Parte I - Detecção de Anomalia

Esta parte versará sobre o desenvolvimento de um sistema de detecção de anomalia em servidores computacionais usando o modelo Gaussiano multivariado.

O arquivo *dado1.mat* contém um conjunto de exemplos de servidores (matriz Y), cada um com sua respectiva latência (tempo que um pacote específico leva para chegar ao destino) na 1ª coluna e taxa de transferência (quantidade de dados transferidos de um lugar para outro) na 2ª coluna. Estas são medidas sabidamente úteis para se determinar se um servidor está se comportando normalmente. No dado em questão, temos uma grande maioria de servidores "normais" e alguns poucos "anômalos".

(a) Mostre em um gráfico de Latência *vs.* Taxa de Transferência os pontos correspondentes aos servidores na matriz Y.

(b) Implemente uma rotina que ajusta uma Gaussiana multivariada aos dados.

(c) No caso, como só temos dois atributos, mostre curvas de contorno da Gaussiana obtida sobrepostas aos pontos do item (a), de forma que mesmo visualmente seja possível identificar as anomalias.

(d) Uma anomalia se caracteriza por uma baixa probabilidade, menor do que um *threshold* ϵ, de aquele dado pertencer à distribuição Gaussiana obtida. Obtenha o valor ideal de ϵ a partir do conjunto de validação (matriz $Xval$) e usando como medida de avaliação o F_1-*score*. Os rótulos correspondentes são dados por $yval$ e considera-se anomalia quando $y = 1$.

(e) Agora, com todos os elementos necessários para o seu sistema de detecção de anomalia implementados, use este sistema para encontrar e circular as anomalias na figura do item (c).

(f) Por fim, execute o sistema desenvolvido sobre um conjunto com mais servidores e mais atributos (arquivo *dado2.mat*). Obtenha os valores ótimos de ϵ e F_1-*score* usando $Xval$ e $yval$ do arquivo *dado2.mat*, assim como o número de anomalias detectadas.

Parte II - Sistemas de Recomendação

Nesta parte, você desenvolverá um sistema de recomendação de filmes usando filtragem colaborativa.

O arquivo *dado3.mat* contém notas de 1 a 5 dadas por usuários para filmes. A matriz Y armazena na linha i e coluna j a nota dada pelo usuário j para o filme i. Já em relação à matriz R, temos $R(i, j) = 1$ se o usuário j deu alguma nota para o filme i e 0 caso contrário.

(a) Implemente o algoritmo de filtragem colaborativa. Este deverá aprender uma matriz X, que em cada linha contém o vetor de atributos $x^{(i)}$ do i-ésimo filme e uma matriz Θ, que em cada linha guarda o vetor de parâmetros $\theta^{(j)}$ para o j-ésimo usuário. Considere que tanto $x^{(i)}$ quanto $\theta^{(j)}$ possuem dimensão 100. Considere a função de custo sem regularização e use gradiente conjugado para minimizá-la (o mesmo do Projeto 2).

(b) Com base nas notas preditas por seu algoritmo, liste os 10 filmes com notas médias mais altas, mostrando o nome e a nota média do respectivo filme. Para obter o nome do filme, use o arquivo *dado4.txt*, que contém o nome correspondente a cada linha na matriz Y.