Machine Learning
CPS 863
Terceiro Trimestre de 2021

Professores:
Edmundo de Souza e Silva, Rosa Leão
Monitor: Gustavo Santos.

Lista 1

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

Os objetivos desta lista e as várias questões subsequentes são: (a) fazer com que você exercite a teoria vista em sala; (b) fazer com que você se familiarize com diferentes modelos (simples); (c) comparar alguns 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. Procure implementar as equações para aprender a teoria. Pode comparar com *libraries* já prontas, mas tente implementar você mesmo antes de usar o que está pronto.

Utilizamos, nessa lista, um dataset contendo o preço de um conjunto de imóveis e seus respectivos atributos (dataset completo em: https://www.kaggle.com/c/house-prices-advanced-regression-techniques) com o objetivo de criar modelos capazes de prever o preço dos imóveis. Para facilitar o problema consideramos um subconjunto de *features* pequeno e fizemos um pré-processamento dos dados, aplicando a escala logarítmica (logaritmo natural) a todas as variáveis consideradas e realizando filtragens. Os dados para treinamento dos modelos estão disponíveis em arquivo separado (*lista1-dados.csv*). O arquivo CSV com os dados preprocessados contém as seguintes variáveis:

- *GroundLivingArea*: Área total de cômodos presentes no primeiro andar (sala de jantar, sala de estar, escritórios, etc...) medida em *square feet* (em escala logarítmica)
- *BasementArea*: Área total do porão da casa medida em square feet (em escala logarítmica)
- *SalePrice*: Valor da propriedade em dólares (em escala logarítmica). Esta variável representa o que desejamos prever com os nossos modelos

IMPORTANTE: Antes de realizar as questões você deve escolher aleatoriamente um subconjunto de 80% das amostras para realizar o treinamento dos modelos, enquanto o subconjunto com 20% de amostras restantes deve ser utilizado para testar o modelo. Indique, no início do seu relatório, o vetor de índices das amostras escolhidas para treinamento e para teste. (Por exemplo: vetor de treinamento $\mathbf{t} = \{1, 0, 0, 1, 1, 1, 0, ...\}$, onde $t(i) = 1$ se a i-ésima amostra foi escolhida para treinamento, e $t(i) = 0$ para teste).

Questão 1

Nesta questão usaremos como modelo regressão linear.

1. Suponha que seus dados tenham apenas uma única *feature*. Suponha ainda que você escolheu $d = 1$ para o modelo, i.e., $\mathbf{w}^T\mathbf{x} = [w_0, w_1 x]$. Mostre, e explique, os passos necessários para calcular os parâmetros w_0 e w_1 usando conceitos de MLE, conforme descrito em sala.
2. (a) Encontre os parâmetros de uma regressão linear que considera como entrada apenas a variável *GroundLivingArea* e tem como objetivo prever o valor da variável *SalePrice*. Neste caso d não é dado, e faz parte do problema discutir a sua escolha.

- Para o caso de $d = 1$ ilustre a equação que você obteve no item anterior com esses dados.
- 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?

(b) Encontre os parâmetros de uma regressão linear que considera como entrada a variável *BasementArea* e tem como objetivo prever o valor da variável *SalePrice* Neste caso também d não é fornecido, e faz parte do problema discutir a sua escolha.

(c) Compare os modelos obtidos nas duas questões anteriores utilizando o conjunto de teste. Qual modelo você escolheria para prever o preço? Justifique a sua resposta.

(d) Encontre os parâmetros de uma regressão linear que considera como entrada conjuntamente as variáveis *GroundLivingArea* e *BasementArea* e tem como objetivo prever o valor da variável *SalePrice*. Você considera este modelo melhor que os modelos obtidos nos itens acima? Justifique.

(e) Calcule, utilizando o melhor modelo obtido, preveja o preço do imóvel quando *GroundLivingArea* = 8 (em escala logarítmica) e *BasementArea* = 7.5 (também em escala logarítmica).

Questão 2

Considere uma distribuição Normal $\mathcal{N}(\mu, \sigma^2)$ e N amostras ($\mathcal{D} = \{x_1, x_2, \ldots, x_N\}$). (Cada amostra tem uma *feature* apenas.)

1. Qual é a *likelihood function* $\mathcal{L}(\theta, \mathcal{D})$ neste caso? Lembre que θ é um vector de parâmetros. Quantos elementos o vetor θ tem?
2. Mostre **todos** os passos para se obter o MLE neste caso simples. Compare seu resultado e a sua prova com aquele no Teorema 4.1.1 do Murphy(2012).

Questão 3

(a) Encontre os parâmetros de uma Gaussiana univariada para cada uma das variáveis de interesse separadamente (*GroundLivingArea*, *BasementArea*, *SalePrice*).
Faz parte do item mostrar todos os passos para encontrar a fórmula que você usou. É essencial que você esteja familiarizado com esses passos.

(b) Encontre os parâmetros de uma Gaussiana de duas dimensões utilizando as variáveis *GroundLivingArea* e *SalePrice*. Note que, neste caso, $\mathbf{x}$ é bidimensional.
Plote os resultados.
Os seu gráfico deve ser semelhante (na ideia) ao da Figura 7.1 (Murphy 2012).
Utilize este modelo para prever o valor esperado de *SalePrice* quando *GroundLivingArea* = 8 (em escala logarítmica).

(c) Encontre os parâmetros de uma gaussiana de duas dimensões utilizando as variáveis *BasementArea* e *SalePrice*. Utilize este modelo para prever o valor esperado de *SalePrice* quando *BasementArea* = 7.5 (em escala logarítmica)

(d) Encontre os parâmetros de uma gaussiana de três dimensões utilizando as variáveis *GroundLivingArea*, *BasementArea* e *SalePrice*. Utilize este modelo para prever o valor esperado de *SalePrice* quando *GroundLivingArea* = 8 e *BasementArea* = 7.5 (em escala logarítmica).

(e) Utilizando o modelo do item anterior calcule a probabilidade de que o preço do imóvel esteja entre 300000 e 500000 dólares quando $GroundLivingArea = 8$ (em escala logarítmica) e $BasementArea = 7.5$ (em escala logarítmica). Lembre-se que o seu modelo foi criado utilizando variáveis em escala logarítmica.

(f) Esse seu modelo faz uma melhor ou pior previsão dos dados do teste em relação ao modelo de regressão? Compare o que você pode prever com esse e o modelo de regressão?