Universidade Fedeal de Minas Gerais

Exercício 4 - Mistura de Gaussianas

A.P. Braga

April 26, 2017

Gaussian Mixture Model

O aluno deverá implementar um classificador utilizando o modelo de mistura de gaussianas para a base de dados Winsconsin Breast Cancer.
Considere a equação de Bayes:

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)} \tag{1}$$

Com esta equação é possível determinar a classe C_k de uma amostra x, de acordo com a probabilidade a priori de cada classe e da probabilidade a posteriori pertinência de cada amostra à pdf de cada classe. Como a probabilidade a a priori das amostras é constante, ela pode ser desconsiderada no classificador, dessa forma o classificador é implementado utilizando a seguinte equação:

$$p(C_k|x) = p(C_k)p(x|C_k) \tag{2}$$

A mistura de gaussianas deve ser utilizadas para definir as probabilidades a posteriori de cada amostra condicional a cada classe, para isto deve ser gerada uma mistura para cada classe utilizando o conjunto de treinamento, podem ser utilizadas as funções desenvolvidas em sala ou o pacote *mclust*

```
> library('mclust')
```

Para o cálculo das probabilidades a priori de cada classe, deve ser determinada a quantidade ocorrências desta classe no conjunto de treinamento em relação à quantidade total de amostras.

Winsconsin Breast Cancer

É possível utilizar a base de dados *Winsconsin Breast Cancer* da UCI através do pacode *mlbench*, que deve ser devidamente instalado no *RStudio*.

```
> library('mlbench')
```

Como a primeira coluna desta base de dados é um digito de identificação, ele pode ser descartado. Além disso nela existem alguns dados faltantes, assim, é sugerido que se substitua o valores não atribuídos (NA) por 0:

```
> data(BreastCancer)
> summary(BreastCancer)
> X <- data.matrix(BreastCancer[,2:10])
> X[is.na(X)] <- 0
> trainY <- as.numeric(BreastCancer$Class)
```

Como a base de dados está no formato de data frame, foi utilizados os comando *data.matrix* e *as.numeric* para transformá-la em numérica.

Treinamento e Teste

A base de dados deve ser divida em dois conjuntos, um de treinamento e outro de teste, de razão a critério do aluno.

- Com o grupo de testes, deve ser utilizada a mistura de gaussianas para determinar um modelo para cada classe e a probabilidade a priori da cada classe.
- O grupo de treinamento deve ser classificado de acordo com os modelos estimados no treinamento. Dado que a $p(x|C_k)$ para cada classe pode ser estimada a partir dos modelos de misturas de gaussianas estimados no treinamento, assim como as probabilidades a priori $p(C_k)$. A classe que apresentar o maior probabilidade $p(C_k|x)$ deve ser considerada como a classe estimada para a amostra.

Instruções

Neste exercício o aluno deverá:

1. Carregar os dados do pacote *mlbench* e substituir os dados faltantes por 0, por exemplo.
2. Dividir de forma aleatória os dados em grupos de treinamento de teste de acordo com uma razão pré definida.

3. Utilizar uma rotina ou função de treinamento que estime os modelos de mistura de gaussianas e as probabilidades a priori. Se for utilizado o pacote *mclust* a fução para mistura de gaussianas é:

   ```
   > model<-densityMclust(trainX)
   ```

4. Utilizar uma rotina função de teste que verifique pertinência de cada amostra a cada mistura de gaussianas e determine, de acordo com a regra de Bayes, a qual classe cada amostra pertence.

   ```
   > PxC<- dens(modelName=model$modelName, data = testX,
   +                    parameters = model$parameters)
   ```

5. Calcular o erro quadrático médio (MSE) percentual do classificador.

6. Repetir 10 vezes os procedimentos 2 ao 5 e estimar o MSE percentual médio e o desvio padrão do classificador.

Forma de Entrega

Relatório em .doc ou .pdf, descrevendo o que foi feito, mostrando os gráficos e as informações pedidas e explicando os resultados obtidos, assim como as partes importantes do código. O relatório deve ser colocado em um arquivo .zip junto com os códigos utilizados e enviado via Moodle.