Detecção de anomalias e regiões centrais

Jacques Wainer

IC – Unicamp

Dezembro 2015

Anomalias e regiões centrais

- Anomalias: outliers, novidade, exceção - dados que não se conformam com o "padrão normal" dos dados. Uma definição circular, pois os padrões normais são os dados sem as anomalias!
- regiões centrais (não existe nome oficial) - regiões onde os dados são mais frequentes - e provavelmente "mais normais"
- se voce tiver que se restringir a poucos pontos, as regiões centrais são uma alternativa.

Regiões centrais

- regiões ce
- mean shift
- algoritmos baseados em Grid (cap 12.6) - CLIQUE, MAFIA, SHIFT
- não há uma separação clara entre os conceitos de cluster (nao completo) e região central - assim os algoritmos em Grid são normalmente descritos como algoritmos de clusterização (nao completo)

Mean shift

- Defina uma bola d-dimensional de raio ϵ
- para os dados dentro da bola, calcule a média, este é o novo centro da bola
- o algoritmo para quando a bola estiver centrada em uma região mais densa que as vizinhas - máximo local (similar ao k-means)

Algoritmos em Grid - CLIQUE

- divida cada dimensão i em n_i 1-células (1d-células) normalmente equi-espacadas
- haverá $n = \Pi_i^d n_i$ d-dimensional células (hiper-retângulos).
- defina qual o numero de dados X em uma d-célula que indica que há uma concentração de dados (média por d-célula é N/n.
- para cada dimensão selecione as 1-células que contem pelo menos X dados, para pares de 1-células, selecione as 2-células que contem pelo menos X dados, e assim por diante.
- no final voce terá um conjunto de d-células cada uma com pelo menos X dados. Normalmente junta-se as d-células que são adjacentes.
- variações do algoritmos trabalham com células de multiplas resoluções
- não sei como escolher o X.

Anomalias

- anomalias, outliers, exceções, erros, novidades
- Anomalias são dados que "quebram" / não se conformam com o padrão que descreve os dados "normais," não anômalos.
- Historicamente, anomalias eram erros dos dados, a serem eliminados das análises.
- modernamente, detecção de anomalias é considerado como uma "boa" forma de detectar fraudes, ataques e intrusão, mudança de padrão de uso de recursos, surgimento de epidemias, novidades.

Anomalias

- Há várias definições de anomalias e portanto varios algoritmos
- normalmente assume-se que anomalias são infrequentes

1. sao dados que tem baixa probabilidade de sair dado a distribuição geradora dos dados "normais"
2. são dados "longes" do dados normais (mas talvez próximos um dos outros)
3. são dados em regiões de baixa densidade em comparação com dados normais
4. dados que se encontram em sub-espaços diferentes de onde se encontram os dados normais
5. dados que "estragam" a descrição sucinta dos dados

Anomalias globais e locais

- As definições mais comuns se referem a anomalias globais
- mas é possível falar em anomalias locais - um dado é anomalo em relação ao dados "normais" mais perto dele
- isso é mais claro para dados em uma serie temporal - um dado é localmente anomalo numa serie temporal se ele representa um salto inesperado tendo em vista os dados locais

Anomalias em uma dimensão

- velha regra: anomalias são dados a uma distancia $> 3\sigma$ da média dos dados
- 1o problema: a regra assume que os dados "normais" estão distribuidos segundo uma gaussiana ($P(x > 3\sigma) < 0.003$)
- isso pode ser modificado para usar outras distribuições
- ou pose-se usar relações (inequalidades) que não dependem das distribuições (Chebyshev, Hoeffding)

Anomalias em uma dimensão

- 2o problema: o calculo da média e do desvio padrão incluem os dados anômalos
- estatística robusta - estatística que são insensíveis a anomalias
- ex: windowed mean: remova os x% maiores e menores dados antes de tirar a média (e o desvio padrão)
- ex: mediana em vez da media e MAD (median absolute deviation) em vez do desvio padrão

Anomalias por distancia

- Anomalias estão isoladas:
- meça a distancia de cada ponto ao seu vizinho mais proximo - anomalias serão outliers nessa medida 1D!
- Problema: anomalias podem destar perto de outras anomalias:
- meça a distancia de cada ponto a seu k-esimo vizinho mais proximo (ja usamos isso no DBScan!)
- O procedimento trata de casos de anomalias que estão próximas de ate k outras anomalias
- Essa tecnica não deve funcionar em dimensões muito altas (maldição da dimensionalidade - *curse of dimentionality*

Anomalias por densidade

- DBScan - dados anômalos são que não estão na vizinhança de dados densos e eles proprios não tem MinPts vizinhos num raio Eps.
- Local Outlier Factor: densidade = media das distancias aos k vizinhos mais próximos.
- LOF = 1/densidade ou LOF = densidade/densidade dos k vizinhos

Anomalias por subespaço (linear)

- Calcule o PCA, e fique apenas com as dimensões importantes
- reconverta os dados projetados para as dimensões originais
- compute a distancia (nas dimensões originais) dos dados e de suas projeções reconvertidas.
- as anomalias devem ser outliers nessa distribuição de distancias

Outras técnicas

- classificadores de uma classe (*one class SVM* - descobre a menor hiper-bola que inclue a maioria dos dados (miolo: normais, fora: anomalias)
- redes neurais que aprendem a repetir os dados (auotencoders) - equivale a deteção de anomalias por subespaço não linear
- aplicação sequencial de vários detetores: 1o algoritmos remove as anomalias mais obvias, 2o só usa os dados que sobraram e busca anomalias menos óbvias, e assim por diante
- aplicação paralela de vários detetores: cada detetor retorna um grau de anomalia para cada dado, os graus são combinados.