Machine Learning

CPS-863, COS-623, MAB-608 Terceiro Trimestre de 2017

Professores:
Edmundo de Souza e Silva, Daniel S. Menascé, Rosa Leão

Terceira e Quarta Listas de Exercícios (Graduação e Pós-Graduação)

ATENÇÃO! Faça as listas de forma que TODAS AS RESPOSTAS sejam DEVIDAMENTE COMENTADAS (passos para se chegar a resposta).

O objetivo desta e das questões subsequentes é comparar vários modelos que mais se adequem a um conjunto de dados. O propósito não é obter "o melhor" modelo dentre os estudados, mas mostrar se você sabe aplicar a teoria aprendida em classe e avaliar as opções e os resultados encontrados.

Nesta lista, iremos utilizar dados **reais** fornecidos gentilmente pelo Professor Claudio Gil Soares de Araujo (até recentemente professor do Instituto do Coração Edson Saad da UFRJ) da CLINIMEX, através da aluna de doutorado da UFRJ Christina G. de Souza e Silva. Os dados foram obtidos a partir de uma extensa base de dados do Prof. Claudio Gil, coletada durante muitos anos e usada em suas pesquisas. Os dados mostram uma medida da condição aeróbica do paciente (o VO_2max) (por quilo de peso do indivíduo) e ainda as variáveis idade, peso e a carga máxima atingida durante um teste de *exercício máximo* ao qual o paciente foi submetido. (Os dados são todos de pacientes masculinos.) De forma bem simples, o VO_2max é a taxa máxima de consumo de oxigênio medida durante um teste de exercício máximo, e reflete a capacidade aeróbica do paciente, expressa em volume de oxigênio por massa corporal por minuto (ml/(Kg.min) [1]. É uma importante métrica usada na avaliação cardiovascular de indivíduos [1].

Nesta lista os modelos devem prever VO_2 máximo de pacientes com uma dada idade, peso e carga máxima atingida durante o teste de *exercício máximo* do paciente. Em alguns dos modelos você deverá encontrar uma função adequada do VO_2 máximo em função das outras variáveis (ou de um subconjunto delas). Outros modelos são mais evidentes para se encontrar a probabilidade de se encontrar o VO_2 máximo dentro de uma faixa de valores, a partir dos dados de entrada ou de um subconjunto deles. Em outra questão será solicitado que seja estimada a idade do paciente dado um subconjunto de variáveis.

Seguindo a notação usada em classe, o i-ésimo vetor de dados $\mathbf{x}$ tem dimensão D e o conjunto de treinamento tem N observações. Indicaremos o significado das características do vetor de dados.

As características dos dados fornecidos são: idade do paciente, peso (kg), carga final (watts) e VO_2 máximo (mg/Kg/min). Fornecemos um (1) arquivo com dados de 1.172 pacientes coletados pelo Professor e sua equipe: *Dados-medicos.csv*.

IMPORTANTE: Antes de fazer as questões Você deve escolher aleatoriamente um subconjunto de 1.000 amostras para treinar os modelos e as 172 restantes serão usados para testar o modelo. Indique, no início do seu relatório, o vetor de índice das amostra escolhidas para treinamento e para teste. (Por exemplo: vetor de treinamento $\mathbf{t} = \langle 1, 0, 0, 1, 1, 1, 0, \ldots \rangle$, onde $\mathbf{t}(i) = 1$ se a i-ésima amostra foi escolhida para treinamento, e $\mathbf{t}(i) = 0$ para teste.

Todas as referências sobre a teoria estão em:
`https://www.land.ufrj.br/~classes/machine-learning-2017`.

Sugestão: use o programa **R** (ou Octave, ou python).

Inclua as linhas de código usadas em um apêndice, mas não no meio das respostas.

[1] `https://en.wikipedia.org/wiki/VO2_max`

Terceira Lista

Questão 1

1. Considere as colunas 3 e 4 (carga e VO_2max) dos dados fornecidos. Seja $p(y|\mathbf{x}, \boldsymbol{\theta}) = \mathcal{N}(y|\mathbf{w}^T\phi(\mathbf{x}), \sigma^2)$ onde y é o VO_2max), $\mathbf{x} = \langle \text{carga} \rangle$, $\phi(x) = 1, x, \ldots, x^d$. Use regressão linear e ache os parâmetros do modelo.
   - experimente diferentes valores de d.
   - plote as funções encontradas, junto com os dados de treinamento.
   - qual é o NLL($\mathbf{w}$) encontrado em cada caso? Explique.
   - compare os valores previstos pelo modelo escolhido com os valores reais dos dados de teste. Você está satisfeito com o modelo escolhido?

   Explique e comente a sua resposta.
2. Repita o item anterior considerando as colunas 2, 3 e 4 (peso, carga e VO_2max).
3. Repita o item anterior considerando as colunas 1, 2, 3 e 4 (idade, peso, carga e VO_2max). Você acha que alguma das variáveis é menos importante para se estimar o VO_2max? Ou todas são importantes? Explique.
4. O VO_2max pode ser estimado pela seguinte equação de acordo com o *American College of Sports Medicine*: $VO_2\text{max} \approx (W \times 11.4 + 260 + \text{peso} \times 3,5)/\text{peso}$, onde W é a carga máxima em watts, o peso corporal é em kg e a constante de 260 mL/min indica o volume de oxigênio correspondente ao gasto energético necessário para mover os pedais do teste de esforço sem qualquer resistência adicionada.
   Compare essa equação com a do modelo que você obteve considerando carga e peso.
   Qual a vantagem que você obteve (se existe alguma) com a equação considerando a idade? Exemplifique.

Questão 2

Considere agora que o modelo a ser usado é uma Gaussiana multivariada.

1. Use as colunas 2, 4 (carga e VO_2max). Neste caso $\mathbf{x}$ é bidimensional. Ache os parâmetros da Gaussiana. Plote os resultados. Os seu gráfico deve ser semelhante (na ideia) ao da Figura 7.1 (Murphy).
2. Use as colunas 2, 3 e 4 (peso, carga e VO_2max). Ache os parâmetros da Gaussiana. Escolha 3 valores de peso e carga e estime o VO_2max. (Como a variável VO_2max é contínua, explique como você faria essa estimativa.) Compare com os resultados da equação do *American College of Sports Medicine* com os resultados deste modelo e da questão anterior. Explique.
3. Use toda a informação dada, isto é, colunas 1, 2, 3 e 4 (idade, peso, carga e VO_2max), ache os parâmetros do modelo, explicando o que foi feito para achar os parâmetros.
   Nota: "o que foi feito" não significa "incluir código" na resposta, mas sim mostrar que você conhece os passos para se obter a solução.
   Esse seu modelo faz uma melhor ou pior previsão dos dados do teste? Compare o que você pode prever com esse e o modelo da questão anterior.

Questão 3

1. Suponha que você quer obter modelos para 3 faixas de idade: $id_1 = [18 - 40), id_2 = [40 - 60), id_3 = \geq 60$.
   Construa 3 modelos Gaussianos diferentes (multivariados).
   Comente as diferenças entre os modelos, se houver, com o objetivo de prever o VO_2max).
2. Suponha que o nosso objetivo é classificar os dados em uma das 3 faixas de idades do item anterior ($id_1 = [18 - 40), id_2 = [40 - 60), id_3 = \geq 60$). Mostre (explique) como você irá construir um *Naive Bayes Classifier* baseado no modelo Gaussiano, isto é, para uma *feature* j e faixa de idade c:
   $$p(\mathbf{x}_j | h = c, \boldsymbol{\theta}_{jc}) = \mathcal{N}(x_j, |y = c, \mu_{jc}, \sigma^2_{jc})$$
   como visto em classe. Comece escrevendo uma expressão para $p(\mathbf{x}_i, y_i = c_j | \boldsymbol{\theta})$, onde $\mathbf{x}_i$ é o vetor de *features* (no caso, peso, carga e VO_2max) e a classe é a idade.
3. Quais a suposição básica feita para um *Naive Bayes Classifier*? Essa suposição é feita no caso do modelo Gaussiano multivariado da Questão 2?
4. Mostre como calcular os parâmetros do modelo *Naive Bayes Classifier*. É parte da questão obter o MLE, o que vai permitir obter os parâmetros do seu modelo.
5. Construa um *Naive Bayes Classifier* com 3 classes de idade conforme acima, e treine o seu modelo com 1000 amostras escolhidas aleatoriamente dos dados desta lista. Você consegue prever adequadamente a faixa etária do restante dos dados? Caso positivo, Como?
6. Você conseguiria prever adequadamente a faixa etária, usando os 3 modelos Gaussianos construídos no início da questão? Sim? Não? Comente.
7. Você consegue prever a faixa etária dos dados de teste usando o seu modelo Gaussiano da questão 2 construído com todas as *features*? Sim? Não? Comente.

Questão 4

O objetivo desta questão é usar construir um modelos supervisionado com uma mistura de 3 Gaussianas, usando os dados do problema.

1. Neste modelo, todas as variáveis são observadas. Considere que os dados são da forma $\langle \mathbf{x}_i, z_i \rangle$ onde $\mathbf{x}_i$ tem dimensão 3 (peso, carga, VO_2max) e z_i é uma das 3 faixas de idade $id_1 = [18 - 40), id_2 = [40 - 60), id_3 = \geq 60$.
2. Explique, em linhas gerais, os passos feitos para encontrar os parâmetros do modelo. Como anteriormente, escolha o conjunto de treinamento de 1000 dados e outro para testes.
3. Para cada um das amostras i de teste, estime a probabilidade de se obter o VO_2max. (Como as variável VO_2max é contínua, explique como você faria essa estimativa.)
4. Compare o modelo obtido nesta questão com o modelo obtido na questão anterior, É parte da questão justificar sua escolha de como comparar os resultados/modelos.
5. Tanto o modelo *Naive Bayes* como o modelo de mistura de Gaussianas são generativos. A partir do melhor modelo obtido, explique como gerar e gere um conjunto de 10 valores de VO_2max, de um determinado conjunto de *features* escolhidas de um dos dados de teste. Compare os 10 valores gerados com o valor de VO_2max do dado de teste escolhido.

Quarta Lista

Questão 5

O objetivo desta questão é construir um modelo não supervisionado, usando uma mistura de Gaussianas. e estudar o algoritmo EM (Expectation Maximization).

Suponha que a variável idade **não** é observada, diferentemente do problema anterior, e você quer verificar se os agrupamentos encontrados formam faixas etárias. Neste caso, suponha que o número de classes seja igual ao número de faixas etárias, e você **não sabe** nada sobre o número de faixas. Note que você não pode usar os dados de idade fornecidos. Portanto, é parte do trabalho a escolha do número de classes a ser usado. (Sugestão: experimente com 2, 3, etc.)

1. Explique, em linhas gerais, os passos feitos para encontrar a solução do modelo incluindo as suas escolhas para usar o algoritmo EM.
2. A partir do modelo obtido, explique como o algoritmo pode agrupar as amostras. É parte da solução explicar como você vai classificar cada amostra $\mathbf{x}_i$ a partir do modelo obtido.
   Os agrupamentos obtidos fazem algum sentido?

É importante justificar suas escolhas, e explicar o desenvolvimento. Não complique!

Questão 6

O objetivo desta questão é encontrar um bom modelo para o VO_2max, usando **todos** os dados fornecidos. Neste caso, $\mathbf{x}$ é multidimensional (idade, peso e carga e VO_2max). Nesta questão você é livre para escolher um modelo, **dentre os estudados no curso até o momento**. Objetivo:

- Dado uma determinada idade, peso e carga, qual seria a distribuição de probabilidade do VO_2max que o modelo fornece.

1. Justifique, em poucas palavras a escolha do modelo, e indique os parâmetros do mesmo.
2. Tente interpretar os resultados encontrados.
3. Escolha duas faixas etárias, por exemplo entre $[30-40)$ e $[50-60)$ anos. Escolha 2 valores de VO_2max possíveis para cada faixa. Você poderia calcular a probabilidade dos indivíduos em cada uma dessas faixas de idade. possuir o valor de VO_2max escolhido, a partir do seu modelo? Justifique.
4. Suponha um indivíduo com cujo $VO_2\text{max} = 32,6$ mL/kg.min, carga de 181 e cujo peso seja de $81,5$ Kg. A partir do modelo, obtenha as probabilidade do indivíduo ter entre $[40-50)$ anos, $(50-60)$ anos e $[60-70]$ anos.

Questão 7

O objetivo desta questão é verificar se é possível encontrar *clusters* de pontos com alguma interpretação que faça sentido. Use os dados de idade, carga e VO_2max de todos os pacientes, e K-means para achar os *clusters*.

1. Para $K = 3$: para cada um dos *clusters*, calcule a fração de pontos do cluster que estão associados a cada uma das seguintes faixas etárias: $[18-30), [30-50), [50-60), [60-70), [70-80), [80-100)$. Por exemplo, para o caso trivial $K = 1$, $68/1172 = 0.0580$ dos pontos deste *cluster* único são associados a faixa $[18-30)$ e $445/1172 = 0.3797$ dos pontos estão associados à faixa $[30-50)$.

2. Repita para Para $K = 4$.
3. Tente perceber alguma interpretação para os *clusters* encontrados, se é que existe alguma.
4. Comente os resultados.

Referências

[1] Christina G. de Souza e Silva and Claudio Gil S. Araujo. Sex-Specific Equations to Estimate Maximum Oxygen Uptake in Cycle Ergometry. *Arquivos Brasileiros de Cardiologia*, 2015.