UDACITY

Sair

‹ Return to "Machine Learning Engineer Nanodegree" in the classroom

Creating Customer Segments

REVISÃO | HISTORY

Meets Specifications

Olá, aqui é o Cláudio! Obrigado por enviar todos os arquivos necessários para o processo de revisão e também todo o código executando sem erros.
Parabéns pelo ótimo trabalho realizado e pela qualidade exibida. Você parece ter dominado os conceitos e ainda aplicou recursos extras para indagação das suas idéias e comparações importantes. Você realmente fez um excelente trabalho. Parabéns.

Caso tenha ficado alguma dúvida, por favor, me envie um email, e estarei feliz em ajudar.

Espero que as sugestões e bonus colocados no seu projeto façam sentido para você.

Espero que tenha curtido fazer este projeto e ter colocado em prática conceitos importantes de machine learning. Deixarei meus contatos abaixo caso tenha qualquer dúvida referente a esta revisão e também para nos mantermos conectados.

É isto, divirta-se aprendendo machine learning.

Por fim, queria compartilhar uma ferramenta interessante do Google que ajuda os engenheiros de aprendizado de máquina a entender os dados muito rapidamente e decidir qual tipo de algoritmo se ajustará melhor a esses dados: Facets - Visualizations (https: // pair -code.github.io/facets/). Definitivamente, confira.

O poder do aprendizado de máquina vem de sua capacidade de aprender padrões a partir de grandes quantidades de dados. Entender seus dados é fundamental para construir um sistema poderoso de aprendizado de máquina.

O Facets contém duas visualizações robustas para auxiliar na compreensão e análise de conjuntos de dados de aprendizado de máquina. Tenha uma ideia da forma de cada recurso do conjunto de dados usando a Visão geral de facetas ou explore observações individuais usando o Facets Dive.

1_bGQMgbCspRFqfcMJM63D_g.gif

Obrigado.
Cláudio

email: cgimenest@uol.com.br
Linkedin: https://www.linkedin.com/in/claudiogimenestoledo/

Exploração dos dados

Três amostras diferentes dos dados são escolhidas, e o que elas representam é proposto com base na descrição estatística dos dados.

Excelente trabalho selecionando três tipos de dados em amostras diferentes:

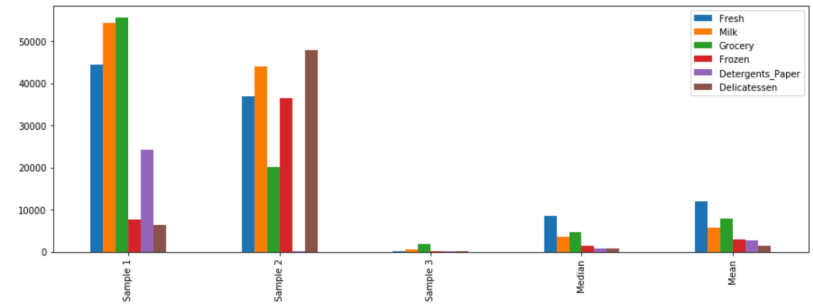
> indices = [43, 202, 402]

Este é um passo muito importante durante a fase de exploração dos dados, no qual nos dá grandes e importantes insights sobre o o dataset para próximos passos.

Sugestões:

- Você pode também plotar estes exemplos de dados e extrair visualizações sobre estatísticas deles, exemplo:

```
import matplotlib.pyplot as plt
import seaborn as sns

samples_for_plot = samples.copy()
samples_for_plot.loc[3] = data.median()
samples_for_plot.loc[4] = data.mean()

labels = ['Sample 1', 'Sample 2', 'Sample 3', 'Median', 'Mean']
samples_for_plot.plot(kind='bar', figsize=(15, 5))
plt.xticks(range(5), labels)
plt.show()
```



A pontuação do atributo removido foi corretamente calculada. A resposta justifica se o atributo removido é relevante.

Ótimo trabalho.

> Grocery e Detergents_Paper possuem pontuação positiva (aproximadamente 0.60). Milk quase não tem pontuação (aproximadamente 0.10). Os outros três não são previsíveis (pontuações negativas).

Atributos correlacionados são corretamente identificados e comparados com o atributo previsto. A distribuição dos dados para esses atributos é discutida.

Muito bom. As correlações estão devidamente indicadas.

> Grocery e Detergents_Paper possuem pontuação positiva (aproximadamente 0.60). Milk quase não tem pontuação (aproximadamente 0.10). Os outros três não são previsíveis (pontuações negativas).

Pré-processamento dos dados

Os valores aberrantes extremos são identificados, e discute-se se eles deveriam ser removidos. A decisão de remover quaisquer dados é corretamente justificada.

Perfeito.

Os valores aberrantes foram devidamente selecionados:

Bonus:

- Aqui deixo em inglês um artigo interessante sobre a estratégia do que se fazer com outliers:
  http://www.theanalysisfactor.com/outliers-to-drop-or-not-to-drop/

O código de dimensionamento de atributos, tanto para os dados como para as amostras, foi corretamente implementado.

Excelente.

Está parte do código foi devidamente implementada conforme esperado:

```
TODO: Escalone os dados utilizando o algoritmo natural
log_data = np.log(data)
```

```
TODO: Escalone a amostra de dados utilizando o algoritmo natural
log_samples = np.log(samples)
```

Bonus:

- Aqui deixo alguns bons artigos sobre feature scaling, em inglês para sua referência:
  http://sebastianraschka.com/Articles/2014_about_feature_scaling.html
  https://stackoverflow.com/questions/26225344/why-feature-scaling
  https://www.quora.com/Why-do-we-use-standardization-for-feature-scaling-for-machine-learning-algorithms

Transformação de atributos

A variância explicada total para duas e quatro dimensões dos dados do PCA é corretamente relatada. As primeiras quatro dimensões são interpretadas como uma representação dos gastos do cliente com justificativa.

Perfeito.

Você explicou e exibiu corretamente os valores acumulados das variâncias:

> A variância cumulativa explicada para duas e quatro dimensões é de cerca de 71% e 93%, respectivamente. Isso pode mudar para um pequeno percentual, com base nos outliers que são removidos.

O código do PCA foi corretamente implementado e aplicado, tanto para os dados dimensionados como para as amostras dimensionadas, no caso bidimensional.

Bom trabalho.

Devidamente implementado.

Bonus:

- Aqui deixo alguns artigos interessantes sobre o tema análise PCA e porquê é importante:
  https://towardsdatascience.com/dimensionality-reduction-does-pca-really-improve-classification-outcome-6e9ba21f0a32
  https://hackernoon.com/supervised-machine-learning-dimensional-reduction-and-principal-component-analysis-614dec1f6b4c
  https://machinelearningmastery.com/calculate-principal-component-analysis-scratch-python/

Clustering

Os algoritmos GMM e k-means são comparados em detalhes. A escolha do aluno é justificada com base nas características do algoritmo e dos dados.

Muito bom!

Sugestões:

- Sempre é uma boa idéia adicionar links para referências, imagens, estudo de caso para que transmita uma mensagem mais clara e consistente com aquilo que está sendo discutido. Aqui deixo algumas sugestões de artigos para dar uma olhada, em inglês:
  http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/mixture.html

http://www.nickgillian.com/wiki/pmwiki.php/GRT/GMMClassifier
http://playwidtech.blogspot.hk/2013/02/k-means-clustering-advantages-and.html
http://www.improvedoutcomes.com/docs/WebSiteDocs/Clustering/K-Means_Clustering_Overview.htm
http://stats.stackexchange.com/questions/133656/how-to-understand-the-drawbacks-of-k-means
http://www.r-bloggers.com/k-means-clustering-is-not-a-free-lunch/
http://www.r-bloggers.com/pca-and-k-means-clustering-of-delta-aircraft/
https://shapeofdata.wordpress.com/2013/07/30/k-means/
http://mlg.eng.cam.ac.uk/tutorials/06/cb.pdf

Amostras dos dados são corretamente relacionadas aos segmentos da clientela, e o grupo a que pertence cada ponto da amostra é discutido.

Um ponto interessante a ser observado no gráfico cluster_visualization é que os dois clusters são essencialmente separados por um valor na primeira dimensão do PCA, o que vimos anteriormente é predominantemente uma combinação de Detergents_Paper, Grocery e Milk. O resto dos recursos, que figuram proeminentemente apenas na segunda dimensão do PCA, não importa realmente!

Diversas pontuações são corretamente relatadas, e o número ótimo de grupos é escolhido com base na melhor. A visualização escolhida mostra o número ótimo de grupos baseado no algoritmo de clustering escolhido.

Excelente trabalho. Você ainda exibiu um gráfico atribuido a diferentes valores obtidos, parabéns.

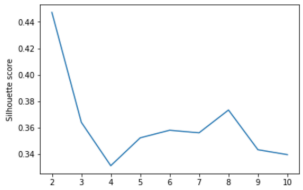
Você também escolheu corretamente o melhor número de clusters.

> Dois clusters quase sempre darão a mesma pontuação Silhouette, de cerca de0.42 (dependendo de quais outliers são removidos).

Sugestão:

- Você também pode exibir um gráfico com valores para cada um dos resultados obtidos, gestão visual facilita o processo de tomada de decisão:

In [19]:
```
from sklearn.cluster import KMeans
from sklearn.mixture import GaussianMixture
from sklearn.metrics import silhouette_score

# TODO: Aplique o algoritmo de clustering de sua escolha aos dados reduzidos
clusters_scores = []
for clusters in np.arange(2,11):

    clusterer = KMeans(n_clusters=clusters).fit(reduced_data)

    # TODO: Preveja o cluster para cada ponto de dado
    preds = clusterer.predict(reduced_data)

    # TODO: Ache os centros do cluster
    centers = clusterer.cluster_centers_

    # TODO: Preveja o cluster para cada amostra de pontos de dado transformados
    sample_preds = clusterer.predict(pca_samples)

    # TODO: Calcule a média do coeficiente de silhueta para o número de clusters escolhidos
    score = silhouette_score(reduced_data, preds)

    #Appending to the list
    clusters_scores.append(score)

    #Setting for the best score
    best_cluster = clusters_scores.index(max(clusters_scores)) + 2
    clusterer = KMeans(n_clusters=best_cluster).fit(reduced_data)
    preds = clusterer.predict(reduced_data)
    sample_preds = clusterer.predict(pca_samples)
    centers = clusterer.cluster_centers_

#ploting the graph of silhouette scores

plt.plot(np.arange(2,11),clusters_scores)
plt.xlabel("Quantity of clusters")
plt.ylabel("Silhouette score")
plt.show()

print("The best cluster number is {} with a silhoute score of {}".format(best_cluster, max(clusters_scores)))
```



Os grupos representados por cada segmento da clientela são propostos com base na descrição estatística do conjunto de dados. Os códigos de transformação e dimensionamento inversos foi corretamente implementado e aplicado para o centro dos grupos.

Muito bem implementado.

```
# TODO: Transforme os bons dados utilizando o ajuste do PCA acima
reduced_data = pca.transform(good_data)

# TODO: Transforme a amostre de log-data utilizando o ajuste de PCA acima
pca_samples = pca.transform(log_samples)
```

Conclusão

O aluno identifica corretamente como um teste A/B pode ser feito com a clientela após uma mudança no serviço de distribuição.

Excelente.

Gostei das explicações exibidas.

Bonus:

- Deixo aqui algumas referências em inglês sobre o assunto e como se complementam com IA:
  https://hackernoon.com/ai-as-complement-to-a-b-test-design-e8f4b5e28d92
  https://www.dynamicyield.com/ab-testing/
  https://www.mediapost.com/publications/article/305837/ab-testing-vs-ai-conversions.html

O aluno discute e justifica como os dados de clustering podem ser usados em um modelo

de aprendizagem supervisionada para fazer novas estimativas.

Muito bom.

> Os rótulos 'customer segment' podem ser usados como feature de input adicional, que uma pessoa que está usando aprendizagem supervisionada por treinar e, então, fazer previsões sobre os novos clientes.

Os segmentos da clientela e os dados em `Channel` são comparados. Os segmentos identificados pelos dados de `Channel` são discutidos, inclusive se essa representação é consistente com resultados anteriores.

Muito bem implementado e o gráfico gerado com perfeição.

BAIXAR PROJETO

RETORNAR