DATA MINING & MACHINE LEARNING (I)

Thiago Marzagão

detecção de anomalias

- Dado um conjunto de amostras, queremos saber quais são "diferentes".

detecção de anomalias

- Dado um conjunto de amostras, queremos saber quais são "diferentes".
- Exemplo: banco te liga à meia-noite p/ confirmar que foi você mesmo que acabou de comprar uma passagem p/ o Havaí no site da American Airlines.

detecção de anomalias

- Dado um conjunto de amostras, queremos saber quais são "diferentes".
- Exemplo: banco te liga à meia-noite p/ confirmar que foi você mesmo que acabou de comprar uma passagem p/ o Havaí no site da American Airlines.
- Possíveis "esquisitices" nessa compra: você geralmente não compra em sites estrangeiros; você geralmente não compra de madrugada; etc. O banco achou essa compra "diferente" e por isso alguém te ligou p/ confirmar.

detecção de anomalias

- Dado um conjunto de amostras, queremos saber quais são "diferentes".
- Exemplo: banco te liga à meia-noite p/ confirmar que foi você mesmo que acabou de comprar uma passagem p/ o Havaí no site da American Airlines.
- Possíveis "esquisitices" nessa compra: você geralmente não compra em sites estrangeiros; você geralmente não compra de madrugada; etc. O banco achou essa compra "diferente" e por isso alguém te ligou p/ confirmar.
- O banco não faz essa análise manualmente: seria impossível olhar cada transação. Aí é que entram as técnicas de detecção de anomalia, que vamos ver hoje.

detecção de anomalias

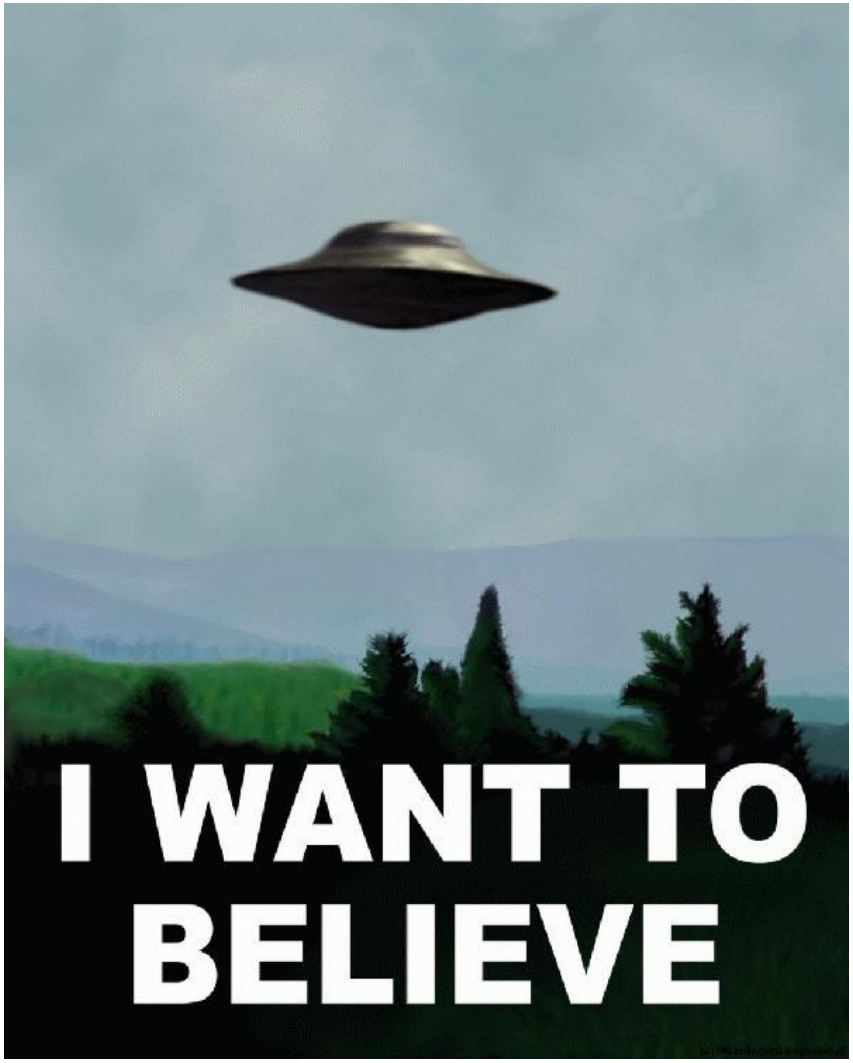
detecção de anomalias

- Outro exemplo: identificação de alienígenas com base em sinais de rádio.

detecção de anomalias

- Outro exemplo: identificação de alienígenas com base em sinais de rádio.
- https://www.seti.org/seti-institute/weeky-lecture/anomaly-detection-data-streams-and-its-implications-radio-astronomy-and

detecção de anomalias



detecção de anomalias

- Vários outros usos!

detecção de anomalias

- Ok, como fazer?

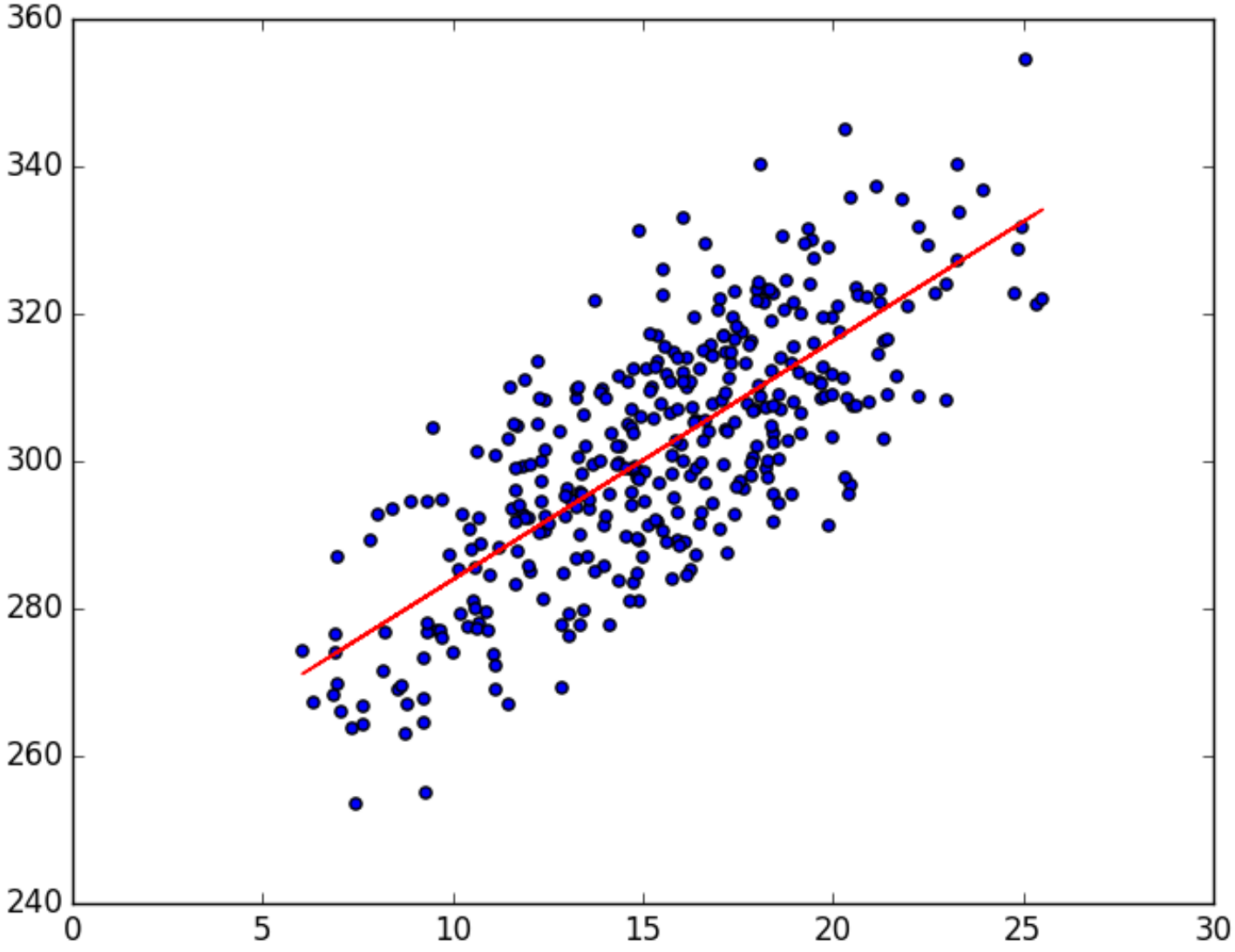
detecção de anomalias

- Ok, como fazer?
- Várias possibilidades: regressão, clusterização, kNN, LOF, autoencoders (não vamos ver tudo).

detecção de anomalias

- Ok, como fazer?
- Várias possibilidades: regressão, clusterização, kNN, LOF, autoencoders (não vamos ver tudo).
- Vamos começar com regressão, que vocês já viram com o Carlos Góes.

detecção de anomalias



usando regressão p/ detectar anomalias

- Os pontos azuis nos dão o valor observado de cada amostra: $y = \alpha + \beta X + \epsilon$

usando regressão p/ detectar anomalias

- Os pontos azuis nos dão o valor observado de cada amostra: $y = \alpha + \beta X + \epsilon$
- A linha vermelha nos dá as previsões do modelo p/ cada amostra: $\hat{y} = \alpha + \beta X$

usando regressão p/ detectar anomalias

- Os pontos azuis nos dão o valor observado de cada amostra: $y = \alpha + \beta X + \epsilon$
- A linha vermelha nos dá as previsões do modelo p/ cada amostra: $\hat{y} = \alpha + \beta X$
- A diferença entre o valor previsto e o valor observado nos dá o erro do modelo: $\epsilon = y - \hat{y}$

usando regressão p/ detectar anomalias

- Os pontos azuis nos dão o valor observado de cada amostra: $y = \alpha + \beta X + \epsilon$
- A linha vermelha nos dá as previsões do modelo p/ cada amostra: $\hat{y} = \alpha + \beta X$
- A diferença entre o valor previsto e o valor observado nos dá o erro do modelo: $\epsilon = y - \hat{y}$
- Quanto maior o ϵ, mais "esquisita" é a amostra.

usando regressão p/ detectar anomalias

- Os pontos azuis nos dão o valor observado de cada amostra: $y = \alpha + \beta X + \epsilon$
- A linha vermelha nos dá as previsões do modelo p/ cada amostra: $\hat{y} = \alpha + \beta X$
- A diferença entre o valor previsto e o valor observado nos dá o erro do modelo: $\epsilon = y - \hat{y}$
- Quanto maior o ϵ, mais "esquisita" é a amostra.
- Olhando os dados você pode tentar identificar a partir de qual ϵ uma amostra pode ser considerada anômala.

usando regressão p/ detectar anomalias

- Os pontos azuis nos dão o valor observado de cada amostra: $y = \alpha + \beta X + \epsilon$
- A linha vermelha nos dá as previsões do modelo p/ cada amostra: $\hat{y} = \alpha + \beta X$
- A diferença entre o valor previsto e o valor observado nos dá o erro do modelo: $\epsilon = y - \hat{y}$
- Quanto maior o ϵ, mais "esquisita" é a amostra.
- Olhando os dados você pode tentar identificar a partir de qual ϵ uma amostra pode ser considerada anômala.
- Mas notem: só faz sentido quando você tem um y, o que nem sempre é o caso.

usando clusterização p/ detectar anomalias

- Nas últimas aulas nós vimos clusterização.

usando clusterização p/ detectar anomalias

- Nas últimas aulas nós vimos clusterização.
- Na clusterização nós tentamos agrupar as amostras semelhantes entre si em clusters.

usando clusterização p/ detectar anomalias

- Nas últimas aulas nós vimos clusterização.
- Na clusterização nós tentamos agrupar as amostras semelhantes entre si em clusters.
- Mas algumas amostras podem não se encaixar bem em nenhum cluster. Elas são diferentes demais.

usando clusterização p/ detectar anomalias

- Nas últimas aulas nós vimos clusterização.
- Na clusterização nós tentamos agrupar as amostras semelhantes entre si em clusters.
- Mas algumas amostras podem não se encaixar bem em nenhum cluster. Elas são diferentes demais.
- O k-means força cada amostra a pertencer a um cluster. Mas outros algoritmos de clusterização, como o DBSCAN, permitem que amostras "diferentes demais" fiquem sem cluster.

usando clusterização p/ detectar anomalias

- Nas últimas aulas nós vimos clusterização.
- Na clusterização nós tentamos agrupar as amostras semelhantes entre si em clusters.
- Mas algumas amostras podem não se encaixar bem em nenhum cluster. Elas são diferentes demais.
- O k-means força cada amostra a pertencer a um cluster. Mas outros algoritmos de clusterização, como o DBSCAN, permitem que amostras "diferentes demais" fiquem sem cluster.
- Essas amostras são possíveis anomalias!

usando clusterização p/ detectar anomalias

- Nas últimas aulas nós vimos clusterização.
- Na clusterização nós tentamos agrupar as amostras semelhantes entre si em clusters.
- Mas algumas amostras podem não se encaixar bem em nenhum cluster. Elas são diferentes demais.
- O k-means força cada amostra a pertencer a um cluster. Mas outros algoritmos de clusterização, como o DBSCAN, permitem que amostras "diferentes demais" fiquem sem cluster.
- Essas amostras são possíveis anomalias!
- Outra possibilidade: lembram da silhueta? Ela mede quão bom ou ruim é o "encaixe" de cada amostra no cluster. Podemos tentar identificar quais valores de silhueta começam a nos dar amostras anômalas.

usando clusterização p/ detectar anomalias

- Nas últimas aulas nós vimos clusterização.
- Na clusterização nós tentamos agrupar as amostras semelhantes entre si em clusters.
- Mas algumas amostras podem não se encaixar bem em nenhum cluster. Elas são diferentes demais.
- O k-means força cada amostra a pertencer a um cluster. Mas outros algoritmos de clusterização, como o DBSCAN, permitem que amostras "diferentes demais" fiquem sem cluster.
- Essas amostras são possíveis anomalias!
- Outra possibilidade: lembram da silhueta? Ela mede quão bom ou ruim é o "encaixe" de cada amostra no cluster. Podemos tentar identificar quais valores de silhueta começam a nos dar amostras anômalas.
- Útil quando você não tem um y.

detecção de anomalias

- Problema comum: dados sujos.

detecção de anomalias

- Problema comum: dados sujos.
- Se um determinado valor foi preenchido erroneamente isso pode resultar numa falsa anomalia.

detecção de anomalias

- Problema comum: dados sujos.
- Se um determinado valor foi preenchido erroneamente isso pode resultar numa falsa anomalia.
- Importante limpar os dados antes.

detecção de anomalias

- Exercícios!

ex. 1: usar regressão p/ encontrar imóveis anômalos em Melbourne

ex. 2: usar clusterização p/ encontrar vinhos anômalos

ex. 3: usar regressão p/ encontrar vinhos anômalos (vocês sozinhos)