IBI5031 - Aprendizagem de Máquina para Bioinformática

Primeira Lista de Exercícios

Docente: Marcelo da Silva Reis[1]

[1] Laboratório de Ciclo Celular, Instituto Butantan

Programa de Pós-Graduação Interunidades em Bioinformática da USP
São Paulo, 16 de outubro de 2020

Instruções

- Esta lista de exercícios é para ser resolvida de forma estritamente individual;
- O(a) aluno(a) pode resolver as questões dissertativas à mão e depois digitalizá-las ou então utilizar algum editor de texto (e.g., LaTeX); em qualquer uma dessas alternativas, deverá ser entregue um único arquivo no formato PDF;
- Para questões que envolvam experimentos computacionais, a parte de programação deverá ser feita no Jupyter notebook - utilizar um único caderno (arquivo) para todos os códigos desta lista. Se preferir, pode escrever as questões dissertativas também no notebook (com Markdown), mas se esta for a sua opção tenha o cuidado de não precisar de pacotes adicionais para visualização de notação matemática;
- É importante que as respostas sejam escritas forma clara e organizada; o mesmo vale para o caderno Jupyter, cujo código deverá ser devidamente comentado (intercale códigos com explicações utilizando o Markdown);
- A entrega dos dois arquivos (PDF e/ou caderno Jupyter) deverá ser feita no eDisciplinas, página oficial desta edição de IBI5031, até o dia do prazo final;
- Prazo final de entrega: **26 de outubro**;
- Bom trabalho!

Questões

1. *(2 pontos)* Existem duas caixas, A e B, cada uma contendo bolinhas vermelhas e verdes. Suponha que a caixa A contém uma bolinha vermelha e duas bolinhas verdes, enquanto que a caixa B contém oito bolinhas vermelhas e duas bolinhas verdes. Considere o seguinte procedimento: uma bolinha é selecionada aleatoriamente da caixa A e outra bolinha é selecionada aleatoriamente da caixa B. Na sequência, a bolinha selecionada na caixa A é transferida para a caixa B e vice-versa. Esse procedimento é iterado indefinidamente.

a) Considerando que A_t^V é uma variável aleatória cuja realização é o número de bolinhas vermelhas na caixa A na t-ésima iteração, construa uma matriz $\mathbf{M}$ definida como:

$$\mathbf{m}_{ij} = Pr(A_{t+1}^V = j - 1 \mid A_t^V = i - 1).$$

b) Suponha que $A_t^V = 3$; qual é a probabilidade de $A_{t+1}^V = A_{t+2}^V = 3$?

c) Suponha novamente que $A_t^V = 3$; qual é a probabilidade de $A_{t+1}^V = 2$ dado que $A_{t-1}^V = 1$?

d) Considere o seguinte vetor:

$$\mathbf{v} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}.$$

Calcule o produto $\mathbf{v}^t\mathbf{M}\mathbf{M} = \mathbf{v}^t\mathbf{M}^2$. Como o resultado desse produto se relaciona com o obtido no item (b)?

2. *(2 pontos)* Considere a versão "vanilla" da desigualdade de Hoeffding, definida como:

$$Pr(|\nu - \mu| > \epsilon) \leq \underbrace{2e^{-2\epsilon^2 N}}_{\delta}.$$

a) Calcule o tamanho N da amostra necessário para que a desigualdade seja satisfeita para $\epsilon = \frac{1}{10}$ e $\delta = \frac{1}{20}$.

b) Suponha que temos uma moeda viciada, que dá cara em 57 porcento dos lançamentos. Considere o seguinte algoritmo, para ϵ com o mesmo valor do item anterior:

1 $\mu \leftarrow 0.57$
2 **for** $N = 1$ **to** 1000
3   **do** $\mathit{violou} \leftarrow 0$
4   **for** $k = 1$ **to** K
5   **do** $\nu \leftarrow$ fração de caras em N sorteios da moeda
6                 $\mathit{violou} \leftarrow \mathit{violou} + [\![|\nu - \mu| > \epsilon]\!]$
7   Imprima $\frac{1}{K}\mathit{violou}$

No código acima, K é um número de repetições (defina ele como 1000) e a função $[\![\mathit{expr}]\!]$ devolve 1 se o argumento *expr* for verdadeiro e 0 caso contrário. Implemente esse pseudocódigo e execute-o no Jupyter notebook. Usando `matplotlib`, imprima os valores obtidos na linha 7 em um gráfico, onde o eixo das abscissas é um N e a ordenada correspondente é a média, para esse N, de violações do erro máximo ϵ permitido. Analise e discuta o resultado obtido, relacionando-o com o resultado obtido no item (a).

c) Repita o procedimento do item anterior, porém utilizando $\mu = 0.26$. Como essa mudança afeta o resultado? E com $\mu = 0.14$? E com $\mu = 0.82$? Analise esses resultados, também relacionando-os ao do item (a).

3. *(6 pontos)* Nesta questão vamos "pôr a mão na massa" em um conjunto de dados biológicos que é clássico em Aprendizagem de Máquina. Utilizando o módulo `sklearn.datasets`, carregue no Jupyter o conjunto de dados Iris:

```
from sklearn import datasets
iris = datasets.load_iris()
```

Trata-se de um conjunto de dados com três classes ($y \in \{0, 1, 2\}$), dimensão quatro nas características ($\mathbf{x}^t = [x_1, x_2, x_3, x_4]$) e 50 pares de observações $(\mathbf{x}, y)$ por classe ($N = 150$). Com `iris.data` e `iris.target` obtemos, respectivamente, os valores observados das características e as classes correspondentes

a) Implemente um algoritmo de aprendizagem de perceptron (PLA) que dê suporte a pares de observações $(\mathbf{x}, y)$ cujo vetor $\mathbf{x}$ tenha dimensão quatro (implementar um algortimo para n dimensões é bem vindo). O seu algoritmo PLA deverá ser um da versão "pocket" (que a cada iteração guarda a melhor solução até então) e também deverá ter um critério de parada, definido por você, para os casos em que o problema não é linearmente separável.

b) Construa três subconjuntos de amostras, com o primeiro, segundo e terceiro subconjuntos contendo observações cujas classes estão em $\{0, 1\}$, $\{0, 2\}$ e $\{1, 2\}$, respectivamente. Para cada subconjunto, execute o seu algoritmo PLA e calcule o erro de classificação dentro da amostra. Analise e comente os resultados obtidos. *Dica:* observe que, dependendo da sua implementação de PLA, poderá ser preciso mapear pares de rótulos para $\{-1, +1\}$.

c) Utilizando os três classificadores obtidos no item (b), crie um quarto classificador g, definido como:

$$g(\mathbf{x}) = \begin{cases} \text{a moda do conjunto } \{g_{0,1}(\mathbf{x}), g_{0,2}(\mathbf{x}), g_{1,2}(\mathbf{x})\}, \text{ se a moda é única;} \\ \text{um valor aleatório em } \{0, 1, 2\}, \text{ caso contrário.} \end{cases}$$

Aplique g sobre toda a amostra (os 150 pares de observações totais) e calcule o erro dentro da amostra. Como existe um fator de aleatoriedade, considere a média de 1000 experimentos. Analise e discuta os resultados obtidos, comparando-os com os do item anterior.