta? Por que?

- Um amigo seu diz que João sabe a posição das urnas e escolhe a urna U_A com probabilidade 0.15.
  Sua aposta mudaria? Você teria confiança na sua aposta? Justifique a resposta.

Questão 3

Considere um *dataset* cujas amostras são obtidas independentemente a partir de uma distribuição **discreta** uniforme $U(1, 5)$ (https://en.wikipedia.org/wiki/Discrete_uniform_distribution).

Considere um *dataset* com as seguintes amostras: $\{2, 2, 4, 3, 2\}$.

1. Qual a verossimilhança (*likelihood*) de observar essas amostras?
2. E o log-likelihood?

Questão 4

Assume you are given a biased coin such that with probability p you obtain heads (H) and $(1 - p)$ tails (T). You toss the coin N times and obtain N_H heads (and $N - N_H$ tails, of course)

1. Obtain likelihood function $\mathcal{L}(\theta|\mathcal{D}) = p(\mathcal{D}|\theta)$ were θ is parameter vector. (The model is clearly very simple in this case.) What is the value of $p(\mathcal{D}|\theta)$ if $\mathcal{D} = \{HHTHTTHTTT\}$ and $p = 0.2$? And if $p = 0.6$?
2. For $\mathcal{D}$ given in the item above, find p that optimizes the log likelihood. In general, find p as a function of N and N_H for any given set $\mathcal{D}$.

Questão 5

Suponha agora que suas amostras são obtidas ou de uma distribuição **discreta** $U(1, 5)$ ou a partir de um dado (seis faces) sendo que todas as amostras são obtidas da **mesma distribuição**. Suponha que o a probabilidade das amostras serem obtidas do dado é igual a p. Considere o conjunto de dados $\{2, 2, 4, 3, 2\}$, e seja $p = 0.7$.

1. Qual a *likelihood* das amostras serem retiradas a partir: (a) do dado se seis faces, ou (b) de uma $U(1, 5)$ discreta?
2. Qual o *posterior distribution*?
3. Uma vez que o dataset acima foi observado, qual a probabilidade de se retirar o número 5?
4. Uma vez que o dataset acima foi observado, qual a probabilidade de se retirar o número 6?
5. Qual o MLE?
6. Qual o MAP?
7. Caso $p = 0.5$, quais das respostas acima mudam de valor? Explique.

Questão 6

Suponha agora que suas amostras são obtidas ou de uma distribuição **discreta** $U(1, 5)$ ou a partir de um dado (seis faces) *com probabilidade* $(1 - p)$ e p, respectivamente. Entretanto, nesta questão, a sequência pode conter amostras de ambas distribuições (mistura de distribuições). Considere o mesmo conjunto de dados $\{2, 2, 4, 3, 2\}$, e seja $p = 0.7$.

1. Qual a *likelihood* de observar essas amostras?
2. Uma vez que o dataset acima foi observado, qual a probabilidade de se retirar o número 5?
3. Uma vez que o dataset acima foi observado, qual a probabilidade de se retirar o número 6?
4. Qual a probabilidade de todas as amostras serem retiradas a partir: (a) do dado se seis faces, ou (b) de uma $U(1, 5)$ discreta?
5. Calcule o *likelihood function* para as amostras em função de p, o log likelihood e obtenha o valor de p que melhor explica o conjunto de dados. Comente a sua resposta.
6. Repita o item anterior, supondo que o conjunto de dados tem cardinalidade 20 e apenas uma única amostra tenha valor igual a 6.